UNIVERSITE DE LA MEDITERRANEE - AIX-MARSEILLE II

MEMOIRE

D'HABILITATION A DIRIGER DES RECHERCHES

présentée par

YSTAD Sølvi

VERS LE SENS DES SONS :
MODELISATION SONORE ET CONTROLE HAUT NIVEAU

JURY

Dr. BESSON, Mireille	Rapporteur
Pr. DAUDET, Laurent	Rapporteur
Pr. JENSEN, Kristoffer	Rapporteur
Pr. KRISTIANSEN Ulf	Examinateur

Remerciements

Je tiens à remercier Mireille Besson, Laurent Daudet et Kristoffer Jensen d'avoir accepté de rapporter ce travail pluridisciplinaire, Ulf Krsitiansen d'avoir accepté de participer à mon jury ainsi que tous mes collègues, collaborateurs et étudiants sans lesquels ce travail n'aurait pas été possible.

Je souhaite tout particulièrement remercier Mireille Besson et Richard Kronland-Martinet d'avoir accepté de se lancer dans la pluridisciplinarité en associant les compétences de leurs équipes, Mitsuko Aramaki pour son investissement sans faille dans la neuro-acoustique, Thierry Voinier pour son expertise temps-réel indispensable, le consortium masquage temps-fréquence, et en particulier Peter Balazs, Bernhard Laback, Sophie Savel, Sabine Meunier et Thibaud Necciari, pour avoir partagé leur savoir faire en psychoacoustique et mathématiques, et mes quatre doctorants (souffre-douleur) Marie-Céline Bezat, Mathieu Barthet, Adrien Merer et Jean François Sciabica pour le travail remarquable qu'ils ont fourni.

En ces temps où les instances d'évaluation tendent à devenir de plus en plus élitistes en cherchant à récompenser l'excellence individuelle plutôt que le dynamisme d'un groupe, je souhaite souligner qu'un travail pluridisciplinaire comme celui que je décrit dans ce document n'aurait pu aboutir sans associer les compétences entre chercheurs de différents horizons. Les résultats présentés dans ce document illustrent que le dynamisme et la motivation d'un groupe cohérent permet des avancées bien plus importantes que la somme des avancées individuelles de chaque participant.

Table des matières

1 Résumé des titres et travaux

- **Nom** : YSTAD
- **Prénom** : Sølvi
- **Date de naissance** : 15 février 1968
- **Fonction** : Chargé de Recherche 1er classe
- **Affectation** : Laboratoire de Mécanique et d'Acoustique, CNRS, 31 Chemin Joseph Aiguier, 13402 Marseille, Cedex 20
- **Téléphone** : 0491164259
- **Courriel** : ystad@lma.cnrs-mrs.fr

Diplômes

- **1998** : Doctorat de l'université Aix-Marseille II et de l'université de Trondheim, discipline : Acoustique, Traitement du Signal et Informatique Appliqués à la Musique. Titre de la thèse : Sound Modeling Using a Combination of Physical and Signal Models (Mention : Très honorable avec félicitations du jury)
- **1994** : Diplôme d'Etudes Approfondies en Mécanique (option Acoustique), Université Aix-Marseille II.
- **1992** : Diplôme d'ingénieur en électronique, NTH (Norges Tekniske Høgskole), Trondheim, Norvège

Mobilité

- Séjour Post-Doctoral à l'Université de Stanford (Californie) : septembre 2001 à septembre 2002
- Etudes d'ingénieur à l'Université de Trondheim (Norvège) : 1988-1992

Principales responsabilités scientifiques et administratives

- Responsable scientifique d'un projet ANR dans le cadre du programme blanc "jeunes chercheuses et jeunes chercheurs". Intitulé du projet : "Vers le sens des sons" (http ://www.sensons.cnrs-mrs.fr/). Période du projet : décembre 2005 à juin 2009.
- Responsable scientifique d'un projet PAI dans le cadre du programme

Amadeus avec l'Université de Vienne (Autriche) - institut ARI. Intitulé du projet : Représentations temps-féquence et perception des sons". Période du projet : janvier 2006 à janvier 2008

- Présidente des comités scientifiques des congrès internationaux CMMR-2007 (Computer Music Modelling and Retrieval) à Copenhague (Danemark), août 2007 (http ://www.lma.cnrs-mrs.fr/ cmmr2007), du CMMR2008 (http ://www.re-new.dk) à Copenhague (Danemark) mai 2008, du CMMR2009 (http ://www.lma.cnrs-mrs.fr/ cmmr2009) à Copenhague (Danemark), mai 2009 et co-présidente du CMMR2010 (http ://www.cmmr2010.etsit.uma.es/) à Malaga (Espagne), juin 2010.
- Membre du comité d'élaboration du réseau Européen MOSART (12 laboratoires Européens), coordinatrice de la section sud du réseau, 2000-2002.
- Membre du comité d'édition du Journal of Music and Meaning (http ://www.musicandmeaning.net)
- Membre des comités de programme de CMMR2003, CMMR2004, CMMR2005 et CMMR2007 (Computer Music Modelling and Retrieval), de Dafx06, Dafx07 (Digital Audio Effect International conferences) et de ICMC07 (International Computer Music Conference)
- Membre du jury de sélection pour un poste de professeur des universités à l'Université de Esbjerg (Danemark) en 2003 et pour 3 postes de maître de conférence en 2006.
- Membre du jury de sélection pour un poste de professeur des universités à NTNU (Trondheim, Norvège) en 2005
- Co-éditrice de 4 livres issus des conférences Computer Music Modelling and Retrieval, paru chez Springer-Verlag en 2006, 2008, 2009 et 2010.
- Co-directrice de 4 thèses (2 CIFRE et 2 MREN)
- Enseignement (4h) dans le cadre du master MPI 2A – Spécialité Acoustique (module sons et musique)
- Rapporteur de la thèse de Torunn Smevik, NTNU, Trondheim, Norvège, mai 2009.

Encadrement scientifique

- co-directrice de la thèse de Mathieu Barthet (financement MREN), école doctorale "Physique, Modélisation et Sciences pour l'Ingénieur". Thèse soutenue le 18 décembre 2008.
- co- directrice de la thèse de Marie-Céline Bezat (financement CIFRE avec la société Peugeot Citroën Automobiles) . Ecole doctorale "Phy-

sique, Modélisation et Sciences pour l'Ingénieur". Thèse soutenue le 19 décembre 2007.
- co- directrice de la thèse de Jean François Sciabica (financement CIFRE avec la société Peugeot Citroën Automobiles) Début de thèse : janvier 2008.
- directrice de la thèse d'Adrien Merer (financement MREN), école doctorale "Physique, Modélisation et Sciences pour l'Ingénieur". Début de thèse : octobre 2007.
- encadrements réguliers de plusieurs stages de Master 2 (français et étrangers)

Coopérations industrielles

Co-responsable scientifique de trois contrats industriels :

- Contrat LMA/France Télécom pour la mise en œuvre d'un système de prédiction de la résistance à la rupture de poteaux téléphoniques basé sur l'analyse de la réponse vibratoire du système à un impact (2000-2001).
- Contrat LMA/Peugeot Citroën Automobiles, dans le cadre de la direction de thèse de M.C Bezat sur l'étude des indices perceptifs associés à l'écoute de bruits de portières automobiles (2005-2007).
- Contrat LMA/Peugeot Citroën Automobiles, dans le cadre de la direction de thèse de J.F. Sciabica sur l'étude des indices perceptifs associés à l'écoute de bruits moteurs (2008-2010).

Liste de publications

Les publications jointes en annexe à la fin du document apparaissent en gras

Editions d'ouvrages :

1. "Auditory Display", Lecture Notes in Computer Sciences (LNCS), 5954, S. Ystad, M. Aramaki, R. Kronland-Martinet, K. Jensen, Springer Verlag Berlin Heidelberg, April 2010.
2. "Genesis of Meaning in Sound and Music" Lecture Notes in Computer Sciences (LNCS) 5493, 286 pages, S. Ystad, R. Kronland-Martinet and K. Jensen Editors, Springer Verlag Berlin Heidelberg , June 2009.
3. "Sense of Sounds", Lecture Notes in Computer Sciences (LNCS), 4969, Eds. R. Kronland-Martinet, S. Ystad and K. Jensen, Springer Verlag Berlin Heidelberg, June 2008.
4. "Computer Music Modelling and Retrieval", Lecture Notes in Computer Sciences (LNCS), 3902, 275 pages, R. Kronland-Martinet, Th. Voinier and S. Ystad Editors, Springer Verlag Berlin Heidelberg , 2006.

Articles soumis

5. Laback B., Balazs P., Necciari T., Savel S., Meunier S., Kronland-Martinet R., Ystad S. "Additivity of Masking for Short Tone Pulses" article soumis à JASA(2009), en révision
6. Necciari T., Savel S., Meunier S., Kronland-Martinet R., Ystad S., Laback B., Balazs P. "Auditory Masking for Gaussian-Windowed Stimuli" article soumis à JASA (2010)
7. Micoulaud-Franchi J.A., Aramaki M., Merer A., Cermolacce M., Ystad S., Kronland- Martinet R., Vion-Dury J. "Categorization and timbre perception of environmental sounds in schizophrenia" soumis à Psychiatry research, juin 2010

Articles à comité de lecture et chapitres d'ouvrage :

8. Sciabica J.F, Bezat M.C., Roussarie V., Kronland-Martinet R., Ystad S. "Timbre Characteristics of Interior Car Sound" Auditory Display, Springer Verlag Berlin Heidelberg, Avril 2010, pp. 377-391.
9. Aramaki. M., Gondre, C., Kronland-Martinet, R. and Ystad, S., "Imagine the Sounds : An Intuitive Control of an Impact Sound Synthesi-

zer", Auditory Display, Springer Verlag Berlin Heidelberg, Avril 2010, pp. 408-421.

10. **Aramaki M., Besson M., Kronland-Martinet R., Ystad S., "Controling the Perceived Material in an Impact Sound Synthesizer : towards a Neuro-Acoustic Approach", IEEE transactions on Audio, Speech and Language Processing (2010), Vol. PP, Issue 99, pp. 1-1, doi : 10.1109/TASL.2010.2047755**

11. **Barthet M., Depalle Ph., Kronland-Martinet R., Ystad S. "Acoustical correlates of timbre and expressiveness in clarinet performance", Journal of Music Perception (2010)**

12. **Barthet M., Depalle Ph., Kronland-Martinet R., Ystad S., "Analysis-by-Synthesis of Timbre, Timing and Dynamics in Expressive Clarinet Performance", Journal of Music Perception (2010)**

13. **Barthet M., Guillemain Ph., Kronland-Martinet R., Ystad S. "From Clarinet Control to Timbre Perception" Acta Acustica, Vol 96, 2010, pp. 678-689**

14. **Aramaki M., Marie C., Kronland-Martinet R., Ystad S., Besson M. "Sound Categorization and Conceptual Priming for Non Linguistic and Linguistic Sounds", Journal of Cognitive Neurosciences, Vol. 0, No. 0, Pages 1-15,
doi : 10.1162/jocn.2009.21398)**

15. **Schön D., Kronland-Martinet R., Ystad S., Besson M. "The evocative power of sounds : Conceptual priming between words and nonverbal sounds", Journal of Cognitive Neuroscience, Vol. 0, No. 0, Pages 1026-1035,
doi : 10.1162/jocn.2009.21302**

16. Aramaki, M., Besson, M., Kronland-Martinet, R. and Ystad, S., "Timbre Perception of Sounds from Impacted Materials" in Genesis of Meaning in Sound and Music, Ystad S., Kronland-Martinet R., Jensen K. Editors, LNCS #5493, Springer Verlag, Berlin Heidelberg, 2009, pp. 1-17

17. Aramaki M., Brancheriau L., Kronland-Martinet R., Ystad S., "Perception of impacted materials : sound retrieval and synthesis control perspectives" in Genesis of Meaning in Sound and Music, Ystad S., Kronland-Martinet R., Jensen K. Editors, LNCS #5493, Springer Verlag, Berlin Heidelberg, 2009, pp. 134-146.

18. Ystad, S. , Kronland-Martinet, R., Schön, D., Besson, M., "Vers une approche acoustique et cognitive de la sémiotique des objets sonores", Les Unités Sémiotiques Temporelles (UST), Nouvel outil d'analyse musicale, Théories et Applications, Collection Musique /Sciences, éditions Delatour 2008, pp. 73-83.

19. Merer A., Ystad S., Kronland-Martinet R., Aramaki M. "Semiotics of Sounds Evoking Motions : Categorization and Acoustic Freatures" in Sense of Sounds, Lecture Notes in Computer Sciences (LNCS) #4969, R. Kronland-Martinet, S. Ystad and K. Jensen Editors, Springer Verlag Berlin Heidelberg , pp 139-158, 2008.

20. Barthet M., Kronland-Martinet R., Ystad S. "Improving Musical Expressiveness by Time-Varying Brightness Shaping" in Sense of Sounds, Lecture Notes in Computer Sciences (LNCS) #4969, R. Kronland-Martinet, S. Ystad and K. Jensen Editors, Springer Verlag Berlin Heidelberg , 2008.

21. **Aramaki M., Baillères H., Brancheriau L., Kronland-Martinet R., Ystad S. "Sound quality assessment of wood for xylophone bars", Journal of the Acoustical Society of America 121(4), pp 2407-2420, April 2007.**

22. **Ystad, S., Magne, C., Farner,S., Pallone, G., Aramaki,M., Besson, M., Kronland-Martinet,R., "Electrophysiological Study of Algorithmically Processed Metric/Rhythmic Variations in Language and Music", EURASIP Journal on Audio, Speech, and Music Processing, numéro spécial Perceptual Models for Speech, Audio, and Music Processing , vol.2007, Article ID 30194, 13 pages, 2007. doi :10.1155/2007/30194.**

23. **Magne C., Astésano C., Aramaki M., Ystad S., Kronland-Martinet R., Besson M., "Influence of Syllabic Lengthening on Semantic Processing in Spoken French : Behavioural and Electrophysiological Evidence", Cerebral Cortex,**

doi : 10.1093/cercor/bhl174, Oxford University Press, janvier 2007, V17, N11, pp. 2659-2668.

24. Farner S., Kronland-Martinet R., Voinier T., Ystad S. "Timbre Variations as an Attribute of the Naturalness in the Clarinet Play", Lecture Notes in Computer Sciences (LNCS), N 3902, Springer Verlag, pp : 45-53, 2006.

25. **Aramaki M., Kronland-Martinet R., Voinier T., Ystad S. "A Percussive Sound Synthetizer Based on Physical and Perceptual Attributes", Computer Music Journal, Vol. 30(2), pp : 34-43, MIT Press, 2006.**

26. Magne, C. , Aramaki, M., Astesano, C., Gordon, R.L., Ystad,S., Farner, S., Kronland-Martinet, R., Besson,M., "Comparison of Rhythmic processing in Language and Music : An interdisciplinary approach", JMM : The Journal of Music and Meaning 3, Fall 2004/Winter 2005, sec.5.1, 2005.
27. D. Schön, P. Regnault, S. Ystad and M. Besson, "Sensory consonance : an Event- Related brain Potential study", Music Perception, 23(2), pp.105-117, ISSN 0730- 7829, University of California Press, décembre 2005.
28. Guillemain, Ph. , Helland, R. T., Kronland-Martinet, R. , Ystad, S., "The Clarinet Timbre as an Attribute of Expressiveness", Lecture Notes in Computer Sciences (LNCS), pp 246-259 Springer Verlag, 2005.

29. **Bensa, J., Dubois, D., Kronland-Martinet, R and Ystad, S., "Perceptive and Cognitive evaluation of a Piano Synthesis Model" Lecture Notes in Computer Sciences (LNCS), pp 232-245 Springer Verlag, 2005.**

30. Gobin, P., Kronland-Martinet, R., Lagesse, G.A., Voinier, Th., Ystad, S.,"From Sounds to Music : Different Approaches to Event Piloted Instruments , Lecture Notes in Computer Science, LNCS 2771, pp 225-246 Springer Verlag, 2003.

31. **Ystad, S., Voinier, Th., "A Virtually-Real Flute", Computer Music Journal (MIT Press), 25 :2, pp 13-24, Summer 2001.**

32. **Ystad,S. "Sound Modeling Applied to Flute Sounds", Journal of Audio Engineering Society, Vol. 48, No. 9, pp. 810-825, September 2000.**

33. Ystad, S. , "De la facture informatique au jeu instrumental", Les nouveaux gestes de la musique, Editions Parenthèses, pp.111-120, ISBN 2-86364-616-8, mai 1999.

34. **Kronland-Martinet, R., Guillemain, Ph., Ystad,S., "Modelling of Natural Sounds Using Time-Frequency and Wavelet Représentations", Organised Sound 2(3) : pp.179-191, Cambridge University Press 1997.**

Articles dans des actes des colloques à comité de lecture

35. Sciabica J.F, Bezat M.C., Roussarie V., Kronland-Martinet R., Ystad S., "Towards the Timbre Modeling of Interior Car Sound", 15th International Conference on Auditory Display, 18-21 Mai 2009, Copenhague, Danemark.
36. Aramaki. M., Gondre, C., Kronland-Martinet, R. and Ystad, S., "Thinking the Sounds : An Intuitive Control of an Impact Sound Synthesizer", 15th International Conference on Auditory Display, 18-21 Mai 2009, Copenhague, Danemark.
37. A.Merer, M. Aramaki, R. Kronland-Martinet and S. Ystad "Toward synthesis tools using 'evocation' as control parameters" Acoustics 08, 29 juin-4 juillet 2008, Paris, France.
38. B. Laback, P. Balazs, G. Toupin, T. Necciari, S. Savel, S. Meunier, S. Ystad and R. Kronland-Martinet "Additivity of auditory masking using Gaussian-shaped tones" Acoustics 08, 29 juin-4 juillet 2008, Paris, France.
39. T. Necciari, S. Savel, S. Meunier, S. Ystad, R. Kronland-Martinet, B. Laback and P. Balazs "Auditory masking using Gaussian-windowed stimuli" Acoustics 08, 29 juin-4 juillet 2008, Paris, France.
40. M. Barthet, P. Guillemain, R. Kronland-Martinet and S. Ystad "Exploration of timbre variations in music performance" Acoustics 08, 29 juin-4 juillet 2008, Paris, France.

41. M.C. Bezat, V. Roussarie, R. Kronland-Martinet and S. Ystad “Relations between acoustic parameters and perceptual properties : an approach by regressions tree applied to car door closure sounds” Acoustics 08, 29 juin-4 juillet 2008, Paris, France.
42. Aramaki M., Brancheriau L., Kronland-Martinet R., Ystad S. “Perception of impacted materials : sound retrieval and synthesis control perspectives “Proceedings of the 5th International Symposium on Computer Music Modeling and Retrieval (CMMR 2008), pp. 1-8, 19-22 May 2008, Copenhagen, Danemark
43. Bézat M.C., Roussarie V., Voinier Th., Kronland-Martinet R., Ystad S. “Car door closure sounds : Characterization of perceptual properties through analysis-synthesis approach” ICA 2007, Madrid.
44. Aramaki M., Kronland-Martinet R., Voinier Th., Ystad S. “Timbre control of real-time percussive synthesizer” conf. invitee, ICA 2007, Madrid.
45. Merer A., Ystad S., Kronland-Martinet R., Aramaki M., Besson M., Velay J.L. “Perceptual categorization of moving sounds for synthesis applications” ICMC 2007, Copenhague.
46. Barthet M, Depalle Ph., Kronland-Martinet R., Ystad S. “ The effect of timbre in clarinet interpretation” ICMC 2007 ,Copenhague, Danemark, 27-31 aout 2007.
47. Aramaki M., Baillères H., Brancheriau L., Kronland-Martinet R., Ystad S. “Relationship between sound classification of xylophone-like bars and wood species properties” CD-ROM Proceedings of the Thirtheenth International Congress on Sound and Vibration (ICSV13), Vienna, Austria, July 2-6, 2006. Eds. : Eberhardsteiner, J. ; Mang, H.A. ; Waubke, H., Publisher : Vienna University of Technology, Austria, ISBN :3-9501554-5-7.
48. Barthet M., Kronland-Martinet R., Ystad S. “Consistency of Timbre Patterns in Expressive Music Performance”, Proc. Of the 9th Int. Conf. On Digital Audio Effects (DAFx-06), September 18-20 2006, Montréal, Canada.
49. Schön, D., Ystad, S., Besson, M. and Kronland-Martinet, R. “An acoustical and cognitive approach to the semiotics of sound objects”, Bologne, Italy August 2006.
50. Bézat M.C., Roussarie V., Kronland-Martinet R., Ystad S., McAdams S. “Perceptual analyses of action-related impact sounds”, EURONOISE 2006, 30 May- 1 June 2006, Tampere, Finland.

51. Ystad, S. Kronland-Martinet, R., Schön D. and Besson, M., "Vers une approche acoustique et cognitive de la sémiotique des objets sonores", Journées USTs, 7-9 Décembre 2005.
52. S. Farner, R. Kronland-Martinet, T. Voinier, and S. Ystad, "Sound fluctuation as an attribute of naturalness in clarinet play", International Computer Music Modelling and Retrieval Conference (CMMR2005), pp. 210-218, Pisa, Italy, Sept. 2005.
53. M. Barthet, P. Guillemain, R. Kronland-Martinet et S. Ystad, "On the Relative Influence of Even and Odd Harmonics in Clarinet Timbre", International Computer Music Conference (ICMC), Barcelona, Spain, pp. 351-354, Sept. 2005
54. M. Aramaki, R. Kronland-Martinet, T. Voinier, S. Ystad, "Synthesis and Perceptual Manipulation of Percussive Sounds", International Computer Music Conference (ICMC), Barcelona, Spain, pp. 335-338, Sept. 2005
55. Julien Bensa, Danièle Dubois, Richard Kronland-Martinet and Sølvi Ystad "A cognitive approach to piano timbre" International Symposium on Computer Music Modeling and Retrieval, Esbjerg, Denmark 26-29, Mai 2004.
56. Ph. Guillemain, R. T. Helland, R. Kronland-Martinet , et S. Ystad "Towards a Better Understanding of the Relationship Between the Clarinet Timbre and the Playing" International Symposium on Computer Music Modeling and Retrieval, Esbjerg, Denmark, 26-29 Mai 2004
57. Reyna Leigh Gordon, Daniele Schön, Cyrille Magne, Corine Astesano, Sølvi Ystad, Richard Kronland-Martinet et Mireille Besson, "An fMRI study of the neural basis of song perception", Conference on Interdisciplinary Musicology, Graz, Austria, 15-18 April 2004.
58. S. Ystad, "Modellering, syntese og kontroll av lyd og musikksignaler", article invité, actes de Norsk Akustisk Selskap, Trondheim, Norvège, octobre 2003.
59. S Ystad, C Magne, S Farner, G Pallone, V Pasdeloup, R Kronland-Martinet and M Besson "Influence of rhythmic, melodic, and semantic violations in language and music on the electrical activity in the brain", Stockholm Music Acoustics Conference, (SMAC03), Stockhom, Sweden, 6-9 August 2003.
60. P. Gobin, R. Kronland-Martinet, G.A. Lagesse, Th. Voinier, S. Ystad, "Musical Interfaces", Computer Music Modeling and Retrieval (CMMR 2003), Montpellier, 26-27 Mai 2003.

61. S. Ystad, “Analysis-Synthesis of Musical Sounds by Hybrid Models", Article invité, Actes du Congrès International de l’ASA (American Society of Acoustics), publication CDROM, Cancun Mexico Décembre 2002.
62. Stefania Serafin, Patty Huang, Sølvi Ystad, Chris Chafe, Julius O. Smith III, “Analysis and Synthesis of Unusual Friction-Driven Musical Instruments”, International Computer Music Conference (ICMC) , Gotheburg, Sweden, Sept. 2002.
63. S. Ystad, Th. Voinier, “Analysis-Synthesis of Flute Sounds Using a Non-Linear Digital Waveguide Model”, International Computer Music Conference (ICMC), La Havana, Cuba, 17-22 September 2001.
64. R. Kronland-Martinet, Ph. Guillemain et S. Ystad, “From Sound Modeling to Analysis-Synthesis of Sounds”, Keynote paper, Workshop on Current Research Directions in Computer Music, pp. 200-208, Barcelona, Spain, 15-17 November 2001.
65. S. Ystad, Th. Voinier “Analysis-Synthesis of Flute Sounds with a Looped Non-linear Model”, Workshop on Current Research Directions in Computer Music, pp. 259-262, Barcelona, Spain, 15-17 November 2001.
66. J. Bensa, K. Jensen, R. Kronland-Martinet, S. Ystad “Perceptual and Analytical Analysis of the effect of the Hammer Impact on the Piano Tones”, International Computer Music Conference (ICMC), pp.58-61, Berlin (Allemagne) 27 août au 1 September 2000.
67. S. Ystad, P. Guillemain, R. Kronland-Martinet, “Sound Modeling of Transient and Sustained Musical Sounds”, International Computer Music Conference (ICMC), pp.112-116, Beijing (China) 22-27 October 1999.
68. S. Ystad, Th. Voinier, “Design of a Flute Interface to Control Synthesis Models ”, International Computer Music Conference (ICMC), pp.228-232, Beijing (China) 22-27 October 1999.
69. S. Ystad, “Musikalsk Akustikk ; konstruksjon av en digital fløyte”, Journal of the Acoustical Society of Norway (NAS Nytt), Vol.20, December 1999.
70. S. Ystad, “Identification and Modeling of a Flute Source Signal”, proc. of the DAFX (Digital Audio Effects) conference, pp. 187-190, 9-11 December 1999, Trondheim, Norway.
71. S. Ystad, Ph. Guillemain, R. Kronland-Martinet, “Sound Modeling From the Analysis of Real Sounds” proc. of the “First COST-G6 Work-

shop on Digital Audio Effects" (DAFX98), Barcelona, Spain, 19-21 November 1998.

72. Ph. Guillemain, R. Kronland-Martinet, S.Ystad, "Physical Modelling Based on the Analysis of Natural Sounds", proc. of the ISMA (International Symposium on Musical Acoustics), Vol.19 part 5, pp.445-450 University of Edinburgh, Scotland, August 1997.
73. S. Ystad, "Simulation of the Response of a Wind Instrument by a Waveguide Synthesis Model" proc. of the French acoustical society (cfa), Volume 1, pp. 581-584, Marseille 14-18 April 1997.
74. S. Ystad, P. Guillemain, R. Kronland-Martinet, "Estimation of Parameters Corresponding to a Propagative Synthesis Model Through The Analysis of Real Sounds" International Computer Music Conference (ICMC), pp.32-35, Hong Kong 19-24 August 1996.
75. S. Ystad, R. Kronland-Martinet, "A Synthesis Model for the Flute Transient by Analysis of Real Sounds Through Time-Frequency Methods" proceedings of the ICA (International Conference on Acoustics), Vol. III, pp. 525-528, Trondheim 26-30 June 1995.

Communications France

76. J.-A. Micoulaud-Franchi, M. Aramaki, A. Merer, M. Cermolacce, S. Ystad, R. Kronland-Martinet, L. Boyer, J Vion-Dury, "Reconnaissance du bizarre et du familier dans la perception de sons inouïs chez le patient schizophrène", Congrès Français de psychiatrie, Nice 2-5 décembre 2009.
77. M-C. Bezat, V. Roussarie, R. Kronland-Martinet, S. Ystad et S. McAdams, "Qualification perceptive des bruits d'impact. Application au bruit de fermeture de porte", Journées fondatrices du groupe Perception Sonore de la SFA, 18-19 janvier 2007.
78. M. Aramaki, M. Besson, R. Kronland-Martinet et S. Ystad, "Catégorisation sonore des matériaux frappés : Approche perceptive et perceptive et cognitive", Journées fondatrices du groupe Perception Sonore de la SFA, 18-19 janvier 2007.
79. S. Ystad, R. Kronland-Martinet, D. Schön et M. Besson, "Vers une approche acoustique et cognitive de la sémiotique des objets sonores", Journées fondatrices du groupe Perception Sonore de la SFA, 18-19 janvier 2007.

80. M. Barthet, S. Ystad et R. Kronland-Martinet, “Evaluation perceptive d’une interprétation musicale en fonction de trois paramètres d’expression : le Rythme, l’intensité et le Timbre”, Journées fondatrices du groupe Perception Sonore de la SFA, 18-19 janvier 2007.

Thèses encadrées et soutenues

81. Marie-Céline Bezat, “Perception des bruits d’impact : Application au bruit de fermeture de porte automobile”, Université de Provence, Aix-Marseille I, 19 décembre 2007.
82. Mathieu Barthet, “De l’interprète à l’auditeur une analyse acoustique et perceptive du timbre musical”, Université de Provence, Aix-Marseille I, 18 décembre 2008

Documents Audio-Visuels

83. C. Donguy, R. Kronland-Martinet, Ph. Guillemain, S. Ystad. “La modélisation du son”, film vidéo d’une durée de 12 minutes.
84. S. Ystad, “Sound Modeling using a Combination of Physical and Signal Models”, disque compact d’exemples sonores illustrant les exemples décris dans la thèse.
85. S. Ystad, Th. Voinier “Design of a Flute Interface to Control Synthesis Models”, film vidéo d’une durée de 14 minutes, présenté à l’ICMC99 Beijing (China), 1999.

2 Introduction

Les relations entre la structure acoustique des sons et la perception qui en résulte constituent le cœur de mes travaux de recherche actuels. Si cette connaissance relève de la recherche fondamentale, elle n'en est pas moins d'une importance capitale dans tous les domaines où l'information sonore est pertinente. On peut ainsi citer de nombreux exemples d'applications liées aux sciences et technologies de l'information et de la communication pour lesquelles ce "langage des sons" est central :

- la réalité virtuelle sonore qui nécessite la synthèse de sons cohérents avec une scène visuelle (bruit d'impact sur une structure de matériau donné par exemple),
- la qualité sonore, fondamentale pour une bonne représentation mentale de l'environnement (notion de qualité des objets tel le jugement porté sur une automobile à partir du bruit produit par ses portières),
- le “design” sonore nécessitant une meilleure compréhension de la sémiotique des sons et le retour sensoriel à une information sonore qui s'avère être souvent cruciale (information sur le type d'évènement, degré d'alarme, interprétation musicale. . .),
- le codage de signaux audio qui pourrait être considérablement optimisé (compression de données) grâce à la prise en compte des processus primaires et centraux liés à l'écoute.

Si les énormes avancées méthodologiques permettent aujourd'hui de construire par synthèse numérique des sons complexes et réalistes, les indices perceptifs et cognitifs repérés lors de l'écoute sont encore relativement peu connus. Ainsi peut-on par exemple re-synthétiser le son produit par divers pianos [Bensa, 2003] sans pour autant bien comprendre ce qui fait la différence auditive entre un bon et un excellent piano [Bensa et al., 2005]. De même est il extrêmement difficile de contrôler un système de synthèse sonore à partir d'une description sémantique des sources (par exemple : bruit d'impact d'une bouteille de champagne sur une coque plastique de bateau) [Gaver, 1993; Aramaki and Kronland-Martinet, 2006]. L'acoustique physique permet de mieux comprendre les relations entre une structure vibrante et le champ de pression rayonné. C'est ainsi que l'on peut prédire le champ acoustique produit par de nombreuses sources sonores. La propagation de ces ondes acoustiques communique à notre système auditif les informations nécessaires à la perception des sources sonores dans leur environnement, mais qu'en est-il de la relation entre les ondes acoustiques et la représentation mentale du phénomène générateur ? La communauté des Neurosciences manifeste un vif intérêt pour comprendre les relations entre les sons (notamment la parole et la musique) et les processus cérébraux mis en jeu lors de l'écoute [Patel et al., 1998; Overy, 2003]. Ces travaux, dont le but principal est de mieux définir l'architecture cérébrale fonctionnelle du cerveau, apportent une nouvelle vision des phénomènes liés à l'écoute des sons [Griffiths and Warren, 2004] ainsi qu'un ensemble de méthodes permettant la mesure objective de l'activité cérébrale (potentiels évoqués, IRMf, ...). Le travail que je décris ici repose sur ma conviction profonde que l'association de ces domaines de recherche ouvre une voie originale à l'abord d'une question jusqu'alors hors de portée : quelle est la relation entre les sons et le sens qu'ils communiquent ? L'approche que je propose ici est axée sur l'établissement d'un lien entre mathématiques, acoustique, traitement du signal, perception et cognition, l'ensemble étant centré sur la synthèse numérique des sons. C'est en effet grâce à la synthèse que le lien entre le son physique et le son perçu peut mieux être établi, en revisitant un concept qui a déjà fait ses preuves dans le domaine de la conception des sons par ordinateur : l'analyse par synthèse [Risset and Wessel, 1999].

Les problèmes que j'aborde dans ce document ne sont pas nouveaux et le son à toujours été l'objet de forts enthousiasmes. La quête de sons nouveaux a passionné compositeurs et musiciens depuis toujours. Le compositeur Edgar Varèse (1883-1965) disait : "*je rêve d'instruments obéissant à la pensée et qui, avec l'apport d'une floraison de timbres insoupçonnés, se prêtent aux combinaisons qu'il me plaira de leur imposer et se plient à l'exigence de mon rythme intérieur*. Bien avant que les premiers sons sur ordinateur soient réalisés, un grand nombre d'instruments électroniques tels que le Telharmonium ou Dynamophone, le Theremin, le Dynaphone ou les Ondes Martenot ont été développés dans les années 1920 et 1930 [Griffiths, 1979; Weidenaar, 1995; Rhea, 1989]. En 1957 Max Mathews a développé un outil permettant de générer des sons à partir d'ordinateurs à l'aide de son programme Music 1 [Mathews, 1963]. Cet événement peut être considéré comme la naissance de la synthèse numérique. L'utilisation de l'ordinateur pour produire des sons a permis aux chercheurs, musiciens et compositeurs d'accéder à un nouveau monde de sons tout en s'affranchissant des contraintes physiques liées aux sources sonores, et ce, avec la précision inhérente au calcul numérique. Même si le contrôle des processus de synthèse faisait déjà parti des préoccupations des chercheurs et musiciens, les ordinateurs de l'époque ne permettaient pas de synthétiser des sons en temps-réel, limitant ainsi les possibilités d'utilisations interactives de la synthèse. Les recherches étaient ainsi essentiellement axées sur les processus de synthèse.

Aujourd'hui les synthétiseurs numériques sont facilement accessibles et la puissance des processeurs qui leur sont associés permet l'implémentation d'algorithmes (complexes) de synthèse temps-réel. Dès lors, de nouveaux champs d'investigations axés sur le contrôle temps-réel prennent une place essentielle et permettent la réalisation et la mise en oeuvre d'un grand nombre d'interfaces et de stratégies de contrôle, visant pour la plupart des applications musicales [Cook, 2002; Gobin et al., 2003; Miranda E. and Wanderley, 2006]. Si la notion de contrôle s'est essentiellement focalisée sur les paramètres clefs de la structure musicale (fréquence, intensité, durée, timbre, ...), de façon à produire de nouveaux instruments calqués sur les instruments acoustiques, la relation entre le son et l'impact perceptif produit a toujours été dans le champ de mire des chercheurs. Cette notion s'est notamment appuyée sur le concept de timbre sonore, attribut subjectif susceptible de différencier des sons de même hauteur tonale et de même sonie.

Grâce à la ductilité de la synthèse, une meilleure compréhension de la pertinence perceptive des paramètres représentatifs du signal peut être obtenue en utilisant notamment le concept de "l'analyse par synthèse" [Risset and Wessel, 1999]. Les premières études sur le timbre se sont appuyées sur une approche analyse-synthèse [Grey, 1977]. Il s'agissait ici de construire des sons de synthèse simulant des instruments de musique, dégrader certaines parties du signal et identifier la pertinence perceptive des paramètres par un test d'écoute. Malgré la simplicité des méthodes d'analyse et de transformation sonore utilisées, certains paramètres tels que le temps d'attaque et le barycentre spectral ont pu être identifiés comme étant pertinents du point de vue perceptif. Des études plus récentes sur le timbre ont permis de mettre en évidence un certain nombre de descripteurs de timbre représentatifs de catégories spécifiques de sons [McAdams et al., 1995; Plomp, 1970; Hajda et al., 1997; Kendall and Carterette, 1991; Caclin et al., 2005] et la communauté scientifique MIR (Music Information Retrieval) a proposé une grande quantité de descripteurs basés sur la structure du signal. Malheureusement, l'ensemble de ces études n'a pas permis de définir de descripteurs universels adaptés aux différentes classes de sons et laisse encore ouvert le problème du "modèle de timbre".

De nouveaux domaines de recherche tels que la réalité virtuelle, la sonification et le design sonore ont émergé depuis peu. Il s'agit ici d'utiliser des sons comme porteurs d'information et donc de dépasser la stricte notion de timbre sonore. Ces domaines connaissent de nombreuses applications, notamment dans le contexte industriel (industrie automobile, téléphone) dans le domaine de la santé (aides aux mal voyants) dans le domaine de jeux vidéo, cinéma d'animation etc. Des sons pré-enregistrés sont le plus souvent utilisés dans ce contexte, bien que la synthèse sonore semblerait être l'outil idéal pour ce type d'applications. Cependant, nonobstant la qualité et le réalisme sonore procurés par la synthèse numérique, la difficulté liée au contrôle des sons "par le ressenti" freine considérablement son utilisation. Ces nouvelles applications procurent un cadre idéal à mes travaux de recherche tant l'enjeu premier vise à mieux comprendre l'adéquation entre la structure interne des sons et le sens généré. Du point de vu de la synthèse numérique, ce problème amène naturellement à la notion de contrôle et de "mapping" entre les paramètres de synthèse et une interface "intuitive" permettant la génération de sons à partir de mots décrivant des objets sonores et leur interactions ou même directement à partir de la scène visuelle correspondant. La construction de méthodologies de contrôle devient alors un domaine de recherche à part entière et nécessite des études perceptives permettant d'identifier les grandeurs responsables d'évocations. On parle dans ce cas de contrôle “haut niveau”. La question du sens porté par des sons devient alors essentielle et fait naturellement référence à des recherches menées dans le domaine des neurosciences cognitives. C'est dans ce contexte que j'ai initié une collaboration avec l'Institut de Neurosciences Cognitives de la Méditerrannée (INCM) à Marseille. Cette collaboration entre chercheurs issus de la modélisation et de l'acoustique d'une part et des neurosciences cognitives d'autre part, s'est avérée très intéressante et fructueuse, et m'a permis d'obtenir un financement auprès de l'Agence Nationale de la Recherche dans le cadre du programme blanc “jeunes chercheuses et jeunes chercheurs” 2005. J'ai ainsi introduit une nouvelle thématique de nature fortement pluridisciplinaire au laboratoire de mécanique et d'acoustique : "le sens des sons". Ce projet a suscité un grand intérêt dès son démarrage, et plusieurs nouveaux collaborateurs s'y sont associés. J'ai ainsi collaboré avec plusieurs institutions françaises et étrangères tels le LATP et le centre de Réalité Virtuelle à Luminy, ARI (Acoustics Research Institute) à Vienne en Autriche, l'Université de Aarhus au Danemark ainsi que l'université McGill au Canada. Depuis 2005, j'ai co-organisé 4 conférences internationales (Computer Music Modelling and Retrieval) de 2005 à 2009 autour des thèmes “PLAY”, “Sense of Sounds”, “Genesis of Meaning in Sound and Music” et “Auditory Display”. Suite à ces conférences, j'ai co-édité 4 livres [Kronland Martinet et al., 2006, 2008; Ystad et al., 2009, 2010]. J'ai également co-encadré 4 thèses autour des problématiques liées à la fois aux aspects fondamentaux et appliqués, en vue des applications musicales [Barthet, 2008], réalité virtuelle [Merer et al., 2008b] et design sonore dans le cadre industriel [Bezat, 2007; Sciabica et al., 2010].

Tout au long de mes recherches, j'ai étudié le "sens des sons" à différents niveaux de complexité (atomique, sons isolés et structures complexes) en associant à la synthèse numérique des protocoles issus de la psychologie expérimentale et des méthodes d'imagerie cérébrale. Ces approches ont permis de construire des méthodologies générales permettant d'identifier les paramètres du signal pertinents du point de vue perceptif et de proposer des stratégies de contrôle haut niveau des processus de synthèse. Les expériences issus des méthodes d'imagerie cérébrale ont également permis de répondre à un certain nombre de questions fondamentales liés au sens attribué aux sons [Aramaki et al., 2009c; Schön et al., 2009].

Bien que les méthodes développées dans le cadre de mes recherches se veulent générales et ne se limitent pas à un domaine spécifique, je présente dans la suite de ce document mes travaux en deux parties en distinguant les applications liées aux sons musicaux et les applications liées aux sons environnementaux et industriels. La deuxième partie est étroitement liée aux recherches effectuées dans le cadre du projet ANR senSons, du projet PAI Amadeus ("Représentations temps-fréquence et perception des sons", programme EGIDE) et des contrats LMA/ Peugeot Citroën Automobile.

3 Les sons musicaux

Durant ma thèse de doctorat [Ystad, 1998], l'essentiel de mes recherches concernait l'analyse-synthèse et contrôle de sons issus d'instruments de musique. J'ai ainsi proposé un modèle hybride de synthèse de sons de flûte, associant un modèle physique prenant en compte le résonateur et un modèle non linéaire de signal pour la source excitatrice. J'ai également développé une méthodologie générale permettant l'extraction des paramètres de synthèse à partir de l'analyse de sons naturels. Les sons obtenus par synthèse ont été jugés très positivement et se rapprochaient très fortement des sons naturels de référence. L'intérêt de la synthèse réside cependant dans les possibilités de dépassement des contraintes physiques, ce qui permet d'étendre la palette de sons générés à des sons nouveaux pour lesquels le contrôle ne peut plus être associé au geste musical traditionnel. Ceci m'a amené vers la construction de nouvelles interfaces adaptées à ces modèles de synthèse [Ystad, 1999; Ystad and Voinier, 2001; Ystad and Voinier, 2001a,b]. Ces interfaces, qui contrairement aux "instruments traditionnels" ne sont pas restreintes par la mécanique, préparent naturellement à de nouveaux gestes visant l'exploitation la plus efficace des possibilités offertes par l'ordinateur. J'ai dans ce cadre participé à l'élaboration d'interfaces de natures très différentes (flûte augmentée, radio baton, modèle percussif. . .) [Gobin et al., 2003]. Si le contrôle des paramètres de synthèse liés à la structure musicale (fréquence, durée, intensité,...) ne pose pas de problème majeur, le contrôle du timbre en tant que vecteur de l'expressivité musicale est apparu extrêmement complexe. En effet, la mise en correspondance des paramètres de synthèse et de contrôle suppose une stratégie de regroupement des paramètres de bas niveau au sein de grandeurs ayant une pertinence perceptive au regard du ressenti escompté.

Le lien entre les caractéristiques du signal et perception a fait l'objet d'un grand nombre d'études depuis la construction des premiers sons de synthèse. La curiosité des compositeurs et musiciens pour ces nouveaux sons qui donnent accès à un monde de timbres sans limitations a poussé les chercheurs à développer des outils de transformations permettant de révéler la pertinence perceptive de certaines structures sonores. Ces informations sont à la fois importantes pour la qualité de la synthèse et peuvent permettre une réduction de donnés facilitant ainsi la resynthèse en temps-réel. C'est par une approche d'analyse de sons réels que Schaeffer a proposé des règles d'identité timbrale de sons de piano [Schaeffer, 1966]. Par une approche analyse-synthèse Risset a montré l'importance de l'évolution temporelle des différentes composantes spectrales des sons de trompette [Risset, 1965]. Cette étude révèle que l'augmentation de la largeur spectrale en fonction de l'amplitude est liée à l'effet "cuivré" de l'instrument. En introduisant des irrégularités fréquentielles aux sons de synthèse, Matthews et collaborateurs [Mathews et al., 1965] ont fortement amélioré le réalisme de l'attaque de simulation de cordes frottées. Par cette même approche j'ai dans le cadre de ma thèse montré que, lors de l'attaque de sons de flûte, les composantes spectrales ne démarrent pas en même temps et leurs temps d'attaque augmente avec le rang harmonique. En ce qui concerne la partie stable du son, qui est dominée par l'effet vibrato dans le cadre de la flûte, j'ai montré que les variations d'amplitude (tremolo) et variations de fréquence sont en phase, impliquant que le vibrato peut être piloté par le jet d'air de l'instrumentiste lors du contrôle du modèle de synthèse.

Depuis les travaux pionnier de Grey sur le timbre, plusieurs études se basant sur des tests perceptifs ont permis de définir des grandeurs acoustiques (descripteurs de timbre) qui caractérisent des catégories de sons [McAdams et al., 1995; Krumhansl, 1989]. La méthodologie repose le plus souvent sur la collecte de mesures de dissimilarités ("distances perceptives") entre paires de stimuli différents. Les jugements de dissimilarité peuvent ensuite être représentés dans un espace de timbre reflétant des distances perceptives entre sons à l'aide d'une analyse multidimensionnelle de proximité (MDS). Nous avons utilisé cette approche lors d'une étude sur l'influence perceptive de paramètres de contrôle de sons de clarinette dans le cadre de la thèse de Mathieu Barthet que j'ai co-dirigée (voir section 3.1).

Les études qui suivent s'appuient sur l'interaction de la synthèse numérique et des sciences de la perception et de la cognition. Elles abordent le point de vue sensoriel lié à l'écoute des sons musicaux en tentant de répondre à la problématique posée par le contrôle des modèles numériques d'instruments de musique (tant pour la note isolée que pour la séquence musicale expressive), mais également à la question du calibrage perceptif des modèles de synthèse et de transformation des sons.

3.1 Du contrôle au timbre : l'exemple de la clarinette

Dans le but de mieux comprendre le lien entre les paramètres de jeu d'un instrument de musique, les sons engendrés et leur perception, des sons de clarinette ont été étudiés. Ce travail a été réalisé dans le cadre de la thèse de Mathieu Barthet [Barthet, 2008]. La clarinette est un bon candidat pour l'étude des relations entre le geste musical et les sons engendrés car cet instrument donne lieu d'une part à des modèles physiques réalistes [Chaigne and Kergomard, 2008; Silva et al., 2008] (contrairement à la flûte) et permet d'autre part un contrôle continu du son lors du jeu instrumental. Un modèle de synthèse basé sur le comportement physique d'une clarinette a été utilisé pour construire des stimuli contrôlés [Guillemain et al., 2005]. Les paramètres de contrôle de ce modèle correspondent aux paramètres principaux de jeu d'un vrai instrument, à savoir la pince ou l'ouverture du canal d'anche, la pression dans la bouche du musicien, et la longueur équivalente du tuyau. Même si ces paramètres ne permettent pas d'accéder aux subtilités du contrôle d'un vrai instrument, les sons obtenus pour différentes combinaisons des paramètres de contrôle balaient une large palette de timbres réalistes, permettant ainsi de mieux appréhender l'influence perceptive de ces contrôles dans une situation de jeu conventionnelle. 15 stimuli ont ainsi été engendrés avec le modèle de synthèse pour une longueur de tuyau constante, et évalués par 16 sujets lors d'un test de dissimilarité. Les jugements ont donné lieu à des analyses statistiques (analyse multidimensionnelle de proximité) qui ont permis de représenter les sons dans un espace de timbre à trois dimensions reflétant leurs distances perceptives. Des corrélations entre les dimensions de l'espace de timbre, des descripteurs de timbre classiques et les paramètres de contrôle ont ensuite été recherchés. Les résultats de l'étude ont révélé que la première dimension de l'espace de timbre est corrélée au logarithme du temps d'attaque et au centre de gravité spectral du son. Cette dimension est également corrélée au paramètre de jeu correspondant à la pince. La deuxième dimension est corrélée à la deuxième composante du tristimulus, tandis que la troisième dimension est corrélée au rapport pair/impair des composantes spectrales et au paramètre de jeu lié à la pression dans la bouche du musicien (Figure 1).

Ces résultats sont importants et permettent de calibrer les paramètres

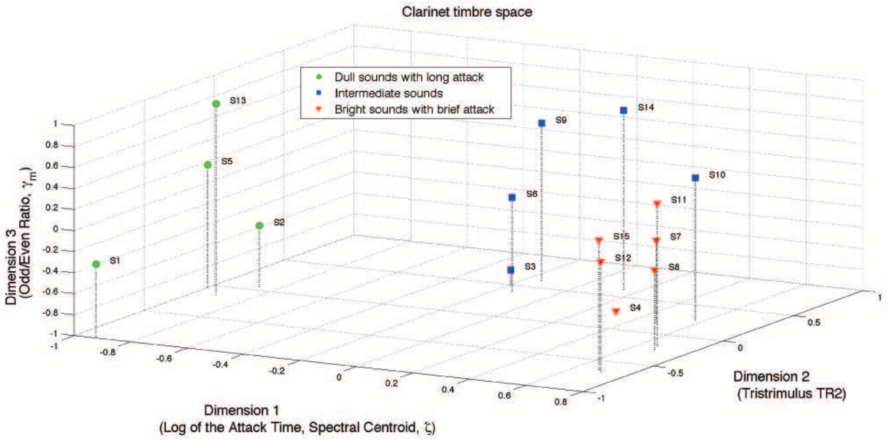


FIGURE 1: Espace de timbre des sons de synthèse de clarinette et corrélats acoustiques et mécaniques.

de contrôle du modèle de synthèse afin d'obtenir un timbre particulier pour un jeu spécifique des paramètres de contrôle. Le lien entre descripteurs de timbre et contrôle peut également faciliter l'identification de l'instrument ou de l'interprète pour automatiquement identifier un instrument ou un instrumentiste dans le cadre d'extraction automatique d'informations contenues dans la musique ("music information retrival"). Cette étude exhibe enfin que l'instrumentiste exerce un contrôle fin sur le timbre, et ce tout au long de la génération du son. Elle est ainsi le point de départ d'une étude plus ambitieuse sur l'influence des variations de timbre dans la structure musicale et en particulier sur la relation entre l'intention expressive de l'interprète et les variations de timbre [Barthet et al., 2010c].

3.2 Timbre et interprétation

L'interprétation musicale est le résultat d'une transformation du signal notationnel du compositeur en un signal acoustique perçu par un auditeur. Elle implique à la fois l'intention du musicien, ses gestes de contrôle sur l'instrument et les possibilités acoustiques de ce dernier. La révélation des paramètres acoustiques corrélés à l'expression musicale revêt un intérêt fondamental pour la compréhension de la perception musicale et trouve de multiples applications en synthèse sonore. De nombreuses études ont montré que les variations de hauteurs, de rythme et d'intensité sont corrélées à l'intention expressive de l'interprète [Seashore, 1938; Gabrielsson, 1999; Friberg, 1995]. Le timbre est rarement étudié dans ce contexte, même si c'est un paramètre essentiel pour les sons pouvant, pour certains instruments, être modifié par l'instrumentiste pendant la production du son. Outre le manque de notations musicales faisant référence au timbre, une des raisons est probablement la complexité de cet attribut encore mal compris. Même si certains descripteurs de timbre permettent de distinguer des catégories de sons, les variations fines de timbre au sein d'un même instrument sont mal connues. De plus, les études concernent essentiellement des sons isolés et stationnaires et ne permettent pas de caractériser la dynamique des sons qui est primordiale dans le cadre de l'interprétation. Pour aborder ce problème, nous avons, dans le cadre de la thèse de Mathieu Barthet [Barthet, 2008], développé une méthodologie à partir de l'analyse du signal. Il s'agissait dans un premier temps d'enregistrer des extraits musicaux issus de répertoires traditionnels (Bach et Mozart) joués sur une clarinette par un clarinettiste professionnel, segmenter les notes, puis calculer des descripteurs de timbre pour chaque note. L'évolution temporelle des descripteurs spectraux a été obtenue par transformée de Fourier à court terme. Pour tester le rôle du timbre dans l'interprétation, les extraits musicaux ont été joués de manière expressive, puis de manière inexpressive. La comparaison par analyses de variances (ANOVA) à deux facteurs (l'intention expressive et la note) entre les deux types d'interprétation a permis de montrer que certains descripteurs de timbre tels que le temps d'attaque, le centre de gravité spectral et le rapport pair/impair (Figure 2) se distinguent de façon significative entre les deux types d'interprétation, ce qui signifie que le timbre joue un rôle important dans l'interprétation [Barthet et al., 2010a].

Pour comprendre la pertinence perceptive des variations de timbre sur l'interprétation, une approche par analyse-synthèse a permis de transformer les interprétations en réduisant les déviations expressives effectuées par l'instrumentiste.Trois transformations ont été définies afin de modifier de manière indépendante le timbre, le rythme et la dynamique. Le timbre a été transformé en éliminant les variations temporelles du centre de gravité spectral. Les déviations rythmiques ont été supprimées pour que les durées des notes soient conforme à une transcription exacte des indications de la partition, tandis que les variations d'énergie ont été limitées par l'implémentation d'un compresseur permettant de réduire l'ambitus de variation de la dynamique. Les trois transformations de base ainsi que les quatre combinaisons (entre les transformations) ont été appliquées aux séquences de clarinette expressives. L'évaluation des effets perceptifs des transformations a été effectuée par 20 musiciens lors d'un test de jugement de préférence issus d'une comparaison par paires des différents stimuli. Les analyses sta-

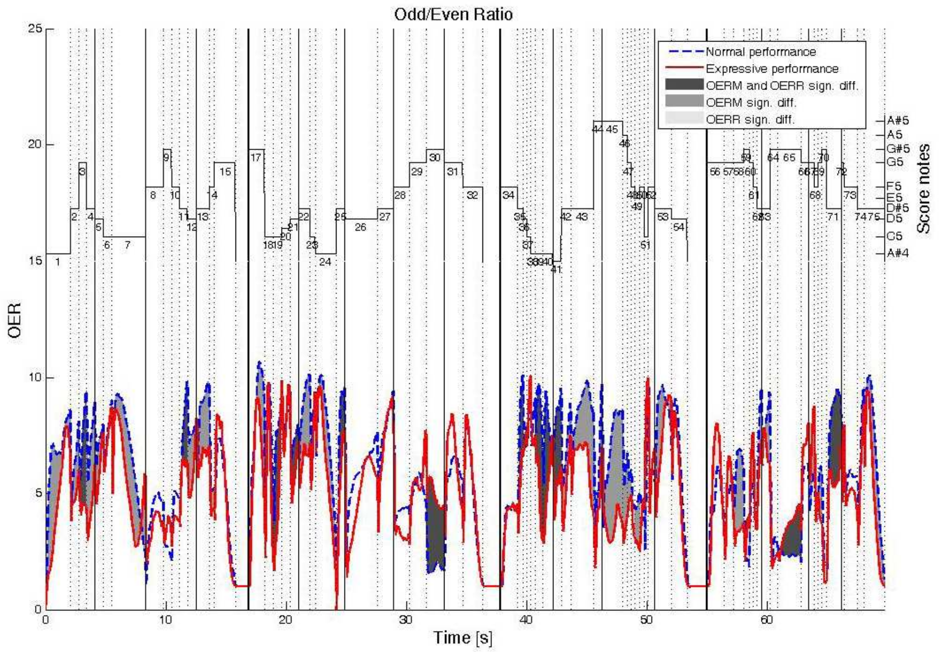


FIGURE 2: Rapport Impair-Pair (Odd Even Ratio). Moyens pour les 20 interprétations scolaires (ligne pointillée) et les 20 interprétations expressives (ligne solide) de l'extrait de Bach. Les notes ou groupes de notes pour lesquelles des différences systématique sont été déterminées par l'ANOVA, sont mises en évidence (zones grisées). Les notes de la séquence musicale sont représentées dans la partie supérieure de la figure. Leur rang au sein de la structure musicale est également reporté. La figure est extraite de la thèse de Mathieu Barthet.

tistiques montrent que le gel du timbre (par le biais du centre de gravité spectral) induit une plus grande perte de qualité musicale que l'annulation des variations rythmiques ou la compression de la dynamique [Barthet et al., 2010b].

Dans l'idée de proposer un modèle d'interprétation prenant en compte les variations de timbre, un deuxième test perceptif a été effectué pour vérifier si l'application de variations expressives de brillance à une séquence initialement inexpressive peut améliorer la qualité musicale. Pour engendrer des séquences musicales, un échantillonneur, contrôlé selon le protocole MIDI (Musical Instrument Digital Interface) qui transmet des informations relatives aux hauteurs et durées des notes, mais également relatives à leur intensité et à leur timbre (le paramètre qui contrôle ces aspects et appelé vélocité pour des raisons historiques liées au fonctionnement du piano) a été utilisé. L'extrait musical de Bach étudié précédemment a été utilisé pour engendrer des séquences musicales avec 5 instruments différents, à savoir clarinette, piccolo, violoncelle, guitare et sitar. Des séquences inexpressives ont été engendrées en figeant le paramètre MIDI vélocité (relative à l'intensité et au timbre), et les séquences expressives ont été obtenues en ajoutant l'évolution temporelle de la brillance extraite lors des analyses précédentes. Conformément au test précédent, un test de jugement de préférence issus d'une comparaison par paires des différents stimuli a été effectué sur 20 sujets. Les résultats montrent que l'application de variations temporelles de timbre améliore l'appréciation des séquences pour les instruments entretenus tels la clarinette, le piccolo et le violoncelle. L'effet de la transformation sur les instruments non-entretenus (guitare, sitar) est moins prononcé. Ce résultat n'est pas surprenant, étant donné que les variations de timbre ont été extraites d'un instrument entretenu (clarinette) ce qui explique leur incohérence avec ces types d'instruments. Cette étude confirme encore une fois l'importance du timbre dans l'interprétation. En vue de l'élaboration de modèles prédicteurs des variations de timbres reposant sur la structure musicale, il conviendrait dans les études futures d'analyser une large variété d'extraits musicaux avec différents instruments pour faire un rapprochement systématique des analyses avec la structure musicale et les spécificités de chaque instrument [Barthet et al., 2008].

3.3 Modélisation sonore et sciences cognitives : une première rencontre

Les sections précédentes illustrent comment l'approche analyse-synthèse permet d'évaluer le rôle du timbre dans l'interprétation. Cependant, lorsqu'il s'agit d'analyser le jugement global porté sur des sons, d'autres méthodes d'investigation sont nécessaires. Il m'a alors semblé naturel de m'inspirer de protocoles expérimentaux issus des sciences cognitives. Dans un premier temps, deux approches différentes faisant l'objet de collaborations avec le Laboratoire d'Acoustique Musicale (LAM) à Paris et l'Institut des Neurosciences Cognitives de la Méditerranée (INCM) à Marseille ont été abordées. Dans le cadre de la collaboration avec le LAM, un protocole de catégorisation libre a été utilisé pour effectuer une évaluation perceptive et cognitive de sons de piano et dans le cadre de la collaboration avec l'INCM les variations du potentiel électrique cérébral (ou Potentiels Evoquées) ont été analysées en vue d'étudier l'influence de l'allongement de syllabes dans le langage et de notes dans la musique.

3.3.1 Catégorisation libre de sons de piano

L'objectif de cette étude réalisée en collaboration avec J. Bensa et D. Dubois du Laboratoire d'Acoustique Musicale (LAM) à Paris, était d'évaluer, par un protocole de classification libre, la qualité de sons de piano engendrés par un modèle de synthèse hybride. Malgré le réalisme du modèle de synthèse développé par J. Bensa durant sa thèse [Bensa, 2003], une question est restée en suspend, à savoir qu'est-ce qui différencie un excellent piano d'un piano standard ? Il ne s'agissait pas seulement d'analyser des corrélations entre catégories de sons et paramètres de timbre, mais d'identifier les jugements perceptifs de façon globale. En effet, pour accéder aux critères associés aux processus d'interprétation chez les auditeurs, il faut inclure les caractéristiques "haut niveau" dans la modélisation des qualités acoustiques de sons de piano. La méthodologie qui a été utilisée dans cette approche a déjà fait ses preuves dans d'autres évaluations perceptives liées à la vision et à l'olfaction [Rosch, 1978; Dubois, 2000]. Cette méthodologie est basée sur des suppositions théoriques concernant des catégories cognitives et leur lien avec le langage. En demandant aux sujets de classer des stimuli librement en fonction de critères personnels et de commenter leur classement final, l'expérimentateur a accès aux représentations cognitives à travers l'analyse psycholinguistique des commentaires verbaux des sujets. A partir de l'analyse sémantique des commentaires, des variations subtiles de timbre qui ne sont pas reconnues par des descripteurs classiques, peuvent alors être identifiées.

Il s'agissait dans cette étude d'analyser l'influence perceptive de deux caractéristiques physiques impliquées dans la production sonore des sons de piano, à savoir la raideur de la corde (responsable de l'inharmonicité) et la modulation de tension qui crée des partiels "fantômes" dans le spectre (modes longitudinaux). Un modèle de synthèse de type hybride basé sur une combinaison entre un modèle physique (de type guide d'onde) et un modèle additif a été utilisé pour engendrer 17 stimuli (pour une même note B1 : f0=61.7Hz) à partir de modifications indépendantes des deux paramètres physiques. Trois catégories de sujets ont participé à ce test : des pianistes, des musiciens non-pianistes et des non-musiciens.

Les résultats montrent que le premier niveau de catégorisation chez les pianistes correspond aux critères physiques liés à l'inharmonicité et le niveau secondaire correspond aux partiels fantômes. Chez les sujets non-pianistes, les catégories ne sont pas de même nature. En effet, trois catégories sont observées à un seul niveau de catégorisation. Ces résultats montrent que les critères de catégorisation dépendent fortement de l'expérience des sujets impliquant que différentes catégories cognitives peuvent émerger du même ensemble de stimuli décrits de façon univoque par la physique. Chez les pianistes, le degré d'expertise se traduit également par la caractérisation des sons par des objets proches (ici différents types de piano). Cette étude a aussi montré que les sons sont classés en fonction de leur distance du son prototypique. En effet, pour des valeurs extrêmes des paramètres de synthèse, on observe une discontinuité dans la catégorisation cognitive qui se traduit par un changement de type de jugement. Dans ces cas les sons sont trop éloignés du prototype pour que les sujets puissent se constituer une représentation d'un son réel, et l'analyse se fait de façon analytique *i.e.* directement liée aux propriétés du signal [Bensa et al., 2005].

3.3.2 Influence de l'allongement de syllabes dans le langage et de notes dans la musique

La discontinuité dans la catégorisation de sons de piano décrite précédemment, soulève le problème lié aux limites perceptives des modifications sonores. Ce problème a également été abordé dans le cadre de la thèse de Gregory Pallone [Pallone, 2003] qui concernait le transcodage de la bande son entre le cinéma et la vidéo. Il s'agissait ici de dilater ou de comprimer le signal sonore sans en altérer le timbre. La méthode retenue est de type temporelle et consiste à enlever ou insérer de petits segments sonores. Le choix de la taille des éléments à insérer était ici crucial, ce qui nous a amené à envisager des expériences sur l'influence perceptive des modifications sonores. Cette problématique rejoint la thématique de l'équipe de recherche, "Langage et Musique" (M. Besson, C. Magne) de l'INCM, Marseille, qui s'intéresse à la comparaison entre traitement cognitif du langage et de la musique. L'association des expertises dans ces deux domaines nous a amené à proposer un protocole expérimental sur la perception du rythme dans le langage et la musique.

Le rythme représente une partie essentielle de la musique. Est-ce le cas aussi en ce qui concerne le langage et est-ce que le traitement cérébral lié au rythme est le même dans le langage et dans la musique ? Pour répondre à ces questions, une étude consistant à analyser les variations du potentiel électrique cérébral (ou Potentiels Evoquées) provoquées par des violations du mètre/rythme ou du sens dans le langage et dans la musique a été initiée. Dans la partie langage de l'expérience, des phrases dont le dernier mot était toujours trisyllabique ont été enregistrées. Pour simuler les violations métriques, l'accentuation qui est toujours sur la dernière syllabe en français a été déplacée de la dernière à l'avant-dernière syllabe. Pour ne pas modifier le timbre de la voix naturelle, un algorithme de time-stretching basé sur les travaux de G. Pallone [Pallone, 2003] a été utilisé. Une adaptation de la méthode initialement conçue pour des facteurs de dilatation temporelle de 4 a été nécessaire pour effectuer de dilatations de l'ordre de 100 (ce qui correspond aux variations de longueur des syllabes dans un mot terminal en français). Une attention particulière a été portée sur la segmentation des signaux, de façon à ne faire porter les allongements temporels que sur les parties voisées du signal (voyelles). Les incongruités sémantiques ont été produites en remplaçant un mot qui a un sens dans la phrase par un mot sémantiquement incongru (Le menuisier répare un tabouret/marabout). Dans la partie musique, un logiciel de synthèse de séquences musicales constituées de triolets basés sur des arpèges a été développé. Ces arpèges sont construits à partir d'accords dont la succession satisfait les règles classiques de contrepoint. Un contrôle précis de la modification locale du rythme et/ou du tempo permet l'introduction d'incongruités rythmiques. Nous avons choisi d'utiliser des triolets de façon à nous rapprocher de la situation retenue pour le langage où le dernier mot est toujours trisyllabique. Les incongruités mélodiques ont été construites en transposant le dernier triolet de façon à nous rapprocher à nouveau de la condition langage, c'est-à-dire, un triolet congruent du point de vue harmonique mais incongru dans le contexte mélodique. Un ensemble de 128 séquences sonores a été construit pour le langage et pour la musique. Le traitement numérique de ces séquences à permis l'obtention de 512 stimuli calibrés contenant des séquences incongrues sémantiquement/mélodiquement, incongrues métriquement/rythmiquement, et incongrues sémantiquement/mélodiquement et métriquement/rythmiquement [Ystad et al., 2003]. Ces stimuli ont été présentés à des sujets qui devaient focaliser leur attention soit sur le mètre/rythme, soit sur la sémantique/harmonie et déterminer si les phrases/mélodies étaient acceptables. Les résultats ont montré que des mots sémantiquement incongrus sont traités indépendamment de l'attention du sujet, suggérant que le traitement sémantique se fait de façon automatique. Les incongruités métriques (déplacement d'accent) semblent perturber le traitement sémantique et pose la question de la transparence des traitements audios de type "time-stretching". En ce qui concerne la musique, les analyses montrent que les incongruités rythmiques sont traitées indépendamment de l'attention, ce qui tend à montrer que le rythme est traité de façon automatique lorsqu'on écoute de la musique [Magne et al., 2007; Ystad et al., 2007].

3.3.3 Bilan de la rencontre

Le bilan des premières études associant modélisation sonore et sciences cognitives a été très positif, même si cette rencontre a renforcé notre conscience de la complexité liée à l'extraction d'indices perceptifs et cognitifs pour la synthèse. Dans le cadre de l'étude sur les sons de piano, les différentes écoutes en fonction du degré d'expertise et en fonction de la dégradation du signal ont été identifiées. Même si ces résultats sont extrêmement intéressants, la méthode d'expérimentation est très laborieuse et ne nous a pas permis d'étudier des situations plus complexes et réalistes mettant en jeu plus qu'une note isolée de piano. Pour répondre à la question initiale concernant la distinction entre un bon et un mauvais piano, il faudrait étendre l'étude à des situations plus complexes de jeu musical, faisant intervenir des aspects dynamiques. En effet, les études sur le timbre et l'interprétation décrites précédemment (section 3.2), montrent que l'appréciation d'un morceau de musique dépend fortement de l'évolution temporelle de la durée, du timbre et de l'intensité des notes ainsi que des genres musicaux. L'analyse doit donc prendre en compte le contexte musical et se focaliser sur l'appréciation de la cohérence entre notes au sein de séquences musicales ainsi que la cohérence des séquences au sein de l'oeuvre musicale, plutôt que sur des sons isolés extraits de tout contexte.

En ce qui concerne la liaison entre les neurosciences cognitives et la modélisation sonore, l'association entre transformations sonores et imagerie cérébrale nous a paru très prometteuse. En effet, pour l'acousticien qui s'intéresse à la modélisation sonore, les neurosciences cognitives apportent des éléments de réponse quant à la construction de modèles sonores pertinents du point de vue perceptif. Pour le cognitien, la modélisation sonore représente un grand intérêt par sa capacité de construire des stimuli parfaitement contrôlés et d'effectuer des transformations sonores en vue de l'étude des processus cérébraux impliqués dans la perception et la cognition auditive. Les premières études sur la perception du rythme dans le langage et la musique ont données un certain nombre de réponses quand aux différences de traitement cérébrale lors des violations de rythme et de sens dans ces deux systèmes d'expression. En ce qui concerne le lien entre ces résultats et les paramètres acoustique, le protocole expérimental a permis de vérifier que le sens peut être perturbé lorsqu'une dilatation trop importante est appliquée aux signaux. Néanmoins, ce protocole expérimental n'a pas permis de tirer des conclusions sur la pertinence perceptive des paramètres acoustiques à cause de la complexité des stimuli faisant intervenir un grand nombre de paramètres. Il nous a alors paru intéressant de définir des protocoles expérimentaux sur des sons parfaitement calibrés à l'aide de l'approche analyse-synhèse. Le grand défi consistait à proposer des études répondant aux problématiques des 2 domaines de recherche. Les réflexions nous ont amenés à proposer un projet de recherche intitulé "Vers le sens des sons", (acronyme senSons) autour de trois thèmes principaux de recherche : la structure des sons et leur perception, la sémiotique des sons et le contrôle haut niveau de processus de synthèse. Ce projet a été soutenu par l'Agence National de la Recherche dans le cadre du programme blanc "jeunes chercheuses et jeunes chercheurs" en 2005. Ce projet m'a permis d'exporter la synthèse vers des applications nouvelles dépassant largement le strict cadre musical.

4 Les sons environnementaux et industriels

L'extension de la synthèse vers de nouvelles applications visant à utiliser les sons comme porteurs d'information et/ou à proposer un contrôle du ressenti, nécessite une interaction forte entre la modélisation sonore et les sciences du cerveau. Le projet ANR "Vers le sens des sons" avait pour but d'aborder cette problématique en associant des partenaires académiques français (LMA, INCM), étrangers (ARI-Autriche, Université de McGill-Canada, Université de Aalborg-Danemark) et industriels (Peugeot Citroën, France Télécom). Les objectifs principaux du projet étaient de montrer qu'il existe une sémiotique des sons et qu'un lien peut être établi entre le comportement physique des sources sonores, la perception des sons engendrés et le sens évoqué par ces sons. Le projet était organisé suivant deux axes principaux dont les buts étaient de :

- trouver un lien entre la structure des sons et leur perception
- comprendre les mécanismes cérébraux responsables de l'attribution d'un sens à un son

Le projet reposait sur une approche pluridisciplinaire associant l'acoustique, le traitement du signal, la psychoacoustique et les neurosciences cognitives dont les complémentarités permettent aujourd'hui d'aborder cet aspect crucial de la communication sonore. Le projet était structuré suivant quatre axes principaux, à savoir l'analyse des signaux audio non-stationnaires, la synthèse, la perception et l'aspect cognitif lié à l'écoute des sons :

- L'analyse des sons fournit les outils et méthodes susceptibles de représenter un signal sonore à partir de concepts mathématiques (notamment de complétude et de conservation de l'information). Elle prend tout son intérêt dans le domaine des sons en lui associant des concepts sensoriels (sonie, timbre, sensations, . . .)

- La synthèse des sons permet la génération et la manipulation de sons calibrés en s'appuyant sur des modèles physiques, de signaux ou hybrides. Elle permet d'étudier l'effet perceptif des paramètres et de s'affranchir des contraintes physiques tout en donnant lieu à des sons réalistes.
- La perception des sons s'attache aux bases mécaniques et physiologiques propres à notre système auditif ainsi qu'aux attributs subjectifs sur lesquels s'appuie notre perception des sons. Ce domaine procure un ensemble classifié de méthodes basées sur le comportement et/ou le jugement de sujets (psycho-physique).
- Les neurosciences cognitives s'intéressent quant à elles aux mécanismes mis en jeu au niveau cérébral lors de l'écoute des sons et permettent d'élargir le champ d'étude par la prise en compte du contexte et des spécificités des sujets (degré de connaissance, spécificités culturelles, ...). Ce domaine procure en outre des méthodes de mesures objectives basées sur l'imagerie cérébrale.

Dans le cadre du projet senSons, ces domaines interagissaient fortement au sein d'un système bouclé qui est représenté dans la Figure 3 ci-dessous.

4.1 Partenaires

Initialement, “l'équipe senSons” regroupait sept personnes dont quatre permanents CNRS. Tous les membres de l'équipe étaient rattachés au Laboratoire de Mécanique et d'Acoustique (LMA) à l'exception d'un chercheur postdoctoral qui dépendait de l'Institut de Neurosciences Cognitives de la Méditerranée (INCM). Huit autres chercheurs participaient au projet comme collaborateurs. Parmi ces collaborateurs, deux d'entre eux étaient rattachés à l'INCM ; deux autres au LMA et un au Laboratoire d'Analyse, Topologie et Probabilités de Marseille (LATP). Deux institutions étrangères collaboraient au projet : l'Acoustics Research Institute (ARI), Vienne, Autriche et l'Université d'Aalborg, Esbjerg, Danemark. Par la suite, le projet a suscité un grand intérêt et plusieurs nouveaux collaborateurs relevant d'institutions Françaises et étrangères s'y sont associés. Tel est le cas du Centre de Réalité Virtuelle de Marseille-Luminy, de l'Université McGill, Montréal, Canada, de l'Université de Newcastle, Australie, ainsi que des entreprises Peugeot-Citröen (PCA) et Orange France-Télécom. Ces collaborations ont permis une ouverture vers de nouveaux domaines de recherche qui s'inscrivent naturellement dans la problématique senSons, à savoir la synthèse

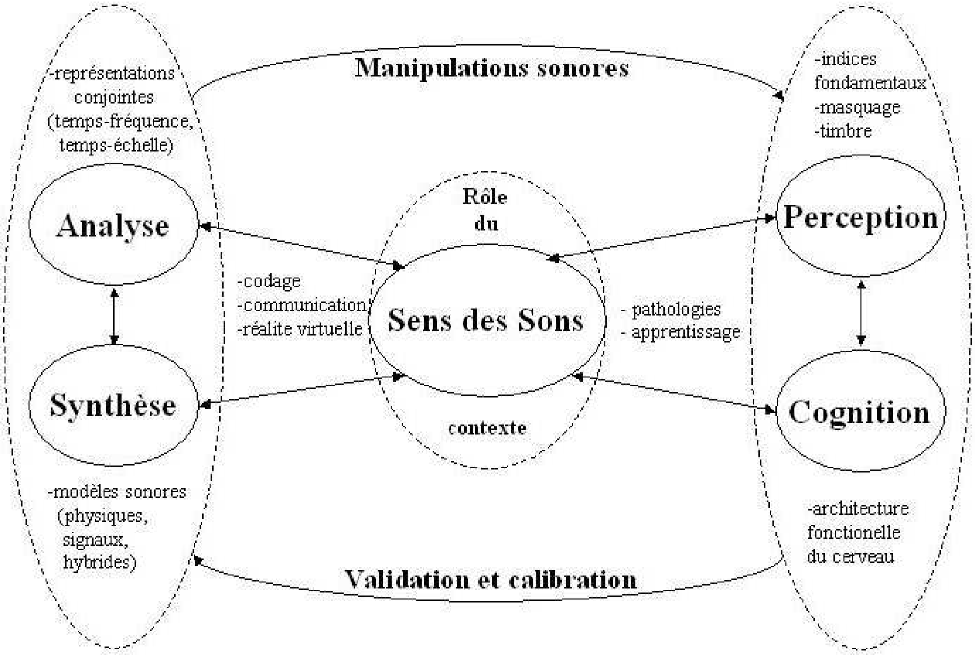


Figure 3: Synoptique, projet senSons

sonore 3D et la Réalité Virtuelle sonore.

4.2 La structure des sons et leur perception

Cet axe du projet visait à caractériser les éléments, tant acoustiques que perceptifs, qui interviennent dans la perception des sons. Cet axe a été abordé suivant 3 niveaux distincts de complexité des sons : “atomique”, sons isolés et structures complexes.

4.2.1 Au niveau atomique

Le domaine de l’analyse des signaux non-stationnaire a connu des avancées importantes durant les années 1980 grâce au développement de méthodes d’analyse donnant lieu à des représentations conjointes de type temps-fréquence ou temps-échelle [Kronland-Martinet et al., 1987; Flandrin, 1993]. Ces représentations sont obtenues par décomposition d’un signal en une somme de fonctions élémentaires ayant de bonnes propriétés de localisation en temps et en fréquence. Aujourd'hui les représentations conjointes de signaux sont disponibles dans la majorité des logiciels destinés au traitement de signaux sonores, et parmi ces représentations, le spectrogramme, obtenu par analyse de Fourier à court terme, est probablement la plus connue. Même si les représentations conjointes constituent des outils indispensables pour identifier et extraire des paramètres à partir d'un signal, elles ne reflètent pas toujours la pertinence perceptive des paramètres et pourraient être optimisées à partir de critères perceptifs. Un phénomène perceptif qui permet de réduire considérablement la quantité de données est le phénomène du masquage (qui définit le degré auquel l'audibilité d'un son est dégradée par la présence d'un ou plusieurs autres sons). Ce phénomène a été largement étudié dans la littérature et est actuellement utilisé dans plusieurs applications tel le codage MPEG-1 Audio Layer 3 (MP3)[Pan, 1995].

Jusqu'à présent l'essentiel des travaux sur le masquage était basé sur l'étude du phénomène dans le domaine fréquentiel [Moore, 2003; Bacon and Viemeister, 1985]. Des travaux basés sur le masquage temporel complètent ces derniers [Fastl, 1979; Widin and Viemeister, 1979], mais très peu étudient simultanément l'interaction des phénomènes fréquentiels et temporels du masquage. Cette approche conjointe est pourtant primordiale, lorsqu'il s'agit de développer des représentations de signaux sonores cohérentes avec la perception. Un tel problème, de nature fondamentale, nécessite une approche pluridisciplinaire. Il s'agit en effet de développer et d'évaluer de nouvelles représentations des signaux audio, de développer de nouveaux outils basés sur les théories mathématiques des représentations temps-fréquence (ou temps-échelle) et d'adapter ces représentations à la perception audio et aux concepts de la psychoacoustique. L'approche proposée ici est basée sur la construction mathématique d'outils de "masquage" basés sur la théorie des multiplicateurs de Gabor et vise à répondre à des questions fondamentales telles que : les effets de masquage sont-ils essentiellement de type additif (concept utilisé dans le codage MP3) ? Comment se caractérise le phénomène de masquage dans le repère temps-fréquence ? Quelles sont les bases auditives (mécaniques et neurales) qui régissent le masquage sonore dans les deux domaines ? Comment peut-on minimiser le nombre de coefficients d'une représentation temps-fréquence tout en assurant une reconstruction perceptivement identique du signal original ? Les réponses à ces questions permettent *in fine* le développement d'un "filtre temps-fréquence de masquage" formalisable du point de vue mathématique en utilisant le concept de multiplicateurs de Gabor.

Deux équipes de recherche du LMA (Psychoacoustique et Modélisation, Synthèse et Contrôle de Signaux Sonores et Musicaux) et une équipe de l'institution ARI - "Acoustics Research Institute" ont travaillé autour de ce projet. J'ai dans ce cadre obtenu un soutien dans le cadre des Programme d'Actions Intégrées (PAI) - AMADEUS, intitulé Représentations temps-fréquence et perception des sons durant la période 2006 à 2008. Ce projet fait l'objet de la thèse de Thibaud Necciari.

Durant la période du projet, 2 expériences psychoacoustiques de masquage temps-fréquence ont été menées. Dans les 2 cas, les signaux utilisés ont été choisis de façon à répondre à une double contrainte : avoir une localisation optimale à la fois en temps et en fréquence et ne produire l'activation que d'une des nombreuses fenêtres d'observation spectro-temporelles du système auditif. Les "atomes" temps fréquence qui répondent au mieux à ces contraintes, sont des sinusoïdes modulées par des Gaussiennes d'une largeur fréquentielle de 600Hz pour une modulation de 4kHz, et une durée "rectangulaire équivalente" de 1.7 ms.

Dans la première expérience, qui a été menée au LMA, un masque centré sur 4000Hz à 60dB SL (i.e. 60dB au-dessus du seuil absolu du sujet) a été utilisé pour étudier l'effet du masquage dans le plan temps-fréquence. 11 séparations de fréquence (allant de -4ERBs à 6ERBs relatives à 4000Hz) et 5 distances temporelles (0, 5, 10, 20 et 30ms) entre cible et masque ont été testées. Les résultats de cette expérience montrent que les courbes de masquage fréquentiel sont conformes aux courbes issues d'expériences traditionnelles de masquage, et mettent en évidence le fait que le masquage s'étend beaucoup plus sur la partie haute fréquence du masque, c'est à dire pour les fréquences cibles supérieures à la fréquence du masque. La variabilité inter individuelle est également plus grande quand les fréquences cibles sont supérieures à la fréquence du masque. En ce qui concerne le masquage temporel, il décroît rapidement avec l'augmentation de la séparation temporelle et devient négligeable à partir de 30ms. Cette décroissance est plus marquée que celle décrite dans la littérature, ce qui peut s'expliquer par la durée du masque (9.6ms) qui est largement inférieure à celle des expériences traditionnelles (de l'ordre de 300ms). La variabilité inter individuelle est importante pour une séparation temporelle de 5ms, probablement due au fait que masque et cible dans ce cas se chevauchent partiellement introduisant des effets de phase (modulation) utilisés comme indice par certains sujets. Les résultats de la condition temps-fréquence (décalages en temps et en fréquence) ont montré que le masquage temps-fréquence ne pouvait pas être déduit en combinant les résultats issus des expériences de masquage fréquentiel et temporel, suggérant une activité complexe de masquage dans le plan temps-fréquence (Figure 4) [Necciari et al., 2008, 2010].

Dans la deuxième expérience, qui a été menée à Vienne (ARI), l'additivité du masquage dans le plan temps-fréquence a été étudiée en utilisant

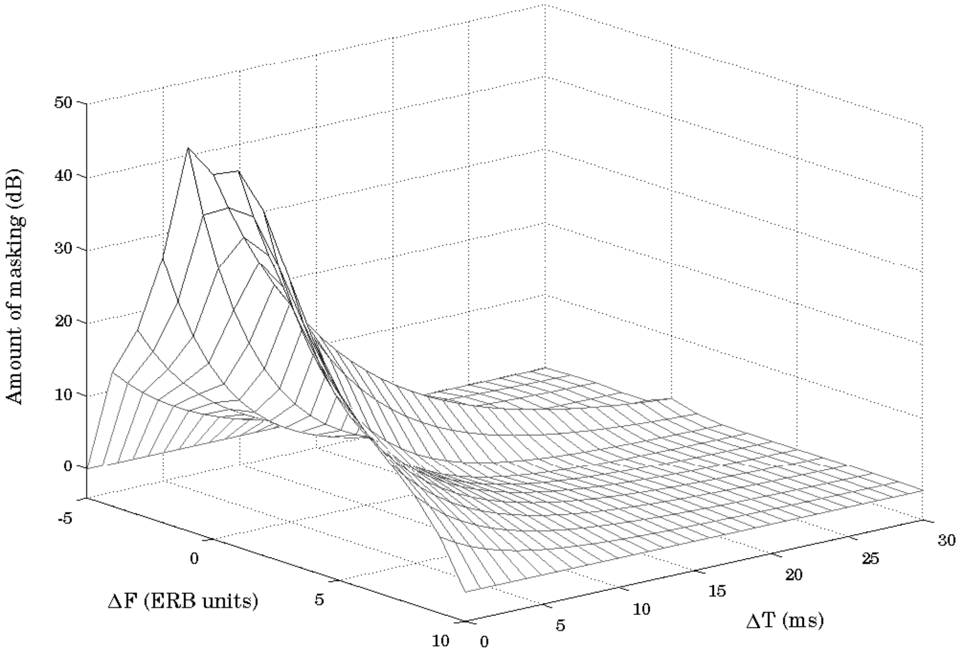


FIGURE 4: Valeur moyenne de masquage (en dB) produit par un masque d'enveloppe Gaussienne en fonction du temps ΔT (en ms) et de fréquence ΔF (en unités ERB). La figure est extrait de la thèse de Thibaud Necciari.

jusqu'à 4 masques d'enveloppes gaussiennes (même signaux que dans expérience 1).

Dans le cas du masquage temporel, les 4 masques sont séparés de la cible de -24ms (M1), -16ms (M2), -8ms (M3), et 8ms (M4). Le seuil du masquage a été mesuré pour 5 combinaisons de masques ; M2-M3, M3-M4, M1-M2-M3, M2-M3-M4 et M1-M2-M3-M4. Les résultats montrent qu'il n'y a pas de différence systématique liée à la combinaison des masques, et que le niveau de masquage augmente avec le nombre de masques. Le masquage est plus important que pour le modèle linéaire additif et dans la condition ou on a 4 masques, le niveau de masque atteint jusqu'à 26dB de plus que le modèle linéaire. L'effet non-linéaire observé dans cette condition, montre que les Gaussiennes utilisées dans cet expérience font l'objet d'une forte compression de la membrane basilaire.

Dans le cas du masquage fréquentiel, la cible avait une fréquence de 5611Hz et les masques étaient séparés de la cible par -7 ERBs (M1), -5 ERBs (M2), -3 ERBs (M3) et +3 ERBs (M4). Contrairement au masquage temporel, les combinaisons de masques ont une influence sur le niveau du masquage. Ainsi, les masques 2-3 contribuent peu au masquage tandis que la combinaison des masques 3-4 contribue beaucoup. En rajoutant le masque M1 à la paire 2-3, on augmente le niveau de masquage, ce qui n'est pas le cas quand on rajoute M2 à la paire 3-4. Pour certaines paires de masques le phénomène est non-linéaire, pour d'autres non. Les résultats sont cohérents avec la littérature et montrent que les effets non-linéaires du masquage interviennent quand on ajoute un masque qui n'a pas de chevauchement d'énergie avec au moins un des masques avec lequel il est présenté [Laback et al., 2008].

Malgré le comportement non linéaire des phénomènes de masquage, la prise en compte des effets de masquage d'un grain temps-fréquence prédominant semble cohérent avec notre perception des sons. Ainsi, les résultats issus de ces tests d'écoute sont actuellement utilisés pour développer un "modèle temps-fréquence de masquage" dans le but d'optimiser les représentations temps-fréquence en les rapprochant d'un point de vue plus perceptif.

4.2.2 Au niveau de sons isolés

La simulation de sons à partir du ressenti (décrit verbalement, par une scène visuelle ou autre) nécessite l'identification de paramètres ou structures morphologiques du signal responsables de la représentation mentale que l'on se fait lorsqu'on écoute un son. Plus précisément on cherche le "squelette" ou l'invariant du son qui permet la reconnaissance du phénomène sonore indépendamment de la situation d'écoute, de la source etc. Cette structure ou consistance timbrale expliquerait pourquoi on reconnaît par exemple un son métallique, même s'il provient de sources très variées (plaques, cloches, barres, de grandes ou petites dimensions) ou la voix d'un proche indépendamment des émotions exprimées et même lorsqu'elle est déformée (par le téléphone par exemple). Dans ce cadre, des sons isolés ont été étudiés dans le but d'identifier les paramètres du signal responsables de leur catégorisation perceptive. A travers des protocoles expérimentaux, la reconnaissance de matériaux et de mouvements ont été étudiés. Le lien entre qualité sonore et paramètres du signal a également été abordé par le biais de bruits de portières automobile et de sons issus de différents espèces de bois jugés par un luthier de xylophone.

Catégorisation perceptive de sons d'impact

Dans cette étude, nous nous sommes intéressés aux évocations de matériaux en vue de la construction d'un synthétiseur de sons d'impacts permettant de contrôler les sons à partir de mots décrivant différentes catégories de matériaux (Bois, Métal, Verre, ...). Cette étude, qui a fait l'objet d'une collaboration étroite avec Mitsuko Aramaki et Mireille Besson à l'Institut de Neurosciences Cognitives de la Méditerranée, nous a amené à effectuer plusieurs expériences axées sous différents angles d'investigation. Les résultats sont présentés dans cette section ainsi que dans les sections (4.3.3) et (4.4.1).

Des sons d'impact issus de différents matériaux frappés (Bois, Métal, Verre) ont été enregistrés et analysés. Une méthode de synthèse additive a été utilisée pour resynthétiser les sons. Pour minimiser des variations éventuelles de timbre induites par des différences de hauteur entre sons, tous les sons ont été égalisés (au même chroma (Do)). Les sons ont également été égalisés en sonie. En interpolant les paramètres de synthèse (amplitudes et coefficients d'amortissement), des continua entre les trois catégories de matériaux (bois-verre, bois-métal et verre-métal) ont été construits. Cette méthode a permis de simuler des transitions progressives entre catégories de matériaux. 5 continua de 20 sons intermédiaires ont été construits pour chaque transition donnant au total 330 sons. Ces sons ont été présentés dans un ordre aléatoire à des sujets qui devaient les classer le plus rapidement possible dans une des trois catégories de matériaux à l'aide d'un boîtier comportant trois boutons de réponse. 22 participants (droitiers et non-musiciens) entre 19 et 35 ans ont été testés dans l'expérience. Des mesures éléctrophysiologiques (Electroencephalogramme, EEG) ont été enregistrées en continu durant l'expérience à l'aide de 32 électrodes localisées sur le salp. Le résultats issus de ces mesures sont décrites dans la section (4.3.3). Les résultats comportementaux correspondant à la catégorisation et au temps de réaction des sujets, ont permis de mettre en évidence une zone de continuum entre catégories de matériau pour chaque transition et de distinguer des sons typiques (classés dans une même catégorie par plus de 70 % de sujets) et des sons ambigus. Cette distinction entre sons typiques et sons ambigus a été utilisée ultérieurement lors d'une expérience d'amorçage de sons environnementaux décrite dans la section (4.3.2). Les résultats montrent que les participants ont plus souvent classé les sons dans la catégorie métal que dans les catégories verre et bois et que les temps de réactions étaient plus grandes pour des sons issus de la catégorie verre que pour ceux issus des catégories bois et métal [Aramaki et al., 2009a].

Pour établir un lien entre paramètres acoustique et catégories issus des tests catégorisation, des corrélations entre catégories de sons typiques et descripteurs acoustiques ont été recherchées. Parmi les descripteurs qui se sont déjà révélés pertinents pour la perception de timbre et l'identification de matériaux, nous avons ici étudié le temps d'attaque (AT), le centre de gravité spectral (CGS), la largeur de bande (SB), la rugosité (R) et l'amortissement normalisé du son (α). Une approche statistique basée sur une méthode de régression logistique binaire a permis de proposer des modèles prédictifs pour chaque catégorie de matériaux. Les résultats de ces analyses ont révélés que α est le facteur prédictif principal de la catégorie bois (85%) et de la catégorie métal (82.7%), mais que des descripteurs spectraux, à savoir le centre de gravité spectral et la largeur de bande pour la catégorie bois et la largeur de bande pour le métal sont également importants pour la catégorisation. En ce qui concerne la catégorie verre, tous les descripteurs testés sauf le temps d'attaque se sont révélés importants et la catégorie verre ne pouvait pas être prédite par seulement un ou deux paramètres. Contrairement aux catégories bois et métal, la catégorie verre est mieux décrite par les propriétés spectrales que par l'amortissement.

Pour définir un espace de contrôle de matériaux permettant un contrôle intuitif des sons, une analyse statistique basées sur une ACP (Analyse en Composantes Principales) a été effectuée. Cette analyse a permis de représenter les trois catégories de sons dans un espace bidimensionnel. En calculant la corrélation entre dimensions et descripteurs acoustiques, on a pu déduire que l'amortissement est responsable de la distinction entre la catégorie Bois et la catégorie Métal et que la rugosité est responsable de la distinction de la catégorie verre. Ces résultats sont exploités dans le synthétiseur de matériaux frappés décrit dans la section (4.4.1) [Aramaki et al., 2009b].

Qualité sonore des bois de luthérie

Les sons d'impact ont également été utilisés dans le cadre d'une étude visant à lier la qualité sonore et les caractéristiques physiques des bois de lutherie. Cette étude a été effectuée en collaboration avec le CIRAD, Montpellier (L. Brancheriau et H. Baillères). Une méthodologie originale a été proposée, associant des processus d'analyse-synthèse à des tests de classification perceptive. Les sons produits en impactant 59 lames de bois issues d'espèces différentes mais toutes de même géométrie, ont été enregistrés et distribués de façon aléatoire sur un écran d'ordinateur sous forme d'icônes simples. Un luthier de renom devait ensuite réorganiser ces sons le long d'un l'axe horizontal en fonction de son jugement de la qualité musicale du son. Les sons jugés de bonne qualité ont été classés à droite et les sons les moins appréciés à gauche. Le luthier pouvait écouter les sons autant de fois qu'il le souhaitait. Pour extraire des descripteurs de timbre responsables de la classification du luthier, les sons ont été resynthétisés, égalisés en pitch et classés de nouveau par le luthier qui ne savait pas que les sons étaient synthétiques.

Des corrélations entre descripteurs acoustiques, les propriétés mécaniques et anatomiques du bois et le classement du luthier ont été obtenus par régression linéaire multivariée [Dillon and Goldstein, 1984]. Les résultats montrent que l'amortissement de la première composante spectrale ainsi que l'étendue spectrale sont les principaux descripteurs acoustiques responsables de la classification. Ce résultat indique que le luthier recherche des sons très résonnants et cristallins. En ce qui concerne les caractéristiques mécaniques du bois, le coefficient de friction interne semble être le principal paramètre corrélé à la classification du luthier [Aramaki et al., 2007].

Cette étude a permis d'identifier les principaux critères perceptifs utilisés par un luthier lors de la classification de différents espèces de bois. Ces résultats donnent des pistes quand aux choix de nouvelles espèces de bois utilisables pour la facture instrumentale et ont permis d'induire des projets de recherches ambitieux, portés par le CIRAD et axés sur la modification génétique des arbres en vue de l'optimisation de leurs propriétés "sonores". L'identification de descripteurs acoustiques corrélés au classement est également importante pour la synthèse et le design de nouveaux instruments numériques.

Sons de fermeture de portières automobiles

Une troisième étude a été effectuée sur des sons d'impacts d'une catégorie bien particulière à savoir les sons de fermeture de portières automobiles. Cette étude a fait l'objet d'une collaboration entre CNRS-LMA et le PSA Peugeot Citröen dans le cadre de la thèse (CIFRE) de Marie-Céline Bezat que j'ai co-encadré. Du point de vue industriel, l'objectif de cette thèse était de mieux corréler les attendes des clients par rapport aux sons de fermeture de portières automobiles et d'utiliser ces connaissances dans le processus de conception en dimensionnant correctement certaines pièces mécaniques constituant les portières. Ce travail nécessite à la fois de connaître le jugement perceptif des bruits de portières et d'expliquer ce jugement par des propriétés acoustiques et mécaniques des sons. Pour aborder ce problème, des études perceptives sur des sons réels issus de différentes marques et gammes automobiles ont été effectuées. Puis une approche par analyse-synthèse a été adoptée afin de tester des stimuli parfaitement calibrés et lier les paramètres du signal au ressenti.

Le jugement perceptif a été étudié sur plusieurs plans. Les sujets ont jugé les sons en situation réelle, en manipulant des portières sur des véhicules de différentes marques et gammes. Puis les sons ont été enregistrés en chambre semi anéchoïque à l'aide d'une tête acoustique pour différentes vitesses de fermeture et présentés aux sujets en laboratoire. Et enfin, une série de vidéos présentant un expérimentateur manipulant des portières de véhicule différentes a été enregistrée pour étudier l'influence de l'image sur le ressenti. Les mêmes marques et gammes de véhicules ont été utilisées tout au long des expériences. Ainsi les évaluations de qualité de véhicule que renvoient ces bruits ont pu être comparées. En confrontant les résultats en laboratoire avec sons uniquement et les résultats IN SITU ou les sujets sont face au véhicule, on observe que l'image du véhicule (surtout la gamme) influence largement l'évaluation de la qualité avec une meilleure évaluation pour les véhicules qui ont une meilleure image et vice versa. Pour analyser plus avant l'influence de l'image du véhicule, les sujets ont ensuite évalué des associations bruits - véhicules obtenus par des montages vidéo. La comparaison des résultats avec ceux obtenus lors d'évaluations avec des sons seuls montre que l'image du véhicule n'influence que légèrement l'évaluation de la qualité que renvoie le bruit de fermeture, mais que l'influence observée sur véhicule réel est bien supérieure. L'introduction du facteur image au laboratoire n'est donc pas représentative du facteur réel. Dans le cadre de l'évaluation de sons au laboratoire, les résultats ont montrés que l'augmentation de la vitesse de fermeture pénalise l'évaluation de la qualité que renvoie le bruit de fermeture. En situation réelle, le comportement des sujets a été filmé et les sujets ont été interrogés sur les évocations que renvoient ces bruits. Cette approche a permis d'identifier trois profils de sujets, à savoir les inquiets, les conviviaux et les esthètes. Trois types d'évocation ont également été identifiés, à savoir la solidité/sécurité (chez les inquiets), le confort (chez les conviviaux) et la distinction (chez les esthètes).

Dans cette étude, 2 types d'écoute ont été distingués selon la définition de Gaver (1993), à savoir l'écoute analytique ("musical listening") et l'écoute écologique ("everyday listening"). L'écoute analytique est caractérisée par une écoute des propriétés intrinsèques du son décrite par les descripteurs de timbre, les descripteurs acoustiques, la durée, etc (paramètres du signal, paramètres sensoriels). L'écoute écologique est caractérisée par les propriétés naturelles qui font référence à la source et aux évocations que renvoie le son (solidité, qualité). Plusieurs tests ont été effectués pour identifier les propriétés analytiques et naturelles impliqués dans l'écoute des bruits de portières.

En ce qui concerne l'écoute analytique, une analyse sensorielle a été effectuée pour caractériser les sons du point de vue sensoriel. Cette approche consiste à entraîner un groupe de sujets à caractériser des sons par quelques composantes perceptives à travers des mots ou d'onomatopées. Dans le cadre de bruits de portières, les "caractéristiques" KE, BOMN, INTENSE, ESPACHOC, Gri-gri, Triangle ont été identifiées par le panel. En ce qui concerne l'écoute écologique, les propriétés naturelles liées à la qualité, la solidité, l'énergie de fermeture, le poids de la porte ont été évalués.

Des experts PSA des ouvrant automobiles ont également évalué les propriétés naturelles en sources organiques : présence serrure, crans de serrure, vibrations de panneaux de porte. Enfin, un lien entre propriétés perceptives a été établi pour construire un réseau perceptif caractéristique du bruit de fermeture de porte.

Pour associer paramètres du signal et évocations liés aux bruits de portières, une approche analyse-synthèse a été adoptée. Les propriétés analytiques évoquées lors des tests perceptifs sont de bons indicateurs concernant l'information perceptivement importante contenue dans le signal. Notamment la caractérisation de propriétés naturelles données par les experts ont permis de mettre en avant l'aspect serrure/fermeture du bruit. Ces deux aspects ne sont pas faciles à distinguer à partir d'analyses traditionnelles du signal. Une méthode modale empirique, "Empirical Mode Decomposition (EMD) introduite par [Huang et al., 1998] et étudiée par [Flandrin et al., 2004] a permis de séparer ces deux sources de façon satisfaisante du point de vue perceptif. Le principe de cette méthode est d'identifier itérativement des modes intrinsèques du signal modulés à la fois en amplitude et en fréquence, en séparant localement (à l'échelle d'une oscillation) une contribution "rapide" d'une tendance plus "lente". Le signal est parfaitement reconstruit par simple addition des modes. Ainsi, le bruit "fermeture" est obtenue à partir des 6 premiers modes et le bruit "serrure" à partir des modes d'ordre supérieurs. La resynthèse des 2 types de bruits a été effectuée à partir d'un modèle de synthèse de type additif développé par Aramaki [Aramaki and Kronland-Martinet, 2006; Aramaki et al., 2006] simulant la répartition spectrale initiale et les lois d'amortissement exponentielles dans 40 bandes de filtres dont la largeur correspond aux bandes critiques de l'oreille. Le bruit "serrure" est simulé par 3 impacts et le bruit fermeture est simulé par un impact. Le synthétiseur a été implémenté en temps-réel à l'aide du logiciel Max/MSP et contrôlé par 28 paramètres indépendants. Ce modèle a permis la génération de stimuli parfaitement calibrés et proches de sons réels. Des tests perceptifs ont été effectués sur les sons de synthèse pour valider sa qualité et lier les paramètres du signal aux évocations. Les résultats sont illustrés dans la Figure 5 ci-dessous .

La figure montre que la propriété BONM est obtenue par une faible présence de la partie "serrure" et une présence plus forte de la partie "fermeture". Inversement, la propriété KE est obtenue par une forte présence de la partie "serrure" par rapport à celui de la fermeture. BONM et KE sont à leur tour responsables des évocations de poids de porte qui ensuite influence l'image de solidité du véhicule. Les paramètres ("bas niveau") du signal qui sont responsables d'évocations de solidité et qualité sont en partie

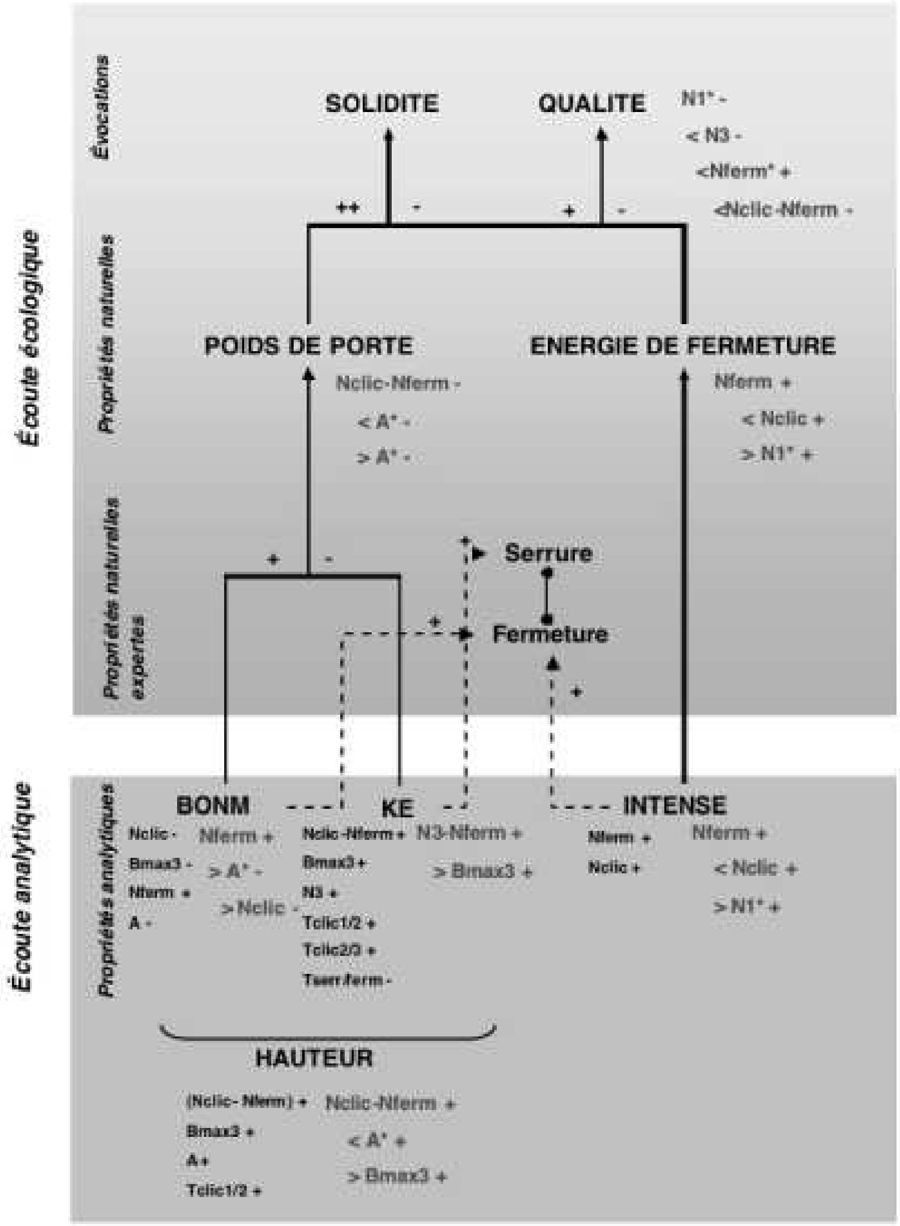


FIGURE 5: La figure montre que la propriété BONM est obtenue par une faible présence de la partie “serrure” et une présence plus forte de la partie “fermeture”. Inversement, la propriété KE est obtenue par une forte présence de la partie “serrure” par rapport à celui de la fermeture. BONM et KE sont à leur tour responsables des évocations de poids de porte qui ensuite influence l’image de solidité du véhicule.

liés aux niveaux des trois impacts du bruit (N1, N2 et N3), de durées inter impacts (Tclic3-Tferm, Tclic3-Tclic2, ...), de sous bandes ERB contenants le maximum d'énergie (Bmax2, Bmax3, ...), du niveau initial de l'impact du bruit "fermeture" (Nferm) et de la lois d'amortissement de l'impact du bruit "fermeture" (A).

Cette étude a permis de montrer que les descripteurs acoustiques proposés permettent de prédire de façon robuste les propriétés analytiques de notre ensemble de sons réels. De plus, afin de mieux comprendre les relations complexes entre ces descripteurs et les propriétés perceptives, les propriétés perceptives peuvent être caractérisées par des arbres de paramètres interdépendants. L'organisation des paramètres permet d'accéder aux critères acoustiques les plus influents d'un bruit particulier en fonction de sa caractérisation acoustique [Bezat et al., 2006; Bezat, 2007; Bezat et al., 2007, 2008].

Dynamique évoquée par les sons

L'étude présentée ci-dessous, concerne l'évocation de la dynamique (ou mouvement au sens du chef d'orchestre) induite par des sons. La notion de dynamique est floue et ne repose pas forcément sur des caractéristiques physiques liées aux sources sonores, ni sur une gestuelle, mais plutôt sur une morphologie sonore complexe. L'utilisation de sons environnementaux pose dans ce genre de problème la question de l'influence de nos acquis cognitifs. En effet, l'association d'une source physique à un son, guide fatalement le ressenti vers une représentation pragmatique du phénomène. Ainsi, personne n'associera au bruit d'une automobile une dynamique de type "s'élève", tant la source physique est bien repérée. Les écoutes analytiques et écologiques évoquées dans le cadre de bruits de portières automobile nous ont, elles aussi, amené à aborder la question du choix des corpus sonores susceptibles de favoriser un type d'écoute. Cette problématique n'est pas nouvelle et se retrouve en dehors du champ scientifique. En effet, depuis les années 1940, des compositeurs et des bruiteurs ont utilisé des sons nouveaux pour favoriser certains types d'écoutes. Différents courants de musique consistant à travailler directement sur le signal sonore ont vu le jour à cette époque, tels la "musique concrète" basée sur l'association de sons enregistrés ou synthétique (Pierre Schaeffer, 1910-1995) et "Elektronische Musik" basée sur l'utilisation de générateurs de signaux et des sons synthétiques (Karlheinz Stockhausen, 1928-2007). Dans son livre "Traité des objets Musicaux" de 1966 [Schaeffer, 1966], Pierre Schaeffer fait également mention de l'écoute acousmatique qui correspond à l'écoute d'un son sans se référer à la cause qui l'a produit. Ce concept avait déjà été utilisé par Pythagore (VIe siècle avant J.C.) qui employait le terme acousmatique pour désigner une situation d'écoute attentive obtenue alors qu'il se plaçait derrière un rideau pour enseigner à ses disciples, dans le noir, et dans un environnement totalement silencieux. Ce type de point de vue peut-il nous guider dans la construction de protocoles expérimentaux visant à mieux comprendre notre perception des sons ? Y-a-t-il des sons qui favorisent des types d'écoute particuliers ? La théorie de Schaeffer s'avère extrêmement intéressante dans ce contexte, et ce pour plusieurs raisons. Tout d'abord du point de vue conceptuel, Schaeffer introduit le terme "objet sonore" qui définit les sons qui une fois enregistrés ont une existence propre sans forcément de lien avec le "corps sonore", c'est à dire l'objet ou le phénomène à l'origine du son. Cette notion débouche ensuite sur une typologie apte à catégoriser l'ensemble des sons en fonction de leur profil spectral ("masse") et de leur dynamique ("facture"). Par son aspect général, cette classification est extrêmement intéressante car elle considère le son non plus comme une entité acoustique, mais comme une structure complexe dotée d'une "morphologie" bien caractérisée. De fait, elle sert de base à la construction de corpus sonores représentatifs de l'ensemble des morphologies sonores rencontrées dans la nature, sans pour autant recourir à des sons environnementaux bien repérés. La typologie de Schaeffer est constituée de 28 catégories de sons dont neuf centrales, définissant les "sons équilibrés". Ces neuf catégories constituent des combinaisons entre trois profils de "facture" (sons entretenues, impulsifs, itératifs) et trois profiles de "masse" (hauteur tonale définie, complexe et variable).

L'étude sur la dynamique des sons s'inspire de la théorie de P. Schaeffer et utilise des sons abstraits qui favorisent une écoute analytique plus apte à l'identification des grandeurs acoustiques responsables d'évocations de dynamique. Les sons utilisés dans cette expérience ont été obtenus à partir de banques de données provenant de compositeurs de musique électroacoustique du conservatoire de Marseille et sont représentatifs des neuf catégories de sons équilibrés de la typologie de Schaeffer. Ce travail fait l'objet de la thèse d'Adrien Merer que je co-encadre actuellement, et d'une collaboration avec l'INCM (M. Aramaki, M. Besson, J.-L. Velay). Les objectifs principaux de l'étude sont de définir des catégories de mouvement, d'identifier des invariants (ou descripteurs) représentatifs des catégories et in fine, d'utiliser ces invariants pour contrôler les évocations de mouvement à travers un synthétiseur numérique (voir section 4.4.2).

Dans le but de caractériser les catégories principales de dynamique sonore, deux tests de catégorisation ont été réalisés. Dans le premier test, 26 participants étaient confrontés à un problème de catégorisation libre de sons monophoniques. 68 sons ont été repartis de façon aléatoire sur un écran d'ordinateur et les participants devaient regrouper les sons qui évoquaient le même mouvement. Ils pouvaient faire autant de catégories qu'ils le souhaitaient et pouvaient écouter les sons autant de fois qu'ils le désiraient. A la fin du test, les participants étaient invités à décrire le mouvement qu'ils associaient à chaque catégorie. L'analyse des résultats a permis de définir 5 catégories principales de mouvement, à savoir : tourne, tombe, approche, passe et monte. Un autre résultat intéressant issu de ce test de catégorisation concerne la description des sons par les sujets. En effet, certains sujets ont choisi de dessiner la forme de la dynamique évoquée par les sons. Il semblerait ainsi que la relation entre la dynamique propre aux sons et une représentation graphique soit naturelle. Cette constatation sera reprise dans la section 4.4.2 où la construction d'un synthétiseur permettant un contrôle haut niveau de la dynamique ressentie est décrite.

Dans le deuxième test, 16 participants (qui ont tous participé au premier test) devaient classer ces mêmes stimuli dans 5 catégories prédéfinies et choisies à partir des résultats du premier test. Les sons étaient repartis de façon aléatoire en bas de l'écran, et le haut de l'écran était divisé en 5 secteurs. Les sujets devaient glisser les sons dans un des 5 secteurs en se basant sur la proximité perceptive (au sens de la dynamique ressentie) du son considéré et de sons prototypiques qui servaient de référence à chaque secteur. Les résultats obtenus ont permis de confirmer les catégories issues du premier test. La majorité des participants ont omis de classer certains sons, mais 62% des sujets ont considéré le nombre de catégories suffisant. Nous avons considéré comme typiques d'une catégorie les sons sélectionnés par plus de 70% des sujets dans une même catégorie. Ainsi, 9 sons ont été sélectionnés dans la catégorie tourne, 5 dans les catégories tombe et passe et 2 dans la catégorie monte. Aucun son n'a été sélectionné dans la catégorie approche. L'analyse acoustique des sons de chaque catégorie a permis de mettre en évidence des descripteurs acoustiques caractéristiques de chaque catégorie. Même si ces résultats sont sujet à caution (faible nombre de données, complexité des sons), on peut mentionner la présence systématique de modulations d'amplitude et/ou de fréquence pour la catégorie tourne, la décroissance de type logarithmique de l'enveloppe d'amplitude du son pour la catégorie passe et la nature impulsive de certains sons de la catégorie tombe [Merer et al., 2007, 2008a].

Contrairement aux sons environnementaux ou musicaux pour lesquels la cause est facilement reconnaissable, les sons abstraits permettent plus facilement des évocations multiples, puisqu'ils ne sont pas associés à une source physique précise. En d'autres termes, ces sons peuvent être évalués dans des contextes différents pour étudier des aspects spécifiques liés aux informations contenues dans les sons. C'est ainsi que nous avons utilisé le même corpus sonore dans deux études différentes liées à la vérification d'une existence d'une sémiotique sonore (voir section 4.3.2) [Ystad et al., 2008] et au jugement du bizarre et du familier chez les patients atteint de schizophrénie [Micoulaud-Franchi, 2009].

4.2.3 Au niveau structures complexes

A travers l'étude des sons abstraits, on a montré qu'un son aussi bref soit-il, peut évoquer un mouvement par sa dynamique intrinsèque. Qu'en est-il au niveau de structures complexes ? Comment les événements interagissent-ils pour que le message sonore soit cohérent avec une situation donnée ? L'étude sur l'interprétation décrite dans section 3.2 aborde cette question du point de vue musical. Cette question est également d'actualité pour le design sonore lorsqu'il s'agit d'améliorer l'ambiance sonore ou augmenter le réalisme d'une scène virtuelle. Une étude concernant la dynamique de bruits de moteurs automobile a été réalisée dans le cadre d'une convention d'étude entre PSA Peugeot Citroën et CNRS-LMA. Ce projet fait l'objet de la thèse (CIFRE) de Jean-François Sciabica que je co-encadre. Il s'agit de mieux comprendre le lien entre le ressenti de la dynamique du véhicule et la dynamique du bruit moteur lors des phases d'accélération. Une meilleure compréhension de l'adéquation entre l'évolution dynamique des descripteurs acoustiques et le ressenti perçu devrait à la fois permettre aux constructeurs d'adapter les réglages du moteur pour optimiser la maîtrise du véhicule et le retour sensoriel de qualité, mais aussi de donner des indications précieuses pour le design de nouveaux sons pour des véhicules hybrides et/ou électriques.

Le bruit perçu à l'intérieur d'un habitacle automobile est le résultat de l'interaction entre trois sources sonores, à savoir le bruit moteur, le bruit aérodynamique et le bruit de roulement. Le bruit moteur est fortement lié aux phénomènes de combustion qui produisent des sons périodiques dont la fréquence est corrélée aux cycles moteur. Le bruit de roulement et le bruit aérodynamique sont deux bruits large bandes, le bruit de roulement étant plus présent à basse vitesse tandis que le bruit aérodynamique augmente avec la vitesse. Lors de l'analyse temps-fréquence du son produit par une automobile en phase d'accélération, on observe la présence de plusieurs harmoniques dans le signal, certaines d'entre elles semblent être altérées par le bruit. Néanmoins, les représentations temps-fréquence ne permettent pas d'identifier les phénomènes perceptifs qui interviennent dans ces signaux complexes. Ces phénomènes sont de deux natures, le masquage des harmoniques moteur par le bruit aérodynamique et de roulement, et le masquage des harmoniques moteur entre elles. Un modèle auditif a été appliqué à ces bruits pour mettre en évidence ces effets. Ce modèle simule les différents étages du système auditif périphérique à l'aide de filtres auditifs correspondants aux bandes critiques de l'oreille [Patterson, 1976]. Contrairement à une représentation temps-fréquence, la représentation auditive nous renseigne sur l'impact perceptif des harmoniques moteur. Les regroupements d'harmoniques par le système auditif apparaissent plus clairement sur le modèle auditif et la distribution relative de leur énergie nous renseigne sur le timbre du bruit moteur (voir figure 6).

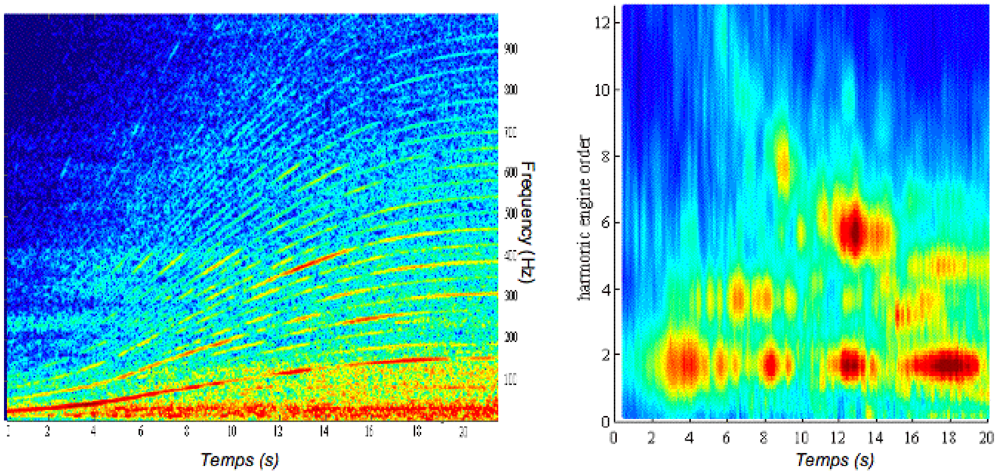


FIGURE 6: Représentation temps-fréquence d'un signal enregistré à l'intérieur d'un voiture en accélération (image de gauche) et représentation du temps en fonction du numéro harmonique de la sortie du modèle auditif du même signal (image de droite).

De façon à mieux interpréter les représentations issues des modèles auditifs, une analyse sensorielle a été réalisée suivant le même protocole expérimental que celui décrit précédemment dans le cadre de bruits de portières automobiles (section 4.2.2). Trois descripteurs principaux ont été mis en évidence par le panel, à savoir "ON" qui caractérise le bourdonnement du bruit moteur et dépend de l'émergence des harmoniques paires, "REU" qui caractérise la rugosité du son et "AN" qui traduit la richesse harmonique du spectre du bruit moteur. Ces résultats laissent penser qu'une structure formantique est contenue dans ces sons. Les représentations auditives sont cohérentes avec cette observation et mettent clairement en évidence un changement de timbre qui correspondrait à la transition du "ON" au "AN" lors de l'accélération. Cette structure, qui est fortement lié à la structure de l'habitacle, est actuellement exploitée dans le but de définir un contrôle intuitif dit de haut niveau (lié à la qualité, la sportivité ou la solidité du véhicule) des bruits moteur à partir d'un modèle de synthèse.

Cette étude a soulevé plusieurs questions qui relèvent de la recherche fondamentale. La première concerne la représentation du caractère dynamique de ce type de sons. Actuellement il n'existe pas de descripteurs de timbre qui permettent de décrire cette dynamique fondamentale du point de vue perceptif et qui doit impérativement être caractérisée pour établir des règles de transformations morphologiques dans le cadre du design sonore. Des études qui ont pour but de proposer des solutions à ce problème sont actuellement en cours dans notre équipe. Les techniques développées dans ce cadre, font l'objet d'une collaboration avec le LATP à Marseille dans le cadre de la thèse d'Anaïk Olivero. Elles sont basées sur l'estimation d'une fonction de transfert temps-fréquence (appelée masque de Gabor) permettant la caractérisation des différences structurelles entre deux sons. Elle a abouti à une méthode de classification hiérarchique des sons et des processus de "morphing sonore" (transition continue entre deux sons). Cette approche devrait aboutir à une description morphologique du timbre tout en procurant des outils mathématiques nouveaux pour la caractérisation des sons. Une deuxième question fondamentale liée à cette étude concerne le sens évoqué par ces sons à travers leurs morphologies spécifiques. Peut-on trouver une morphologie sonore "idéale" qui renvoie une information précise sur le comportement du véhicule et jusqu'à quel niveau de précision l'information peut-elle être transmise par les sons ? Autrement dit, existe-t-il un véritable sens ou une sémiotique des sons ? Ces questions ont abordées dans la section suivante.

4.3 La sémiotique des sons

L'identification de descripteurs ou morphologies du signal qui donnent lieu à une interprétation cognitive du message sonore est primordiale pour la construction de modèles sonores réalistes dans le cadre du design sonore, la sonification ou la réalité virtuelle et/ou augmentée. Cette approche conjecture l'existence d'une sémiotique sonore en lien avec la structure acoustique des signaux[1]. La deuxième action du projet senSons avait pour but de vérifier cette hypothèse à travers une approche "neuro-acoustique" associant la modélisation sonore et des méthodes d'imagerie cérébrale (mesures éléctrophysiologiques). Ces études ont fait l'objet d'une collaboration resserrée

[1] La sémiotique est l'étude de signes ou de symboles et leur signification. De façon simpliste, la sémiotique sonore telle qu'elle est présentée dans ce document sous-tend l'étude du sens communiqué par des sons.

avec l'équipe “Langage, musique et motricité” de l'INCM (M. Besson, M. Aramaki et D. Schön). Deux protocoles expérimentaux ont été utilisés à ces fins ; un protocole d'amorçage et un protocole de catégorisation. Des sons environnementaux ainsi que des sons abstraits issus de l'approche acousmatique ont été étudiés.

4.3.1 Méthode des Potentiels Evoqués

Les méthodologies propres aux neurosciences cognitives n'étant pas bien connues dans le domaine de l'acoustique et du traitement des sons, je présente dans cette section les bases d'une méthode d'imagerie cérébrale : les potentiels évoqués. En plaçant des électrodes sur le scalp de volontaires humains, il est possible d'enregistrer les variations de l'activité cérébrale, connues sous le nom d'électroencéphalogramme (EEG) [Berger, 1929]. L'EEG ayant une faible amplitude (de l'ordre de 100 μV), les signaux analogiques sont amplifiés, puis convertis en signaux numériques. N'importe quel stimulus, tel qu'un son, une lumière, une odeur, etc., provoque des changements ou variations du potentiel dans l'EEG. Néanmoins, ces variations sont très petites (de l'ordre de 10 μV), par rapport à l'activité électrique de base. De façon à extraire le signal utile, il est alors nécessaire de synchroniser l'enregistrement EEG par rapport au début de la stimulation, et de moyenner un grand nombre d'essais obtenus dans les mêmes conditions expérimentales et pour des stimuli de même type. Ainsi, le signal, c'est-à-dire les variations de potentiel évoquées par la stimulation (Potentiels Évoqués ou PEs), pourront émerger de l'activité électrique de base. Typiquement, dans des expériences perceptives et cognitives, 20 à 30 stimuli pour chaque condition expérimentale sont nécessaires pour obtenir un rapport signal sur bruit satisfaisant par “moyennage”. Les Potentiels Évoqués contiennent une série de déflections positives et négatives par rapport à la ligne de base, appelés composantes. La ligne de base correspond au niveau moyen de l'activité cérébrale pendant 100 ou 200 millisecondes (ms) avant la stimulation. Les composantes sont définies par leur polarité (négative, N, ou positive, P), leur latence à partir du début du stimulus (100, 200, 300, 400 ms, etc.), leur distribution sur le scalp (localisation de l'amplitude maximale sur le scalp), ainsi que leur sensibilité liée aux facteurs expérimentaux. Typiquement, les PEs enregistrés dans des expériences perceptives et cognitives comprennent des composantes précoces qui reflètent les étapes sensorielles et perceptives du traitement de l'information (e.g., P100, N200, etc.) et des composantes plus tardives qui reflètent les étapes cognitives et décisionnelles (P300, N400, etc.). Les PEs sont analysés dans des fenêtres de latence différentes, centrées autour de l'amplitude maximale de la composante étudiée. Par exemple, pour analyser la composante N400, une fenêtre de latence comprise entre 300 et 600 ms est généralement utilisée. L'amplitude de la composante est obtenue à partir de la moyenne des acquisitions dans ce rang de latence.

La méthode des Potentiels Évoqués a été largement utilisée dans de nombreuses études portant sur la sémantique du langage. Cette approche a permis en 1980 à Kutas et Hillyard [Kutas and Hillyard, 1980] de démontrer l'existence d'une déflection négative de l'activité cérébrale électrique (composante N400), dont l'amplitude est modulée par le degré de congruence sémantique. Ce résultat est illustré dans la Figure 7.

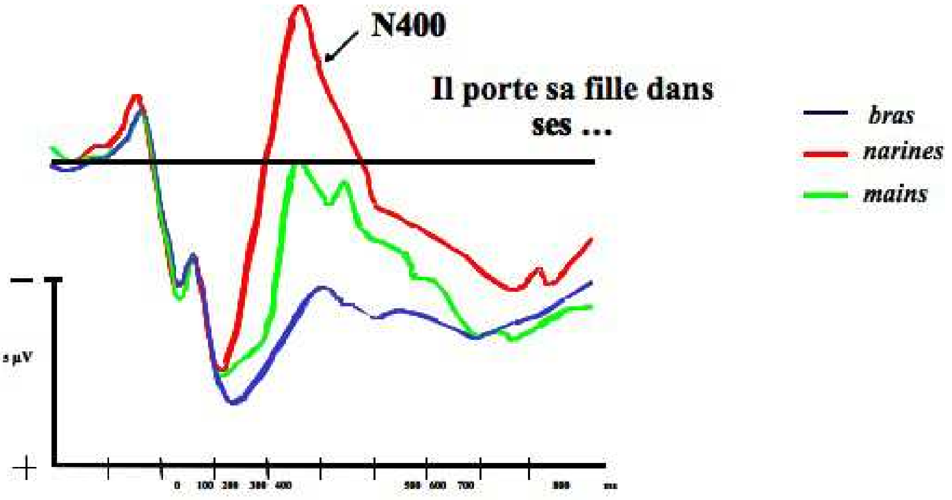


FIGURE 7: N400 et élaboration du sens (la négativité est vers le haut).

Dans cet exemple, des phrases dont le dernier mot peut être congruent ou incongru, ont été présentées à des sujets pendant que leur activité cérébrale était mesurée. Les résultats ont montré une augmentation de l'amplitude de la composante N400 lorsque le dernier mot était incongru. De plus, cette augmentation dépend du degré d'incongruité. Cet exemple montre que la méthode des Potentiels Évoqués permet de mesurer le degré de congruence sémantique d'un mot dans un contexte. D'autres études sur des couples de mots isolés appartenant ou non à la même catégorie sémantique ont également mis en évidence l'existence de la composante N400. Ainsi, lorsque deux mots appartiennent à la même catégorie sémantique (par exemple "chien-chat" ; "docteur-infirmière"), le second mot de la paire évoque une composante N400 de plus faible amplitude que s'il appartient à une catégorie sémantique différente du premier mot (par exemple "poire-chat" ; "abeille-infirmière") [Bentin et al., 1985].

Si de nombreuses études ont permis de mettre en évidence une activité cérébrale spécifique au traitement sémantique dans le langage, il n'en est pas de même pour la sémiotique des sons. Plusieurs études ont tenté de mettre en évidence des situations où la sémantique (ou sémiotique) sonore ferait l'objet du même traitement cognitif que la sémantique dans le langage. Van Petten et Rheinfelder [Van Petten and Rheinfelder, 1995] ont été parmi les premiers à utiliser des sons environnementaux dans le cadre d'un protocole d'amorçage utilisant des sons en amorce et des mots reliés, non reliés ou des pseudo-mots comme cibles. Malgré une différence de distribution d'activité cérébrale sur le scalp (par rapport au langage), ils ont observé une composante N400 plus ample pour les paires sons-mots non-reliées que pour les paires son-mots reliées, ce qui indiquerait un traitement cérébrale similaire pour les sons et pour les mots. Plus récemment, Orgs et collaborateurs [Orgs et al., 2006] ont utilisé des mots comme amorce et des sons environnementaux comme cibles. Les résultats montrent les mêmes tendances lorsque les mots précèdent les sons, à savoir une augmentation de la composante N400 lorsque amorce et cible sont non-reliées. Ce même protocole a été utilisé avec des extraits musicaux par Koelsch et ses collègues [Koelsch et al., 2004]. Des extraits de musique classique suivis par des mots judicieusement choisis ont été présentés aux sujets qui devaient juger la cohérence entre extraits et mots. Les auteurs ont montré que l'amplitude de la N400 liée au mot varie en fonction de sa cohérence avec l'extrait musical qui le précède. Bien que cette étude ait permis de montrer l'existence d'une composante N400 dans une expérience alliant musique et langage, ces résultats ne permettent pas d'affirmer que les sons (ou la musique) induisent le même type de traitement cognitif que la sémantique dans le langage. En effet, l'augmentation de l'amplitude N400 est provoquée par un mot non associé à un extrait musical et non par un son. De plus, l'utilisation d'extraits de musique classique peut refléter l'influence de notre culture sur ces résultats. Dans les études décrites ci-après, nous avons dans un premier temps souhaité limiter l'influence culturelle en associant des sons abstraits aux mots. Puis, dans un second temps, des sons environnementaux ont été utilisés à la fois comme cible et amorce.

4.3.2 Vers une sémiotique de sons isolés

Sons abstraits

Dans l'étude présentée ici nous nous sommes focalisés sur l'association d'un mot à un son, de façon à analyser l'activité cérébrale en fonction de la cohérence entre ces derniers. De manière à dissocier les sons des associations éventuelles liées à notre culture, nous nous sommes efforcés de travailler avec des stimuli “neutres”. Ainsi, nous nous sommes basés sur la théorie de P. Schaeffer (section 4.2.2) en utilisant des sons abstraits (dont la source est difficilement identifiable), favorisant ainsi une écoute analytique. Cette approche permet également de minimiser la médiation linguistique induite par la reconnaissance des sources. En regardant par exemple une image d’un oiseau ou en écoutant son chant, on peut difficilement s’empêcher d’associer le mot “oiseau” aux stimuli visuels ou auditifs. Dans ce cas, l’activité cérébrale pourrait refléter un traitement linguistique qui ne serait pas forcément en lien avec le traitement cognitif visuel ou auditif. Les sons abstraits permettent de limiter cet effet.

En se basant sur la typologie de Schaeffer, un corpus sonore représentatif de la majorité des structures sonores que l’on rencontre dans la nature a été constitué (sons itératifs, continus, impulsifs, complexes, tonals, à hauteurs variables...) à partir de 45 sons concrets ou produits par synthèse. Les sons ont été présentés à 11 sujets dont la tâche était d’associer un (ou plusieurs) mot(s) à chaque son. Certains sons ont suscités un grand nombre d’associations très variées, indiquant que l’association mots/sons est souple pour les sons abstraits. Les mots les plus fréquemment évoqués ont ensuite été choisis, formant ainsi 45 paires mots/sons associés. Une deuxième liste de mots, dont la signification n’est pas associée aux sons a ensuite été constituée formant ainsi un corpus de 45 triplets “mots/sons associés/sons non associés”.

Ces stimuli ont ensuite été utilisés dans une expérience électrophysiologique. La tâche du sujet consistait à juger (le plus rapidement possible) si le son était relié au mot ou non en appuyant sur un des 2 boutons de réponse. Les données comportementales (pourcentage d’erreurs et Temps de Réaction, TR) ainsi que l’activité cérébrale (EEG) ont été enregistrées. Les données comportementales montrent que le TR est relativement long et ne semble pas dépendre de la nature des couples (mots/sons associés et mots/sons non-associés). Le pourcentage d’erreur est dans l’ensemble faible et moins élevé pour les mots non reliés aux sons (5%) que pour les mots reliés aux sons (16%). Ainsi il serait plus facile de décréter que les sons ne sont pas associés aux mots, plutôt que d’affirmer qu’ils le sont. Il est intéressant de noter que certains couples mots/sons sont jugés comme associés par tous les sujets alors que d’autres couples sont beaucoup plus ambigus (50% d’erreurs).

Les résultats issus des mesures électrophysiologiques montrent une augmentation relative de négativité entre 300 et 600 ms lorsque le son n’est pas associé au mot qui le précède, par rapport à la condition où le son est associé au mot (Figure 8). Ces résultats sont similaires à ceux obtenus avec des mots sémantiquement reliés et non reliés dans le langage [Schön et al., 2006, 2009].

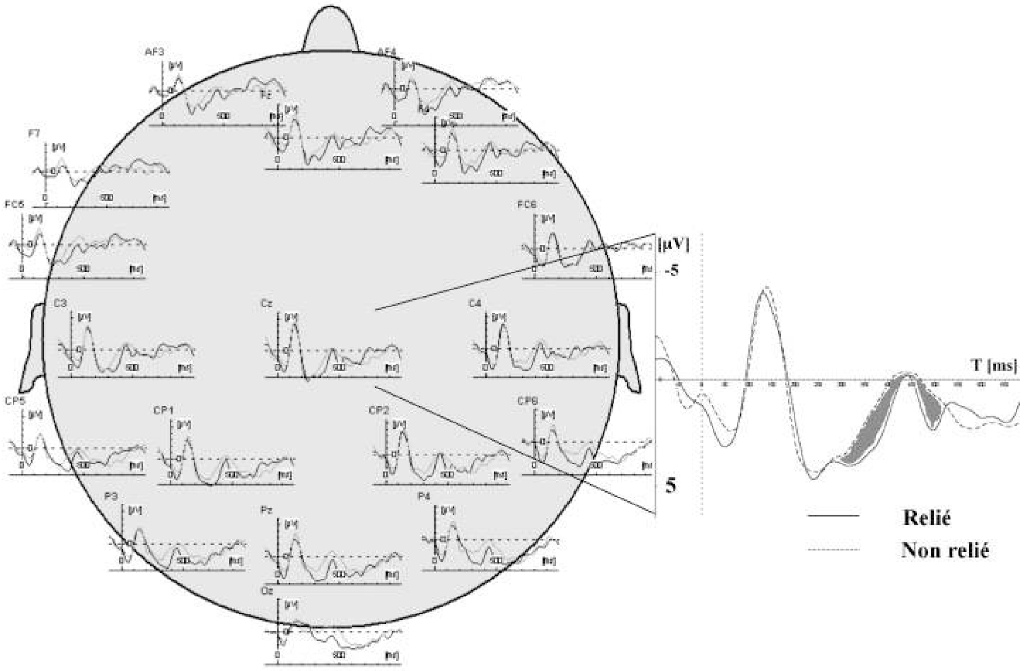


FIGURE 8: Potentiels Évoques par le second stimulus (le son) et enregistrés à partir de différentes électrodes sur le scalp (la négativité est vers le haut)

En ce qui concerne les analyses acoustiques, les résultats montrent que les sons riches en composantes spectrales et variants peu au cours du temps (entretenu et à hauteur complexe selon la typologie de Schaeffer), sont très souvent associés à des mots décrivant des situations désagréables (frisson, froid, douleur, etc.). Les sons dont la densité spectrale est plus faible et dont les composantes manifestent de plus une tendance non-stationnaire avec des évolutions de fréquence au cours du temps (sons entretenues à hauteur variable selon la typologie de Schaeffer) sont très souvent associés aux mouvements (rebond, monté, roulement, rotation, etc.)[Ystad et al., 2008].

Sons environnementaux
Dans l'expérience précédente des sons abstraits ont été choisis pour minimiser la médiation linguistique. Néanmoins, cette expérience fait intervenir des

mots en association avec des sons, n'excluant pas complètement une éventuelle médiation linguistique. Une deuxième expérience utilisant le même protocole avec des sons environnementaux a été effectué pour contourner ce problème. Des sons issus de l'expérience de la catégorisation perceptive de sons d'impact (section 4.2.2) ont été utilisés dans ce cadre. Dans l'expérience précédente, des continua sonores entre trois différentes catégories de matériaux (Bois, Métal, Verre) étaient construits par synthèse. Les sujets devaient catégoriser tous les sons dans une des trois catégories. Deux classes de sons ont ainsi été définies : des sons typiques et des sons ambigus. Les sons catégorisés par plus que 70% des participants ont été considérés comme typiques, et ceux catégorisés par moins de 70% des participants comme ambigus. Les sons typiques et ambigus ont ensuite été utilisés dans le protocole d'amorçage sémantique. Des sons typiques étaient systématiquement utilisés comme amorce et les sons cibles étaient soit des sons typiques de la même catégorie que l'amorce, soit des sons typiques d'une autre catégorie, soit des sons ambigus. Les sujets devaient déterminer si cible et amorce appartenaient à la même catégorie ou non.

Les résultats montrent des similitudes avec ceux obtenus sur le même groupe de participants avec des stimuli linguistiques (mot français comme amorce, mots ou pesudo-mot ou non-mots, comme cible). Les données électrophysiologiques ont révélé une négativité fronto-centrale autour de 450 ms après le début de la cible plus ample pour les cibles ambiguës que pour les cibles reliées, et une positivité (composante P300) dans les régions pariétales pour les cibles non reliées. Ainsi, l'ensemble de ces résultats reflèterait la mise en jeu de processus cérébraux d'amorçage conceptuel communs aux stimuli sonores et linguistiques et iraient dans le sens d'une véritable sémiotique des sons [Aramaki et al., 2009c].

4.3.3 Bases neuronales de la catégorisation sonore

Les résultats obtenus dans le cadre des expériences précédentes indiquent l'existence d'une sémiotique sonore et nous permettent d'envisager la construction de synthétiseurs qui puissent être contrôlés à partir de descripteurs verbaux de haut niveau. Pour identifier les paramètres qui permettent un tel contrôle, il faut identifier des liens entre le comportement physique des sources sonores, la perception des sons engendrés et le sens évoqué par ces sons. Le protocole de catégorisation utilisé précédemment nous a permis d'aborder ce problème en examinant les bases neuronales de la catégorisation et dégager, à travers les tracés éléctrophysiologiques, des différences de traitement cognitif pour différentes catégories de sons. L'analyse des résultats montre que les processus cérébraux mis en jeu lors du traitement de sons typiques de la catégorie Métal diffèrent de façon significative de ceux mesurés avec des sons typiques des deux autres catégories, dès 150 ms et jusqu'à 700 ms après la présentation du son (Figure 9).

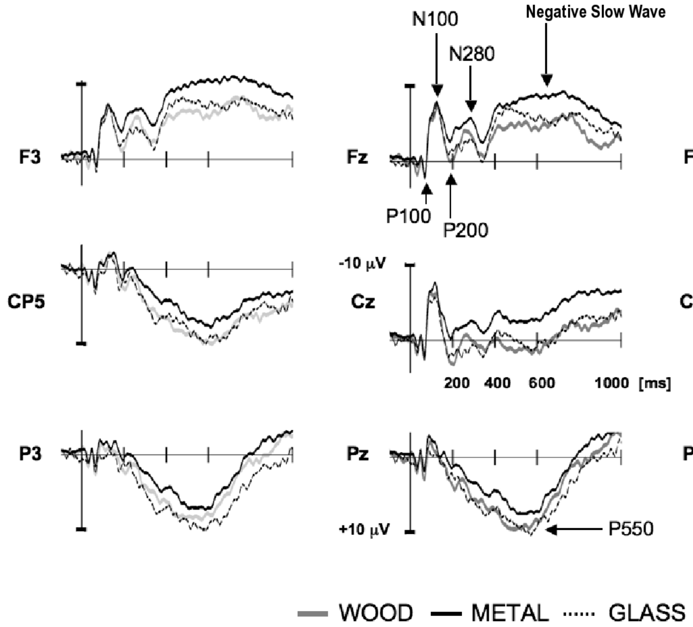


FIGURE 9: Potentiels Evoqués de sons typiques de Bois (tracés gris), Métal (tracés noir) et Verre (pointillé) des électrodes centrales (Fz, Cz, Pz) et latérales ({F3, CP5, P3} / {F4, CP6, P4})). L'amplitude (en microvolts, μV) est représentée en ordonné et la négativité est vers le haut. Le temps à partir du début du son est représenté en abscisse (en millisecondes, ms).

Les analyses statistiques des corrélations entre descripteurs acoustiques et catégories de matériaux montrent que la différentiation entre les sons issus de la catégorie Métal et les sons des autres catégories peut s'expliquer par des descripteurs liés au contenu spectral du son (largeur de bande, barycentre spectral, flux spectral et rugosité) et la durée (liée à l'amortissement). L'absence de différentiation précoce (antérieure à 150ms) dans les ERPs nous incite à rejeter les descripteurs tels que le temps d'attaque [Caclin et al., 2006]. Les différences autour de 200ms (P200) sont susceptibles de refléter un traitement lié à la "complexité" spectrale [Shahin et al., 2005; Kuriki et al., 2006] tandis que les différences plus tardives refléteraient des différences de durée (i.e. des différences d'amortissement [Kushnerenko et al., 2001; Alain et al., 2002]). La rugosité semble ainsi jouer un rôle important dans la distinction du Métal par rapport aux 2 autres catégories. Nous verrons plus loin que cette conjecture a été vérifiée à l'aide de la synthèse grâce à un contrôle uniquement basé sur ce descripteur. L'association de données comportementales, electrophysiologiques et acoustiques nous a permis de mieux comprendre la catégorisation de matériaux et d'entrouvrir la voie à une approche nouvelle : la "neuro-acoustique"[Aramaki et al., 2009a]. Ces résultats nous permettent d'aborder le contrôle haut niveau des processus de synthèse.

4.4 Vers un contrôle haut niveau des processus de synthèse

Le contrôle intuitif de processus de synthèse permettant aux non-experts de construire des sons à partir d'évocations ou d'événements nécessite une mise en correspondance à plusieurs niveaux entre paramètres de synthèse et paramètres de contrôle. C'est la raison pour laquelle nous avons introduit le terme contrôle haut niveau qui reflète le niveau de complexité des paramètres de contrôle. Le contrôle haut niveau est d'un grand intérêt dans les domaines où, plus que l'exactitude du son, l'impact perceptif est d'importance. L'identification de paramètres pertinents du point de vue perceptif et la mise en correspondance entre ces paramètres et les paramètres du contrôle est un problème très complexe qui est rarement abordé, ce qui explique le faible nombre d'interfaces intuitives qui existent à ce jour. Certains projets ont néanmoins abordé ce problème, tel le projet Européen "The Sounding Object" (http ://www.soundobject.org/, 2001-2003), basé sur une approche pluridisciplinaire associant la synthèse et la perception et qui a permis d'adapter les caractéristiques de systèmes de synthèse interactifs aux actions (gestes) ainsi qu'aux attentes (retour sonore) de l'utilisateur. Plus récemment, le projet Européen CLOSED, (http ://closed.ircam.fr/, 2006-2009), a également permis la mise en place d'interactions sonores temps-réel entre utilisateurs et objets par une approche pluridisciplinaire basée sur la synthèse sonore et la psychologie cognitive. Cependant, aucun de ces projets n'a pu conclure sur un contrôle haut niveau adapté à des utilisateurs non experts et basé sur une description factuelle des scènes sonores.

L'objectif de notre approche est de réaliser une plate forme de synthèse sonore temps-réel conçue autour d'une architecture logicielle évolutive dont le but est de générer des sources sonores diverses. Les études précédentes liées aux sons d'impact et aux sons abstraits nous ont permis de construire les fondations de cette plate-forme qui actuellement permet le contrôle haut niveau de sons d'impacts issus de différents matériaux. Des sons environnementaux simulant des effets naturels (vent, pluie, vagues) ont également été intégrés dans cette plate-forme dans le cadre de la thèse de Charles Verron [Verron, 2010], et les évocations de mouvement vont prochainement y être ajoutées suites aux expériences perceptives effectués dans le cadre da la thèse d'Adrien Merer (section 4.2.2).

4.4.1 Sons d'impact

Dans le but de construire un synthétiseur de sons d'impact permettant un contrôle haut niveau, une stratégie de mise en correspondance entre paramètres ("mapping") basée sur une architecture à trois niveaux a été proposée. Le niveau le plus élevé permet un contrôle à partir de descripteurs verbaux liés au type de matériau, à la taille et à la forme de l'objet et au système d'excitation. Le niveau intermédiaire permet un contrôle axé sur des descripteurs acoustiques (amortissement, rugosité...) tandis que le bas niveau permet le contrôle direct sur les paramètres de synthèse (amplitudes, fréquences, phases ...). Le lien entre les différents niveaux est illustré par la Figure 10.

Le lien entre les contrôles de haut niveau et ceux du niveau intermédiaire est basé sur des résultats issus des tests perceptifs. Ainsi, il a été montré que la taille de l'objet est fortement liée à la hauteur perçue du son, tandis que la forme de l'objet est corrélée à la distribution des composantes spectrales. Les expériences effectuées dans le cadre de la catégorisation de sons d'impact issus de différents matériaux (sections 4.2.2 et 4.3.3) montrent que la reconnaissance des matériaux est essentiellement liée à l'amortissement des composantes spectrales (et à leur dépendance fréquentielle) ainsi qu'à la rugosité. L'association de ces descripteurs permet ainsi de proposer une stratégie de contrôle permettant une navigation intuitive dans un espace de matériaux [Aramaki et al., 2006; Aramaki et al., 2010b,a].

4.4.2 Evocations de la dynamique

Les sons d'impacts sont bien repérés du point de vu auditif car ils correspondent à des sons produits dans la vie de tous les jours. Le contrôle des processus de synthèse associés ont ainsi pu être bâtis sur des connaissances a priori des caractéristiques physiques et perceptives de ces sons. La dynamique contenue dans un son est un concept beaucoup moins précis et contrôler un tel attribut nécessite une approche adaptée. Comme discuté précédemment, la dynamique des bruits moteurs ou encore les variations

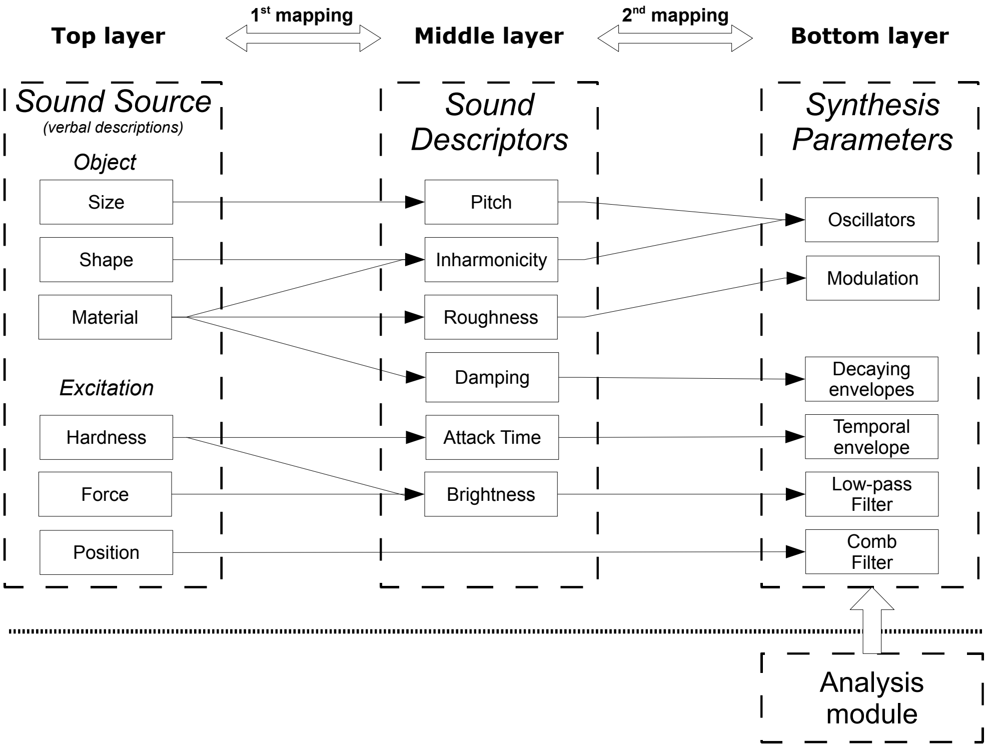
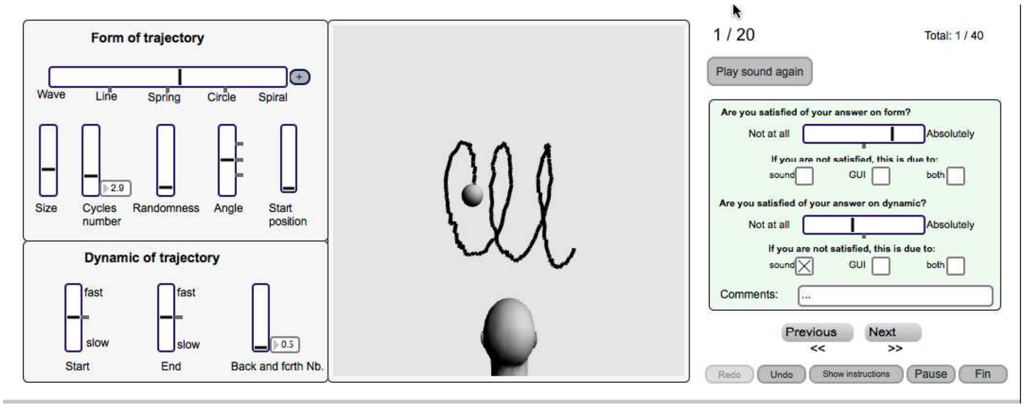


FIGURE 10: Stratégie de contrôle du synthétiseur de sons d'impact

de timbre dans l'interprétation sont primordiales pour le jugement apporté aux sons. L'étude présentée ici aborde le problème du contrôle des aspects dynamiques contenus dans les sons en se basant sur les résultats issus des études perceptives des sons abstraits décrites précédemment (section 4.2.2). Ce travail fait l'objet de la thèse d'Adrien Merer que je dirige.

Les tests de catégorisation libre décrit précédemment ont permis d'identifier des catégories liées aux évocations dynamiques procurées par l'écoute des sons. Cependant, aucune information pertinente ne peut être déduite quant à la façon de contrôler ces évocations dans le cadre de la synthèse. Pour aborder ce difficile problème, nous nous sommes appuyés sur une constatation intéressante axée sur le fait que certains sujets ont spontanément fait des dessins décrivant le mouvement perçu sans que cela soit explicitement demandé dans les consignes (section 4.2.2). Nous avons alors imaginé un protocole consistant à demander aux sujets de décrire la trajectoire dynamique qu'évoque un son à l'aide d'un dessin construit à partir d'une interface graphique telle que présentée sur la Figure 11. Nous nous sommes limités ici à six paramètres principaux, à savoir forme, taille, fréquence, irrégularité, orientation et direction. Cinq formes prédéfinis étaient contenues dans le paramètre forme, à savoir onde, ligne, ressort, cercle et spirale.



1 / 20 Total: 1 / 40

Are you satisfied of your answer on form?

Not at all Absolutely

If you are not satisfied, this is due to:

sound □ GUI □ both □

Are you satisfied of your answer on dynamic?

Not at all Absolutely

If you are not satisfied, this is due to:

sound ☒ GUI □ both □

Comments: ...

FIGURE 11: Interface graphique du test d'écoute.

Les sujets avaient également la possibilité d'agir sur la vitesse et le nombre d'aller-retours pour lesquels la visualisation était assurée par une boule mobile le long de la trajectoire. Pour chaque réponse, les sujets devaient en outre évaluer leur dessin, donc leur propre jugement de l'adéquation qu'ils avaient obtenus entre la dynamique perçue et le dessin produit. Pour minimiser une éventuelle influence liée à la reconnaissance des sources sonores, des sons abstraits (pour lesquels la source était difficilement reconnaissable) ont été utilisés dans ce test. Un premier test a été préalablement effectué pour vérifier que les sujets n'associaient pas facilement des sources aux sons et que ces derniers évoquaient bien un mouvement ou une dynamique.

L'analyse des réponses des 29 sujets qui ont participé à ce test nous a permis d'identifier les paramètres pertinents du point de vue perceptif et d'identifier les paramètres qui semblent pertinents au niveau du contrôle. Les résultats montrent que plusieurs formes peuvent être utilisées pour décrire le même son. Les sujets semblent s'accorder par rapport à la différentiation entre formes oscillantes ou non et entre les oscillations régulières et circulaires. Les précisions liées aux formes des oscillations (circulaire, ressort, spirale) ne sont pas pertinentes. Trois types de formes semblent alors suffire pour le contrôle de la synthèse : linéaire, oscillations régulières et oscillations circulaires. Le nombre d'oscillations ou cycles ne sont pas précisément décrites par les sujets, et un contrôle à deux niveau (peu ou beaucoup) semble suffisant. En ce qui concerne l'orientation de la trajectoire, seule la distinction horizontale/verticale semble pertinente. Les sujets distinguent la direction haut/bas, mais gauche/droite est confondue. Le contrôle de la taille des dessins ainsi que l'irrégularité, est déterminée directement à partir du réglage des paramètres de l'interface effectué par les sujets. Enfin, en ce qui concerne la dynamique, les sujets distinguent les trajectoires à vitesse constante et les trajectoires à vitesse variable, mais les précisions au niveau des variations des vitesses ne sont pas cohérentes entre sujets. Ces résultats nous permettent de proposer un contrôle haut niveau permettant aux utilisateurs de synthétiser des sons soit à partir de descripteurs verbaux (forme, taille, direction etc...), soit à partir de dessins. En effet, les résultats issus de cette étude devraient permettre l'extraction d'une trajectoire "squelette" contenant des formes perceptivement pertinentes à partir de n'importe quel dessin et de synthétiser le son correspondant, soit en se basant sur une texture sonore existante, soit sur un modèle de synthèse.

Cette étude ouvre de nouvelles perspectives quant aux protocoles expérimentales. Contrairement aux protocoles traditionnels où les sujets doivent effectuer un jugement sur les sons, soit en terme de qualité, de dissemblance, ou de catégorisation, les interfaces interactives peuvent permettre aux sujets de calibrer les sons à leur convenance en leur donnant accès aux paramètres du signal.

5 Conclusion et perspectives

Les travaux présentés dans ce document sont le résultat d'une démarche scientifique visant à mieux comprendre la relation entre la structure (ou morphologie) acoustique des sons et le sens communiqué par ces derniers. Ils s'appuient sur une approche pluridisciplinaire associant sciences dures et sciences humaines où la synthèse numérique des sons constitue le point central. Ce travail aborde les aspects fondamentaux liés à la compréhension du traitement cognitif et perceptif des sons mais aussi les aspects appliqués pour lesquels le contrôle haut niveau des processus de synthèse joue un rôle privilégié. Les principaux résultats obtenus ont été déclinés suivant deux axes : les sons musicaux et les sons environnementaux et industriels. Pour les sons musicaux, il s'agissait de mieux comprendre comment les musiciens transmettent une émotion à travers les sons qu'ils produisent. Une étude basée sur les relations entre le jeu musical et le timbre produit a été réalisée sur la base de sons issus d'un modèle physique de synthèse. Cette étude a permis de conclure sur une méthodologie susceptible d'identifier les descripteurs de timbre pertinents et leur relation directe avec les paramètres de jeu. Ces descripteurs ont ensuite été utilisés pour décrire les variations de timbre au cours du jeu musical et montrer ainsi l'influence du timbre dans l'expressivité. Les premières approches associant l'acoustique et les sciences cognitives ont également été décrites dans le cas de la catégorisation libre des sons de piano. Cette dernière étude a montré toute la complexité des processus d'écoute et de catégorisation (contexte, expertise, ...). Elle a de plus confirmé les conclusions émises dans le cadre de l'étude sur l'interprétation musicale, à savoir que le message musical se décline par l'interaction des éléments (notes) qui le constituent, plutôt que par la structure des notes indépendantes. Finalement, une approche associant l'acoustique et l'imagerie cérébrale a été présentée dans le cadre de l'étude de l'influence de l'allongement temporel des syllabes dans le langage et des notes dans la musique. Cette étude a permis d'une part d'intégrer des données liées à l'activité cérébrale à des données liées à des transformations fines du signal, et d'autre part de mettre en évidence les effets de telles manipulations sur l'accès au sens. Elle ouvre de fait la voie à une interaction pluridisciplinaire pertinente.

Dans le cadre de l'étude des sons environnementaux et industriels, il s'agissait de mieux comprendre comment les sons peuvent être utilisés pour procurer une information appropriée. Ce problème se décline de façon naturelle dans la construction d'outils de synthèse numérique contrôlés à partir de descripteurs verbaux ou intuitifs issus, par exemple, de formes graphiques. Ce problème repose sur l'identification de structures invariantes, caractéristiques de catégories sonores. Lorsque les sons considérés relèvent de processus physiques bien repérés, l'association de concepts physiques, perceptifs et cognitifs permet de conjecturer des invariants que la synthèse numérique peut valider. Tel est le cas des sons d'impacts sur des objets de nature et de matériaux différents, pour lesquels seuls quelques descripteurs permettent la synthèse à partir de la description factuelle des objets et des interactions. Lorsque le ressenti est de nature émotionnelle, seul le recours aux tests perceptifs et/ou cognitifs est possible, même si la physique peut parfois suggérer des hypothèses. Tel est le cas de la dynamique interne des sons. Dans ce cas, l'utilisation de sons favorisant une écoute analytique présente de grands avantages. En s'inspirant des travaux de P. Schaeffer [Schaeffer, 1966] des sons "abstraits" ont été introduits dans nos protocoles expérimentaux, permettant d'une part d'étudier la perception des sons en minimisant l'influence de notre culture sur ces résultats et d'autre part de caractériser l'ensemble de morphologies sonores grâce à la typologie proposée. Les protocoles ainsi construits ont permis d'identifier 5 catégories principales de dynamique et de proposer un contrôle graphique de cette dernière dans le cas de la synthèse numérique. La notion d'invariant structurel et de sens communiqué est également pertinente dans le cas des sons industriels. Des applications aux sons produits par les portières automobiles et aux bruits engendrés par les moteurs en phase dynamique ont été abordées et ont permis de préciser les corrélation entre la structure acoustique des sons et le ressenti produit.

La question centrale de l'existence d'une sémiotique sonore a enfin été abordée en utilisant des protocoles propres aux neurosciences cognitives : l'amorçage sémiotique et la catégorisation. L'utilisation de sons abstraits dans un protocole d'amorçage sémiotique a permis de montrer l'augmentation d'une composante N400 lorsque le mot présenté et le son produit ne sont pas associés. Cette augmentation reflète une incongruité sémantique dans le langage et va dans le sens d'une sémiotique sonore. Cet aspect a été précisé grâce à un protocole d'amorçage associant uniquement des sons d'environnement. Ici encore une augmentation de la composante N400 a été observée lorsque des sons appartenant à des catégories différentes sont présentés au sujet.

L'existence d'une sémiotique sonore et de méthodes d'extraction d'invariants structurels responsables de certaines évocations permettent d'envisager la construction de systèmes de synthèse contrôlés à haut niveau. Deux réalisations visant au contrôle de sons d'impacts et du ressenti dynamique des sons ont été mise en œuvre. Les contrôles intuitifs proposés ont montré la possibilité de construire des scènes sonores interactives à partir de données simples et factuelles. Dès lors la synthèse sonore ouvre une voie captivante dans l'utilisation des sons comme moyen de communication. Les perspectives de ce travail vont dans ce sens et visent à étendre notre expertise à la construction d'un véritable langage des sons et au développement d'outils de génération de métaphores sonores basées sur les invariants et permettant d'évoquer une image mentale spécifique à partir d'attributs perceptifs et cognitifs. Ces métaphores seront obtenues par conformation de textures sonores initialement inertes et manipulées à partir de contrôles intuitifs haut niveau. En pratique, la réalisation d'une plateforme évolutive de démonstration temps-réel est en préparation dans le cadre du projet ANR METASON auquel je participe et dont le démarrage est prévu au premier novembre 2010.

Bibliographie

Alain, C., Schuler, B. M., & McDonald, K. L. (2002). Neural activity associated with distinguishing concurrent auditory objects. *Journal of Acoustical Society of America*, 111(2) :990–995.

Aramaki, M., Baillères, H., Brancheriau, L., Kronland-Martinet, R., & Ystad, S. (2007). Sound quality assessment of wood for xylophone bars. *Journal of the Acoustical Society of America*, 121(4)(4) :2407–2420.

Aramaki, M., Besson, M., Kronland Martinet, R., & Ystad, S. (2009a). Timbre perception of sounds fromimpacted materials : behavioral, electrophysiological and acoustic approaches. In : *Computer music modeling and retrieval : genesis of meaning in sound and music*, J. Ystad, Kronland-Martinet, ed., Lecture Notes in Computer Science, pages 1–17. Springer Berlin / Heidelberg.

Aramaki, M., Besson, M., Kronland Martinet, R., & Ystad, S. (2010a). Controlling the perceived material in an impact sound synthesizer. *IEEE Transactions on Audio, Speech, and Language Processing*, PP(99) :1–1.

Aramaki, M., Brancheriau, L., Kronland Martinet, R., & Ystad, S. (2009b). Perception of impacted materials : sound retrieval and synthesis control perspectives. In : *Computer music modeling and retrieval : genesis of meaning in sound and music*, J. Ystad, Kronland-Martinet, ed., Lecture Notes in Computer Science, pages 134–146. Springer Berlin / Heidelberg.

Aramaki, M., Gondre, C., Kronland Martinet, R., Voinier, T., & Ystad, S. (2010b). Imagine the sounds : anintuitive control of an impact sound synthesizer. In : *Auditory display*, Ystad, Aramaki, Kronland Martinet, & Jensen, ed., Lecture Notes in Computer Sciences, pages 408–422. Springer Berlin / Heidelberg.

Aramaki, M. & Kronland-Martinet, R. (2006). Analysis-synthesis of impact sounds by real-time dynamic filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2) :695–705.

Aramaki, M., Kronland-Martinet, R., Voinier, T., & Ystad, S. (2006). A percussive sound synthesizer based on physical and perceptual attributes. *Computer Music Journal*, 30(2) :32–41.

Aramaki, M., Marie, C., Kronland Martinet, R., Ystad, S., & Besson, M. (2009c). Sound categorization and conceptual priming for nonlinguistic and linguistic sounds. *Journal of Cognitive Neuroscience*, pages 1–15.

Bacon, S. P. & Viemeister, N. F. (1985). Simultaneous masking by gated and continuous sinusoidal maskers. *The Journal of the Acoustical Society of America*, 78 :1220–1230.

Barthet, M. (2008). *De l'interprète à l'auditeur : une analyse acoustique et perceptive du timbre musical.* PhD thesis, Université Aix-Marseille II.

Barthet, M., Depalle, P., Kronland Martinet, R., & Ystad, S. (2010a). Acoustical correlates of timbre and expressiveness in clarinet performance. *Music Perception.*

Barthet, M., Depalle, P., Kronland Martinet, R., & Ystad, S. (2010b). Analysis-by-synthesis of timbre, timing, and dynamics in expressive clarinet performance. *Music Perception.*

Barthet, M., Guillemain, P., Kronland Martinet, R., & Ystad, S. (2010c). From clarinet control to timbre perception. *Acta Acustica united with Acustica*, 96 :678–689.

Barthet, M., Kronland Martinet, R., & Ystad, S. (2008). Improving musical expressiveness by time-varying brightness shaping. In : *Computer music modeling and retrieval : sense of sounds*, J. Kronland Martinet, Ystad, ed., Lecture Notes in Computer Science, pages 313–336. Springer Berlin / Heidelberg.

Bensa, J. (2003). *Analysis and synthesis of piano sounds using physical and signal models.* PhD thesis, Université de Provence - Aix-Marseille I.

Bensa, J., Dubois, D., Kronland Martinet, R., & Ystad, S. (2005). Perceptive and cognitive evaluation of a piano synthesis model. In : *Computer music modeling and retrieval*, Uffe Kock Wiil, ed., Lecture Notes in Computer Science, pages 232–245. Springer Berlin / Heidelberg.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision, and semantic priming. *Electroencephalography and Clinical Neurophysiology*, 60(4) :343–355.

Berger, H. (1929). Über das electrenkephalogramm das menchen. *Archiv für Psychiatrie*, 87 :527–570.

Bezat, M.-C. (2007). *Perception des bruits d'impactApplication au bruit de fermeture de porte automobile.* PhD thesis, Université de Provence - Aix-Marseille I.

Bezat, M.-C., Roussarie, V., Kronland Martinet, R., & Ystad, S. (2008). Relations between acoustic parameters and perceptual properties : an apporach by regression tree applied to car door closure sounds. In : *Proc. of Acoustics'08 Acoustics '08.*

Bezat, M.-C., Roussarie, V., Kronland Martinet, R., Ystad, S., & Mcadams, S. (2006). Perceptual analyses of action-related impact sounds. In :

Proceedings of the 6th European Conference on Noise Control Euronoise 2006. thèse CIFRE, PSA, LMA-CNRS.

Bezat, M.-C., Roussarie, V., Voinier, T., Kronland Martinet, R., & Ystad, S. (2007). Car door closure sounds : characterization of perceptual properties through analysis-synthesis approach. In : *Proceedings of the 19th International Congress on Acoustics 19th International Congress on Acoustics*.

Caclin, A., Brattico, E., Tervaniemi, M., Näätänen, R., Morlet, D., Giard, M.-H., & McAdams, S. (2006). Separate neural processing of timbre dimensions in auditory sensory memory. *Journal of Cognitive Neuroscience*, 12(18) :1959–1972.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions : A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America*, 118(1) :471–482.

Chaigne, A. & Kergomard, J. (2008). *Acoustique des instruments de musique*. Belin.

Cook, P. R. (2002). *Real Sound Synthesis for Interactive Applications*. A. K Peters Ltd.

Dillon, W. R. & Goldstein, M. (1984). *Multivariate Analysis*. Wiley series in probability and mathematical statistics. John Wiley & Sons.

Dubois, D. (2000). Categories as acts of meaning : the case of categories in olfaction and audition. *Cognitive Science Quarterly*, 1 :35–68.

Fastl, H. (1979). Temporal masking effects : Iii. pure tone masker. *Acustica*, 43 :282–294.

Flandrin, P. (1993). *Temps-fréquence*. Traité des nouvelles technologies, série traitement du signal. Hermès.

Flandrin, P., G., R., & Gonçavels, P. (2004). Empirical mode decomposition as a filter bank. *IEEE signal processing letters*, 11(2) :112–114.

Friberg, A. (1995). *A Quantative Rule System for Musical Performance*. PhD thesis, Department of Speech, Music and Hearing, Royal Institute of Technology, Stockholm.

Gabrielsson, A. (1999). The performance of music. In : *Psychology of Music*, (2nd ed.). Academic Press.

Gaver, W. W. (1993). What in the world do we hear ? an ecological approach to auditory event perception. *Ecological Psychology*, 5(1) :1–29.

Gobin, P., Kronland-Martinet, R., Lagesse, G.-A., Voinier, T., & Ystad, S. (2003). From sounds to music : Different approaches to event piloted instruments. In : *Computer music modeling and retrieval*, Uffe Kock Wiil, ed., Lecture Notes in Computer Science, pages 225–246. Springer Berlin / Heidelberg.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of theAcoustical Society of America*, 61 :1270–1277.

Griffiths, P. (1979). *A Guide to Electronic Music*. W. W. Norton & Co Inc.

Griffiths, T. D. & Warren, J. D. (November 2004). What is an auditory object ? *Nature Reviews Neuroscience*, 5 :887–892.

Guillemain, P., Kergomard, J., & Voinier, T. (2005). Real-time synthesis of clarinet-like instruments using digital impedance models. *J. Acoust. Soc. Am.*, 118(1) :483–494.

Hajda, J. M., Kendall, R. A., Carterette, E. C., & Harshberger, M. L. (1997). *Methodological issues in timbre research*, (2nd ed.), pages 253–306. Psychology Press.

Huang, N. E., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., Yen, N., Tung, C., & Liu, H. H. (1998). The empirical mode decomposition and hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. Roy. Soc. London A*, 454 :903–995.

Kendall, R. A. & Carterette, E. C. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Perception*, 8 :369–404.

Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., & Friederici, A. (2004). Music, language and meaning : brain signatures of semantic processing. *Nature Neuroscience*, 7(3) :302–307.

Kronland-Martinet, R., Morlet, J., & Grossman, A. (1987). Analysis of sound patterns through wavelet transforms. *International Journal of Pattern Recognition and Artificial Intelligence*, 11(2) :97–126.

Kronland Martinet, R., Voinier, T., & Ystad, S. (2006). *PLAY, Lecture Notes in Computer Science (LNCS 3902)*. Springer Verlag Berlin Heidelberg. ISBN 3-540-34027-0.

Kronland Martinet, R., Ystad, S., & Jensen, K. (2008). *Sense of Sounds*. Springer Verlag Berlin Heidelberg.

Krumhansl, C. L. (1989). *Why is musical timbre so hard to understand ?*, pages 43–53. Excerpta Medica.

Kuriki, S., Kanda, S., & Hirata, Y. (2006). Effects of musical experience on different components of meg responses elicited by sequential piano-tones and chords. *The Journal of Neuroscience*, 26(15) :4046–4053.

Kushnerenko, E., Ceponiene, R., Fellman, V., Huotilainen, M., & Winkler, I. (2001). Event-related potential correlates of sound duration : similar pattern from birth to adulthood. *NeuroReport*, 12(17) :3777–2781.

Kutas, M. & Hillyard, S. A. (1980). Reading senseless sentences : Brain potentials reflect semantic incongruity. *Science*, 207 :203–204.

Laback, B., Balazs, P., Toupin, G., Necciari, T., Savel, S., Meunier, S., Ystad, S., & Kronland Martinet, R. (2008). Additivity of auditory masking using gaussian-shaped tones. In : *Acoustics'08*.

Magne, C., Astesano, C., Aramaki, M., Ystad, S., Kronland Martinet, R., & Besson, M. (2007). Influence of syllabic lengthening on semantic processing in spoken french : behavioral and electrophysiological evidence. *Cerebral Cortex / Cerebral Cortex (Cary)*, 17(11) :2659–2668.

Mathews, M. (1963). The digital computer as a musical instrument. *Science*, 142(3592) :553–557.

Mathews, M. V., Miller, J. E., Pierce, J. R., & Tenney, J. (1965). Computer study of violin tones. *The Journal of the Acoustical Society of America*, 38(5) :912–913.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres : common dimensions, specificities, and latent subject classes. *Psychological Research*, 58 :177–192.

Merer, A., Aramaki, M., Kronland Martinet, R., & Ystad, S. (2008a). Towards a synthesis tool using 'evocation' as control parameters. In : *Acoustics'08*.

Merer, A., Ystad, S., Kronland Martinet, R., & Aramaki, M. (2008b). Semiotics of sounds evoking motions : categorization and acoustic features. In : *Computer music modeling andretrieval : sense of sounds*, J. Kronland

Martinet, Ystad, ed., Lecture Notes in Computer Science, pages 139–158. Springer Berlin / Heidelberg.

Merer, A., Ystad, S., Kronland Martinet, R., Aramaki, M., Besson, M., & Velay, J.-L. (2007). Perceptual Categorization of Moving Sounds For Synthesis Applications. In : *Proceedings of the 2007 International Computer Music Conference International Computer Music Conference*, volume 1, pages 69–72.

Micoulaud-Franchi, J. (2009). Troubles de la reconnaissance du bizarre et du familier dans la perception auditive chez le patient schizophrène.

Miranda E. & Wanderley, M. (2006). *New Digital Musical Instruments : Control and Interaction beyond the Keyboard.* A-R Editions.

Moore, B. C. J. (2003). *An introduction to the psychology of hearing.* Academic Press Inc.

Necciari, T., Savel, S., Meunier, S., Kronland Martinet, R., & Ystad, S. (2010). Masquage auditif temps-fréquence avec des stimuli de forme gaussienne. In : *Actes du 10ème Congrès Français d'Acoustique 10ème Congrès Français d'Acoustique.*

Necciari, T., Savel, S., Meunier, S., Ystad, S., Kronland Martinet, R., Laback, B., & Balazs, P. (2008). Auditory masking using gaussian-windowed stimuli. In : *Acoustics'08.*

Orgs, G., Lange, K., Dombrowski, J., & Heil, M. (2006). Conceptual priming for environmental sounds and words : An erp study. *Brain and Cognition*, 62(3) :267–272.

Overy, K. (2003). Dyslexia and music : From timing deficits to musical intervention. *Ann. N.Y. Acad.Sci.*, 999 :497–505.

Pallone, G. (2003). *Dilatation et transposition sous contraintes perceptives des signaux audio : application au transfert cinéma-vidéo.* PhD thesis, Université de la Méditerranée - Aix-Marseille II.

Pan, D. (1995). A tutorial on mpeg/audio compression. *IEEE Multimedia Journal*, 2(2) :60–74.

Patel, A., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. (1998). Processing syntactic relations in language and music : An event-related potential study. *Journal of Cognitive Neuroscience*, 10 :717–733.

Patterson, R. (1976). Auditory filter shapes derived with noise stimuli. *The Journal of the Acoustical Society of America*, 59(3) :640–654.

Plomp, R. (1970). *Timbre as a multidimensional attribute of complex tones.* Frequency Analysis and Periodicity Detection in Hearing. A. W. Sijthoff.

Rhea, T. (1989). Clara rockmore : The art of the theremin. *Computer Music Journal*, 13(1) :61–63.

Risset, J. C. (1965). Computer study of trumpet tones. *The Journal of the Acoustical Society of America*, 38(5) :912–912.

Risset, J.-C. & Wessel, D. L. (1999). Exploration of timbre by analysis and synthesis. In : *Psychology of Music*, D. Deutsch, ed., (2nd ed.). Academic Press.

Rosch, E. (1978). Principles of categorization. In : *Cognition and categorization*, B. Lloyd & L. Erlbaum, ed., pages 27–47. Hillsdale.

Schaeffer, P. (1966). *Traité des objets musicaux.* Seuil.

Schön, D., Ystad, S., Besson, M., & Kronland Martinet, R. (2006). An acoustical and cognitive approach to the semiotics of sound objects. In : *Music Perception and Cognition 9th International Conference on Music Perception and Cognition*, page pp 1862.

Schön, D., Ystad, S., Kronland Martinet, R., & Besson, M. (2009). The evocative power of sounds : conceptual priming between words and nonverbal sounds. *Journal of Cognitive Neuroscience*, pages 1–11.

Sciabica, J.-F., Bezat, M.-C., Roussarie, V., Kronland Martinet, R., & Ystad, S. (2010). Towards timbre modeling of sounds inside accelerating cars. In : *Auditory Display*, Ystad, Aramaki, Kronland Martinet, & Jensen, ed., Lecture Notes in Computer Science, pages 377–392. Springer Verlag/Heidelberg.

Seashore, C. E. (1938). *Psychology of Music.* McGraw-Hill - Reprinted 1967 by Dover Publications.

Shahin, A., Roberts, L. E., Pantev, C., Trainor, L. J., & Ross, B. (2005). Modulation of p2 auditory-evoked responses by the spectral complexity of musical sounds. *NeuroReport*, 16(16) :1781–1785.

Silva, F., Kergomard, J., Vergez, C., & Gilbert, J. (2008). Interaction of reed and acoustic resonator in clarinetlike systems. *Journal of the Acoustical Society of America*, 124(5) :3284–3295.

Van Petten, C. & Rheinfelder, H. (1995). Conceptual relationships between spoken words and environmental sounds : Event-related brain potential measures. *Neuropsychologia*, 33(4) :485–508.

Verron, C. (2010). *Synthèse Immersive de Sons d'Environnement*. PhD thesis, Université de la Méditerranée - Aix-Marseille II.

Weidenaar, R. (1995). *Magic Music from the Telharmonium*. The Scarecrow Press, Inc.

Widin, G. P. & Viemeister, N. F. (1979). Intensive and temporal effects in pure-tone forward masking. *The Journal of the Acoustical Society of America*, 66 :388–395.

Ystad, S. (1998). *Sound Modeling Using a Combination of Physical and Signal Models*. PhD thesis, Université de la Méditerrannée - Aix-Marseille II, Faculté des Sciences Luminy.

Ystad, S. (1999). De la facture informatique au jeu instrumental. In : *Les nouveaux gestes de la musique*, H. G. et Raphaël de Vivo, ed., pages 111–120. Parenthèses.

Ystad, S., Aramaki, M., Kronland Martinet, R., & Jensen, K. (2010). *Auditory display*. Springer Verlag Berlin Heidelberg. ISBN : 978-3-642-12438-9.

Ystad, S., Kronland Martinet, R., & Jensen, K. (2009). *Genesis of meaning in sound and music*. Springer Verlag Berlin Heidelberg. ISBN : 978-3-642-02517-4.

Ystad, S., Kronland Martinet, R., Schön, D., & Besson, M. (2008). Vers une approche acoustique et cognitive de la sémiotique des objets sonores. In : *Vers une sémiotique générale du temps dans les arts*, M. F. Emmanuelle Rix, ed., Musique/sciences, pages 79–90. Delatour.

Ystad, S., Magne, C., Farner, S., Pallone, G., Aramaki, M., Besson, M., & Kronland Martinet, R. (2007). Electrophysiological study of algorithmically processed metric/rhythmic variations in language and music. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007(30194) :13.

Ystad, S., Magne, C., Farner, S., Pallone, G., Pasdeloup, V., Kronland Martinet, R., & Besson, M. (2003). Influence of rhythmic, melodic, and semantic violations in language and music on the electrical activity in the brain. In : *Stockholm Music Acoustics Conference, (SMAC03)*.

Ystad, S. & Voinier, T. (2001a). Analysis-synthesis of flute sounds using a non-linear digital waveguide model. In : *International Computer Music Conference (ICMC).*

Ystad, S. & Voinier, T. (2001b). Analysis-synthesis of flute sounds with a looped non-linear model. In : *Workshop on Current Research Directions in Computer Music*, pages 200–208.

Ystad, S. & Voinier, T. (2001). A virtually real flute. *Computer Music Journal*, 25(2) :13–24.

Publications principales

Controlling the Perceived Material in an Impact Sound Synthesizer

Mitsuko Aramaki, *Member, IEEE*, Mireille Besson, Richard Kronland-Martinet, *Senior Member, IEEE*, and Sølvi Ystad

***Abstract*—In this paper, we focused on the identification of the perceptual properties of impacted materials to provide an intuitive control of an impact sound synthesizer. To investigate such properties, impact sounds from everyday life objects, made of different materials (wood, metal and glass), were recorded and analyzed. These sounds were synthesized using an analysis–synthesis technique and tuned to the same chroma. Sound continua were created to simulate progressive transitions between materials. Sounds from these continua were then used in a categorization experiment to determine sound categories representative of each material (called *typical* sounds). We also examined changes in electrical brain activity (using event related potentials (ERPs) method) associated with the categorization of these typical sounds. Moreover, acoustic analysis was conducted to investigate the relevance of acoustic descriptors known to be relevant for both timbre perception and material identification. Both acoustic and electrophysiological data confirmed the importance of damping and highlighted the relevance of spectral content for material perception. Based on these findings, controls for damping and spectral shaping were tested in synthesis applications. A global control strategy, with a three-layer architecture, was proposed for the synthesizer allowing the user to intuitively navigate in a "material space" and defining impact sounds directly from the material label. A formal perceptual evaluation was finally conducted to validate the proposed control strategy.**

***Index Terms*—Analysis–synthesis, control, event related potentials, impact sounds, mapping, material, sound categorization, timbre.**

I. Introduction

THE current study describes the construction of a synthesizer dedicated to impact sounds that can be piloted using high-level verbal descriptors referring to material categories (i.e., wood, metal and glass). This issue is essential for sound design and virtual reality where sounds coherent with visual scenes are to be constructed. Control strategies for synthesis (also called mapping) is an important issue that has interested the computer music community ever since it became possible to produce music with computers [1]. A large number of interfaces and control strategies have been proposed by several authors [2]–[10]. Most of these interfaces were designed for musical purposes and are generally not adapted to build environmental sounds used in sound design and virtual reality. As opposed to music-oriented interfaces that generally focus on the control of acoustic factors such as pitch, loudness, or rhythmic deviations, a more intuitive control based on verbal descriptors that can be used by non-experts is needed in these new domains. This issue requires knowledge on acoustical properties of sounds and how they are perceived. As a first approach towards the design of such an environmental sound synthesizer, we focus on the class of impact sounds and on the control of the perceived material. In particular, our aim is to develop efficient mapping strategies between words referring to certain material categories (i.e., wood, metal and glass) and signal parameters to allow for an intuitive sound synthesis based on a smaller number of control parameters.

To point out perceptual properties that characterize the categories, a listening test was conducted. Stimuli were created first by recording impact sounds from everyday life objects made of different materials. Then, these recorded sounds were synthesized by analysis–synthesis techniques and tuned to the same chroma. Finally, we created continua from the tuned sounds to simulate progressive transitions between the categories by interpolating signal parameters. The use of sound continua was of interest to closely investigate transitions and limits between material categories. Sounds from these continua were used in a categorization task so as to be classified by participants as Wood, Metal, or Glass. From the percentage of responses, we determined sound categories representative of each material (called sets of *typical* sounds).

Then, we examined the acoustic characteristics that differ across typical Wood, Metal, and Glass sounds. For this purpose, we considered acoustic descriptors known to be relevant both for timbre perception and for material identification. Previous studies on the perception of sound categories have mainly been based on the notion of timbre. Several authors have used dissimilarity ratings to identify timbre spaces in which sounds from different musical instruments can be distinguished [11]–[14]. They found correlations between dimensions of these timbre spaces and acoustic descriptors such as attack time (the way the energy rises at the sound onset), spectral bandwidth (spectrum spread), or spectral centroid (center of gravity of the spectrum). More recently, roughness (distribution of interacting frequency components within the limits of a critical band) was considered

Manuscript received April 27, 2009; revised October 08, 2009; accepted March 18, 2010. Date of publication April 08, 2010; date of current version October 27, 2010. This work was supported by the Human Frontier Science Program under Grant HFSP #RGP0053 to M. Besson and a grant from the French National Research Agency (ANR, JC05-41996, "senSons") to S. Ystad. The work of M. Aramaki was supported by a postdoctoral grant, first from the HFSP and then from the ANR. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gaël Richard.

M. Aramaki and M. Besson are with CNRS-Institut de Neurosciences Cognitives de la Méditerranée, 13402 Marseille Cedex 20 France, and also with Aix-Marseille-Université, 13284 Marseille Cedex 07 France (e-mail: aramaki@incm.cnrs-mrs.fr; besson@incm.cnrs-mrs.fr).

R. Kronland-Martinet and S. Ystad are with the CNRS-Laboratoire de Mécanique et d'Acoustique, 13402 Marseille Cedex 20 France (e-mail: kronland@lma.cnrs-mrs.fr; ystad@lma.cnrs-mrs.fr).

Digital Object Identifier 10.1109/TASL.2010.2047755

as a relevant dimension of timbre since it is closely linked to the concept of consonance in a musical context [15], [16]. In the case of impact sounds, the perception of material seems mainly to correlate with the frequency-dependent damping of spectral components [17], [18] ([19], [20] in the case of struck bars), due to various loss mechanisms. Interestingly, damping remains a robust acoustic descriptor to identify macro-categories (i.e., between wood–Plexiglas and steel–glass categories) across variations in the size of objects [21]. From an acoustical point of view, a global characterization of the damping can be given by the sound decay measuring the decrease in sound energy as a function of time.

The above-mentioned timbre descriptors, namely attack time, spectral bandwidth, roughness, and normalized sound decay, were considered as potentially relevant signal features for the discrimination between sound categories. An acoustic analysis was conducted on these descriptors to investigate their relevance. At this stage, it is worth mentioning that signal descriptors that are found to be significant in traditional timbre studies may not be directly useful in the case of sound synthesis and control. Some descriptors might not give access to a sufficiently fine control of the perceived material. It might be necessary to act on a combination of descriptors. To more deeply investigate perceptual/cognitive aspects linked to the sound categorization, we exploited electrophysiological measurements for synthesis purposes since they provide complementary information regarding the nature of sound characteristics that contribute to the differentiation of material categories from a perceptual/cognitive point of view. In particular, we examined changes in brain electrical activity [using event related potentials (ERPs)] associated with the perception and categorization of typical sounds (we refer the reader to a related article for more details [22]).

Based on acoustic and electrophysiological results, sound characteristics relevant for an accurate evocation of material were determined and control strategies related to physical and perceptual considerations were proposed. The relevance of these strategies in terms of an intuitive manipulation of parameters was further tested in synthesis applications. High-level control was achieved through a calibration process to determine the range values of the damping parameters specific to each material category. In particular, the use of sound continua in the categorization experiment highlighted transition zones between categories that allowed for continuous control between different materials.

The paper is organized as follows: first the sound categorization experiment with stimuli construction and results is presented. Statistical analyses are further carried out on the set of sounds defined as *typical* to determine the acoustic descriptors that best discriminate sound categories. Then, sound characteristics that are relevant for material perception are obtained from physical considerations, timbre investigations and electrophysiological measurements. Control strategies allowing for an intuitive manipulation of these sound characteristics, based on these findings and on our previous works [23]–[25], are proposed in a three-layer control architecture providing the synthesis of impact sounds directly from the material label. A formal perceptual evaluation of the proposed control strategy is finally presented.

II. Sound Categorization Experiment

A. Participants

Twenty-five participants (13 women and 12 men, 19 to 35 years old, mean age $= 22.5$) were tested in this experiment that lasted for about one hour. They were all right-handed, nonmusicians (no formal musical training), had normal audition and no known neurological disorders. They all gave written consent and were paid to participate in the experiment.

B. Stimuli

We first recorded 15 sounds by impacting everyday life objects made of three different materials (wooden beams, metallic plates, glass bowls) that are five sounds per material. Synthetic versions of these recorded sounds were generated by an analysis–synthesis process and tuned to the same chroma. Then, we created J-step sound continua that simulate progressive transitions between two sounds of different materials by acting on amplitudes and damping parameters. The different stages of the stimuli construction are detailed below.

1) Analysis–Synthesis of Natural Sounds: Recordings of natural sounds were made in an acoustically treated studio of the laboratory using a microphone placed 1 m from the source. The objects from different materials were impacted by hand. We tried to control the impact on the object by using the same drumstick and the same impact force. The impact position on the different objects was chosen so that most modes were excited (near the center of the object for wooden beams and metallic plates; near the rim for glass bowls). Sounds were digitally recorded at 44.1-kHz sampling frequency.

From a physical point of view, the vibrations of an impacted object (under free oscillations) can generally be modeled as a sum of M exponentially damped sinusoids:

$$s(t) = \theta(t) \sum_{m=1}^{M} A_m \sin(\omega_m t + \Phi_m) e^{-\alpha_m t} \tag{1}$$

where $\theta(t)$ is the Heaviside function and the parameters A_m, α_m, ω_m, and Φ_m, the amplitude, damping coefficient, frequency, and phase of the mth component, respectively. Based on the signal model corresponding to (1), we synthesized the recorded sounds at the same sampling frequency. Several different techniques allow precise estimating the signal parameters $\{A_m, \alpha_m, \omega_m\}_{m=1,\ldots,M}$ based on high-resolution analysis such as the Steiglitz–McBride technique [26] or more recently Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT), MUltiple SIgnal Classification (MUSIC), Least Squares or Maximum-Likehood techniques [27]–[30] (see also [31], [32]). These latter methods provide an accurate estimation and can be used to conduct spectral analysis. We here used a simplified analysis technique based on discrete Fourier transform (DFT) since we aimed at reproducing the main characteristics of the original sounds in terms of perceived material rather than achieving a perfect resynthesis.

The number of components M to synthesize was estimated from the modulus of the spectral representation of the signal.

Only the most prominent components, which amplitudes were larger than a threshold value fixed at 30 dB below the maximum amplitude of the spectrum, were synthesized. In addition, to keep the broadness of the original spectrum, we made sure that at least the most prominent component in each critical bandwidth was synthesized. Since Wood and Glass sounds had relatively poor spectra (i.e., few components), most of the components were synthesized. By contrast, Metal sounds had rich and broadband spectra. Some components were due to the nonlinear vibrations of the impacted object (favored by a low dissipation for Metal) and could not be reproduced by the signal model that only considers linear vibrations. Thus, the number of components for synthetic Metal sounds were generally inferior to the number of components of the original sound.

The frequency values ω_m were directly inferred from the abscissa of the local maxima corresponding to the prominent components. Since the spectrum was obtained by computing a fast Fourier transform (FFT) over 2^{16} samples, the frequency precision of each component was equal to 0.76 Hz $(= 44\,100/2^{16})$. Each component m was isolated using a gaussian window centered on the frequency ω_m. The frequency bandwidth of the gaussian window was adapted to numerically minimize the smoothing effects and to avoid the overlap of two successive components which causes interference effects. The gaussian window presents the advantage of preserving the exponential damping when convolved with an exponentially damped sine wave. Then, the analytic signal $\hat{s}_m(t)$ of the windowed signal was calculated using the Hilbert transform and the modulus of $\hat{s}_m(t)$ was modeled by an exponentially decaying function

$$|\hat{s}_m(t)| = A_m e^{-\alpha_m t} \tag{2}$$

Thus, by fitting the logarithm of $|\hat{s}_m(t)|$ with a polynomial function of degree 1 at best in a least-squares sense, the amplitude A_m was inferred from the ordinate at the origin while the damping coefficient α_m was inferred from the slope. Finally, the phases Φ_m were set to 0 for all components. This choice is commonly adopted in synthesis processes since it avoids undesirable clicks at sound onset. It is worth noticing that this phase adjustment does not affect the perception of the material because phase relationships between components mainly reflect the position of the microphone relative to the impacted object.

2) Tuning: The pitches of the 15 synthetic sounds (five per material category) differed since they resulted from impacts on various objects. Consequently, sounds were tuned to the same chroma to minimize pitch variations. Tuning is needed to build homogeneous sound continua with respect to pitch (Section II-B4) and to accurately investigate acoustic descriptors (Section III). In particular, the relationships between descriptors will be better interpreted if they are computed on a set of tuned sounds with normalized pitches rather than on a set of sounds with various pitch values.

We first defined the initial pitch of the sounds from informal listening tests: four participants (different from those who participated in the categorization experiment) listened to each sound and were asked to evaluate the pitch by playing the matching note on a piano keyboard. For each sound, the pitch was defined by the note that was most often associated with the sound. Thus, we defined the pitches $G\sharp 3$ (fundamental frequency of 415.30 Hz), $C\sharp 3$ (277.18 Hz), $F\sharp 3$ (369.99 Hz), $C\sharp 3$, and $C\sharp 3$ for the 5 Wood sounds; $A3$ (440.00 Hz), $F\sharp 3$, $D5$ (1174.65 Hz), $E4$ (659.25 Hz), and $E3$ (329.62 Hz) for the 5 Metal sounds, and $C5$ (1046.50 Hz), $E6$ (2637.02 Hz), $C\sharp 6$ (2217.46 Hz), $D5$, and $F5$ (1396.91 Hz) for the 5 Glass sounds. Then, we tuned the sounds to the closest note C with respect to the initial pitch to minimize signal transformations applied on the sounds: Wood sounds were tuned to the pitch C3, Metal sounds to C3 and C4 and Glass sounds to C5 and C6. Therefore, sounds differed by 1, 2, or 3 octaves depending upon the material. Based upon previous results showing high similarity ratings for tone pairs that differed by octaves [33], an effect known as the octave equivalence, we assume that the octave differences between sounds belonging to a same category should have little influence on sound categorization.

In practice, tuned sounds were generated using the previous synthesis technique [(1)]. The amplitudes and phases of components were kept unchanged but the frequencies (noted $\tilde{\omega}_m$ for tuned sounds) and damping coefficients (noted $\tilde{\alpha}_m$) were recalculated as follows. The tuned frequencies $\tilde{\omega}_m$ were obtained by transposing original ones $\{\omega_m\}_{m=1,\ldots,M}$ with a dilation factor η defined from the fundamental frequency values (in Hz), noted F and $\tilde{F}$, of the sound pitches before and after tuning, respectively,

$$\tilde{\omega}_m = \eta\omega_m \quad \text{with} \quad \eta = \frac{\tilde{F}}{F}. \tag{3}$$

The damping coefficient $\tilde{\alpha}_m$ of each tuned component was recalculated by taking into account the frequency-dependency of the damping. For instance, it is known that in case of wooden bars, the damping coefficients increase with frequency following an empirical expression of a parabolic form where parameters depend on the wood species [34]–[36]. To achieve our objectives, we defined a general expression of a damping law $\alpha(\omega)$ chosen as an exponential function

$$\alpha(\omega) = e^{(\alpha_G + \alpha_R \omega)}. \tag{4}$$

The exponential expression presents the advantage of easily fitting various and realistic damping profiles with a reduced number of parameters. $\alpha(\omega)$ is defined by two parameters α_G and α_R characteristic of the intrinsic properties of the material. The parameter α_G reflects global damping and the parameter α_R reflects frequency-relative damping (i.e., difference between high-frequency component damping and low-frequency component damping). Thus, a damping law $\alpha(\omega)$ was estimated on the original sound by fitting the damping coefficients $\{\alpha_m\}_{m=1,\ldots,M}$ with the (4) at best in a least-squares sense. Then, the damping coefficient $\tilde{\alpha}_m$ of the mth tuned component was recalculated according to this damping law (see also [37])

$$\tilde{\alpha}_m = \alpha(\tilde{\omega}_m). \tag{5}$$

3) Gain Adjustment: Sounds were equalized by gain adjustments to avoid the influence of loudness in the categorization judgments. The gain adjustments were determined on the basis of a pretest with four participants (different from those who participated in the categorization experiment). They were asked to balance the loudness level of the tuned sounds. These tuned sounds were previously normalized by a gain of reference $\Gamma_0 = 1.5 \times A$ with A corresponding to the largest value of the maxima of the signal modulus among the 15 tuned sounds. The coefficient 1.5 is a safety coefficient commonly used in gain adjustment tests to avoid the saturation of the signals after the adjustment. The gain values Γ to be applied on the 5 Wood sounds were equal to [70, 20, 30, 15, 30], on the 5 Metal sounds were equal to [3.5, 1.1, 1, 1.5, 1.3] and on the five Glass sounds were equal to [35, 15, 15, 30, 10].

Finally, the four participants were asked to evaluate the final sounds in terms of perceived material. Results showed that sounds were categorized in the same material category as the original sounds by all participants thereby showing that the main characteristics of the material were preserved.

4) Sound Continua: To closely investigate transitions between material categories, we created 15 J-step sound continua noted Ω_i with five continua for each material transition. The five Wood-Metal continua were indexed from Ω_1 to Ω_5, the five Wood–Glass continua from Ω_6 to Ω_{10} and finally, the five Glass-Metal continua from Ω_{11} to Ω_{15}. Each continuum was composed of 22 hybrid sounds ($J = 22$) that were obtained by mixing the spectra and by interpolating the damping laws of the two extreme sounds. We chose to mix spectra to fix the values of the frequency components which allows minimizing pitch variations across sounds within a continuum (it is known that shifting components modifies pitch). We chose to interpolate damping laws to gradually modify the damping that conveys fundamental information on material perception. Thus, the sound H_j(t) at step j of the continuum is expressed by

$$H_j(t) = \gamma_1(j)\frac{\Gamma_1}{\Gamma_0}\sum_{m=1}^{M} A_m \sin(\omega_m t)e^{-\alpha^j(\omega_m)t} + \gamma_2(j)\frac{\Gamma_2}{\Gamma_0}\sum_{n=1}^{N} A_n \sin(\omega_n t)e^{-\alpha^j(\omega_n)t} \quad (6)$$

where $\{A_m, \omega_m\}_{m=1,\ldots,M}$ and $\{A_n, \omega_n\}_{n=1,\ldots,N}$ correspond to the sets of amplitudes and frequencies of the two extreme sounds and j varies from 1 to 22. The gains Γ_1 and Γ_2 correspond to the gains of the extreme sounds defined from the gain adjustment test according to a gain of reference Γ_0 (see Section II-B2). The gains $\gamma_1(j)$ and $\gamma_2(j)$ vary at each step j on a logarithmic scale, according to the dB scale

$$\begin{aligned}\gamma_1(j) &= 1 - \frac{\log(j)}{\log(J)}\\ \gamma_2(j) &= 1 - \frac{\log(J-j+1)}{\log(J)}.\end{aligned} \quad (7)$$

The damping variation along the continua is computed by interpolating the damping parameters α_G and α_R of the damping law [defined in (4)] estimated on the two extreme sounds (located at step $j = 1$ and $j = 22$, respectively), leading to the determination of a hybrid damping law $\alpha^j(\omega)$ that progressively varies at each step j of the continuum (see Fig. 1)

$$\alpha^j(\omega) = e^{(\alpha_G^j + \alpha_R^j \omega)} \quad (8)$$

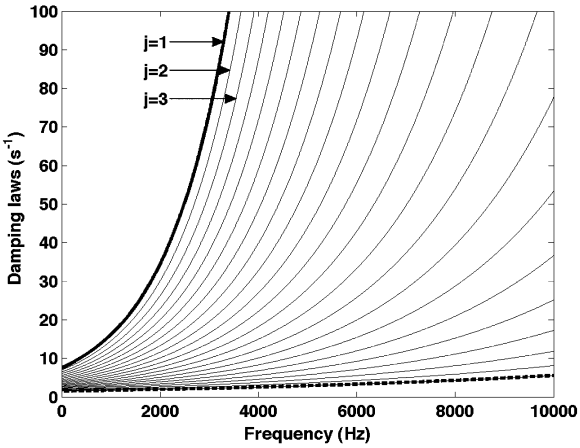


Fig. 1. Damping laws $\alpha^j(\omega)$ for $j = 1, \ldots, 22$ as a function of frequency corresponding to a Wood–Metal continuum. Bold curves correspond to damping laws of Wood (in bold plain) and Metal (in bold dashed) sounds at the extreme positions.

with

$$\begin{aligned}\alpha_G^j &= \left(\alpha_G^{22} - \alpha_G^1\right)\frac{j-1}{J-1} + \alpha_G^1\\ \alpha_R^j &= \left(\alpha_R^{22} - \alpha_R^1\right)\frac{j-1}{J-1} + \alpha_R^1.\end{aligned} \quad (9)$$

The use of an interpolation process on the damping allowed for a better merging between the extreme sounds since the spectral components of the two spectra are damped following the same damping law $\alpha^j(\omega)$. As a consequence, hybrid sounds (in particular, at centered positions of the continua) differed from sounds obtained by only mixing the extreme sounds.

The obtained sounds had different signal lengths (Metal sounds are longer than Wood or Glass sounds). To restrain the lengths to a maximum of 2 seconds, sound amplitudes were smoothly dropped off by multiplying the temporal signal with the half decreasing part of a Hann window.

A total of 330 sounds were created. The whole set of sounds are available at [38]. The averaged sound duration was 861 ms for all sounds and 1053 ms in the Wood–Metal continua, 449 ms in the Wood-Glass continua, and 1081 ms in the Glass–Metal continua.

C. Procedure

The experiment was conducted in a quiet Faradized (electrically shielded) room. Sounds were presented once (i.e., no repetition of the same sound) in random order through one loudspeaker (Tannoy S800) located 1 m in front of the participant. Participants were asked to categorize sounds as Wood, Metal, or Glass, as fast as possible, by pressing one response button out of three on a three-buttons response box[1] (right, middle, and left buttons; one button per material category label). The association between response buttons and material categories was balanced across participants to avoid any bias linked with the

[1] Since participants were not given the option to choose that sounds did not belong to either one of these three categories, results may be biased, but this potential ambiguity would be raised only for intermediate sounds.

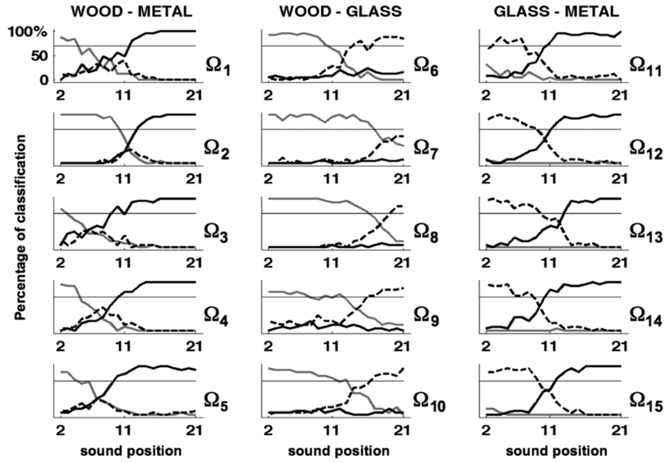


Fig. 2. Percentage of classification as Wood (gray curves), Metal (black curves), or Glass (dashed curves) for each sound as a function of its position j on the continuum for the 15 continua Ω_i. Sounds were considered as typical if they were classified in one category by more than 70% of participants (threshold represented by an horizontal line). No data were collected for extreme sounds ($j = 1$ and $j = 22$) since they were used in the training session.

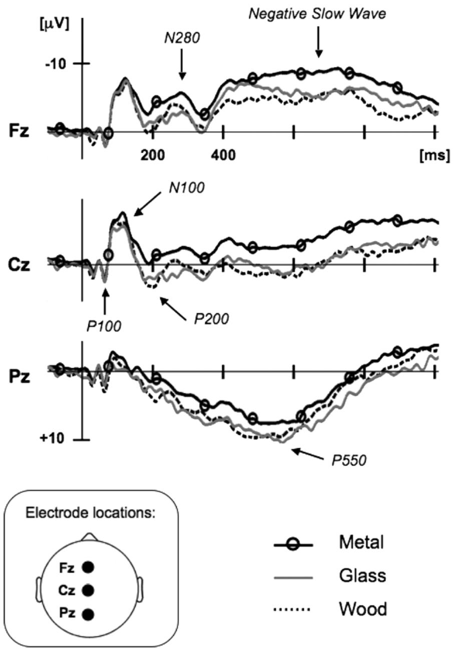


Fig. 3. ERPs to typical sounds of Wood (dotted), Metal (circle marker), and Glass (gray) at electrodes (Fz, Cz, Pz) along the midline of the scalp (see [22] for lateral electrodes). The amplitude (in microvolts) is represented on the ordinate and negativity is up. The time from sound onset is on the abscissa (in milliseconds).

order of the buttons. A row of 4 crosses, i.e., "XXXX," was presented on the screen 2500 ms after sound onset during 1500 ms to give participants time to blink. The next sound was presented after a 500-ms silence. Participants' responses were collected for all sounds except for the extreme sounds since they were used in the training session. The electrical brain activity (Electroencephalogram, EEG) was recorded continuously for each participant during the categorization task.

D. Results

1) Behavioral Data: Percentages of categorization as Wood, Metal, or Glass were obtained for each sound by averaging responses across participants. Fig. 2 allows visualization of these data as a function of sound position along the continua. From these data, we determined a set of *typical* sounds for each material category: sounds were considered as *typical* if they were categorized as Wood, Metal, or Glass by more than 70% of participants (we refer the reader to [39] for more details on the determination of this threshold of percentage value). In addition, sound positions delimiting categories along the continua can be defined from the position ranges of typical sounds. Note that due to the different acoustic characteristics of sounds, the category limits are not located at the same position for all continua (see Fig. 2).

2) Electrophysiological Data: We examined ERPs time-locked to sound onset to analyze the different stages of information processing as they unfold in real time.[2] ERP data were averaged separately for typical sounds of each material category (Fig. 3). We summarize here the main findings (we refer the reader to a related article for more details [22]).

Typical sounds from each category elicited small P100 components, with maximum amplitude around 65-ms post-sound onset, large N100, and P200 components followed by N280 components and Negative Slow Wave (NSW) at fronto-central sites or large P550 components at parietal sites. Statistical analyses revealed no significant differences on the P100 and N100 components as a function of material categories. By contrast, they showed that typical Metal sounds elicited smaller P200 and P550 components, and larger N280 and NSW components than typical sounds from Wood and Glass. From an acoustic point of view, Metal sounds have richer spectra and longer durations (i.e., lower damping) than Wood and Glass sounds. The early differences on the P200 components most likely reflect the processing of spectral complexity (see [42] and [43]) while the later differences on the N280 and on the NSW are likely to reflect differences in sound duration (i.e., differences in damping; see [44] and [45]).

III. Acoustic Analysis

The typical sounds as defined based upon behavioral data (Section II-D1) form a set of sounds representative of each material category. To characterize these typical sounds from an acoustical point of view, we investigated the following descriptors: attack time (AT), spectral bandwidth (SB), roughness (R), and normalized sound decay (α) that are defined below. Then, we examined the relationships between acoustic descriptors and their relevance to discriminate material categories.

A. Definition of Acoustic Descriptors

Attack time is a temporal timbre descriptor which characterizes signal onset. It is defined by the time (in second) necessary

[2] The ERPs elicited by a stimulus (a sound, a light, etc.) are characterized by a succession of positive (P) and negative (N) deflections relative to a baseline (usually measured within the 100 ms or 200 ms that precedes stimulus onset). These deflections (called components) are characterized by their polarity, their latency of maximum amplitude (relative to stimulus onset), their distribution across different electrodes located at standard positions on the scalp and by their functional significance. Typically, the P100, N100, and P200 components reflect the sensory and perceptual stages of information processing, and are obligatory responses to the stimulation [40], [41]. Then, depending on the experimental design and on the task at hand, different late ERP components are elicited (N200, P300, N400, etc.).

for the signal energy to raise from a threshold level to the maximum energy in the temporal envelope (for percussive sound) or to the sustained part (for a sustained sound with no decay part) [46], [47]. Different values have been proposed in the literature for both minimum and maximum thresholds. For our concern, we chose to compute the attack time from 10% to 90% of the maximum amplitude of the temporal envelope as in [48]. This descriptor is known to be relevant to distinguish different classes of instrumental sounds. For instance, sounds from percussive and woodwind instruments have respectively short and long AT.

Spectral bandwidth (in Hz), commonly associated with the spectrum spread, is defined by [49]

$$SB = \frac{1}{2\pi}\sqrt{\frac{\sum_k |\hat{s}(k)|\left(\omega(k) - 2\pi \times SC\right)^2}{\sum_k |\hat{s}(k)|}} \tag{10}$$

where SC is the spectral centroid (in Hz) defined by [50]

$$SC = \frac{1}{2\pi}\frac{\sum_k \omega(k)\,|\hat{s}(k)|}{\sum_k |\hat{s}(k)|} \tag{11}$$

and where ω represents frequency, $\hat{s}$ the Fourier transform of the signal estimated using the FFT algorithm and k the FFT bin index. The FFT was calculated on 2^{16} samples.

Roughness (in asper) is commonly associated with the presence of several frequency components within the limits of a critical band. From a perceptual point of view, roughness is correlated with tonal consonance based on results from experiments on consonance judgments conducted by [51]. From a signal point of view, [52] have shown that roughness and fluctuation strength are proportional to the square of the modulation factor of an amplitude modulated pure tone. We computed roughness based on Vassilakis's model by summing up the partial roughness r_{mn} for all pairs of frequency components contained in the sound [53]

$$r_{mn} = 0.5(A_m A_n)^{0.1} \times \left(\frac{2\min(A_m, A_n)}{A_m + A_n}\right)^{3.11} \times \left(e^{-3.5v|\omega_m - \omega_n|} - e^{-5.75v|\omega_m - \omega_n|}\right) \tag{12}$$

with

$$v = \frac{0.24}{0.0207 \times \min(\omega_m, \omega_n) + 2\pi \times 18.96}$$

and where A_m and A_n are amplitudes and ω_m and ω_n are the frequencies of components m and n, respectively.

Finally, the sound decay D (in s^{-1}) quantifies the amplitude decrease of the whole temporal signal and globally characterizes the damping in the case of impact sounds. In particular, D approximately corresponds to the decay of the spectral component with the longest duration (i.e., generally the lowest frequency one). The sound decay is directly estimated by the slope of the logarithm of the temporal signal envelope. This envelope is given by calculating the analytic signal using the Hilbert transform and by filtering the modulus of this analytic signal using a second-order low-pass Butterworth filter with cutoff frequency of 50 Hz [36]. Since damping is frequency dependent (Section II-B1), sound decay depends on the spectral content of the sound. Consequently, we considered a normalized sound decay denoted α with respect to the spectral localization of the energy and we defined the dimensional descriptor α as the ratio of the sound decay D to the SC value

$$\alpha = \frac{D}{SC}. \tag{13}$$

TABLE I
COEFFICIENTS OF DETERMINATION BETWEEN THE ATTACK TIME AT, THE SPECTRAL BANDWIDTH SB, THE ROUGHNESS R, AND THE NORMALIZED SOUND DECAY α. SINCE THE MATRIX IS SYMMETRIC, ONLY THE UPPER PART IS REPORTED. THE P-VALUES ARE ALSO REPORTED BY *** ($p < .001$) WHEN COEFFICIENTS ARE SIGNIFICANT (WITH BONFERRONI ADJUSTMENT)

	AT	SB	R	α
AT	1	0.05***	0.07***	0
SB	–	1	0.25***	0.02
R	–	–	1	0.23***
α	–	–	–	1

B. Relationships Between Acoustic Descriptors

As a first step, we examined the relationships between the acoustic descriptors estimated on typical sounds. Table I shows the coefficients of determination that are the square of the Bravais–Pearson coefficients between pairs of descriptors. We found highest significant correlation (although not high in terms of absolute value) between the two spectral descriptors SB and R. Lowest correlations were found between AT and the other ones, reflecting the fact that sound onset has little influence on the spectral characteristics and does not depend on the decaying part of the sound (described by α).

Second, a principal component analysis (PCA) was conducted on standardized values of acoustic descriptors (i.e., values centered on the mean value and scaled by the standard deviation value). Results showed that the first two principal components explained about 72% of the total variance (the first component alone explained about 48%). As shown in Fig. 4, the first component was mainly correlated to the spectral descriptors (SB and R) and the second component to the temporal descriptors (AT and α). Thus, PCA revealed that sounds could reliably be represented in a reduced bi-dimensional space which orthogonal axes are mainly correlated to spectral (Component I) and temporal descriptors (Component II), respectively. This result confirmed that spectral and temporal descriptors bring complementary information on the sound characterization from an acoustic point of view.

C. Discrimination Between Material Categories

We examined the relevance of acoustic descriptors to discriminate material categories using a discriminant canonical analysis. This analysis was conducted using Materials (Wood, Metal, and Glass) as groups and standardized values of acoustic descriptors $\{\mathrm{AT}, \mathrm{SB}, \mathrm{R}, \alpha\}$ as independent variables. Since three sound categories were considered, two discriminant functions that allow for the clearest separation between sound categories were computed (the number of discriminant functions is equal to the number of groups minus one). These

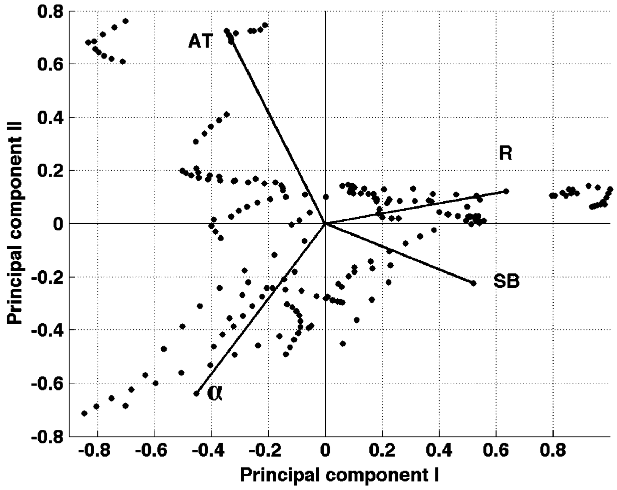


Fig. 4. Biplot visualization: the observations corresponding to typical sounds are represented by dots and the acoustic descriptors by vectors. The contribution of descriptors to each Principal Component (PC) can be quantified by the R^2 statistics given by a regression analysis: Attack time AT ($R^2 = .23$ for PC I and $R^2 = .52$ for PC II), Spectral bandwidth SB ($R^2 = .52$ for PC I and $R^2 = .05$ for PC II), Roughness R ($R^2 = .77$ for PC I and $R^2 = .01$ for PC II) and Normalized sound decay α ($R^2 = .39$ for PC I and $R^2 = .40$ for PC II).

functions C_1 and C_2 were expressed as a combination of the independent variables

$$C_1 = 1.04\alpha + 0.76SB - 0.58R + 0.47AT$$
$$C_2 = 0.70R - 0.15SB - 0.56AT + 0.38\alpha. \quad (14)$$

The Wilks's Lambda show that both functions C_1 (Wilks's $\Lambda = .15$; $\chi^2 = 366.65$; $p < .001$) and C_2 (Wilks's $\Lambda = .87$; $\chi^2 = 28.02$; $p < .001$) are significant. The first function C_1 explains 96% of the variance (coefficient of determination $= 0.82$) while the second function C_2 explains the remaining variance (coefficient of determination $= 0.13$). The coefficient associated with each descriptor indicates its relative contribution to the discriminating function. In particular, the first function C_1 is mainly related to α and allows clear distinction particularly between typical Wood and Metal sounds as shown in Fig. 5. This result is in line with previous studies showing that damping is a fundamental cue in the perception of sounds from impacted materials (see the Introduction). The second axis C_2 is mainly related to the spectral descriptor R and allows for a distinction of Glass sounds.

IV. Control Strategy for the Synthesizer

Results from acoustic and electrophysiological data are now discussed in the perspective of designing an intuitive control of the perceived material in an impact sound synthesizer. In particular, we aim at determining relevant sound characteristics for an accurate evocation of different materials and at proposing intuitive control strategies associated with these characteristics. In practice, the synthesis engine and the control strategies were implemented using Max/MSP [54] thereby allowing for the manipulation of parameters in real-time and consequently, providing an easy way to evaluate the proposed controls. The observations and conclusions from these synthesis applications are also reported.

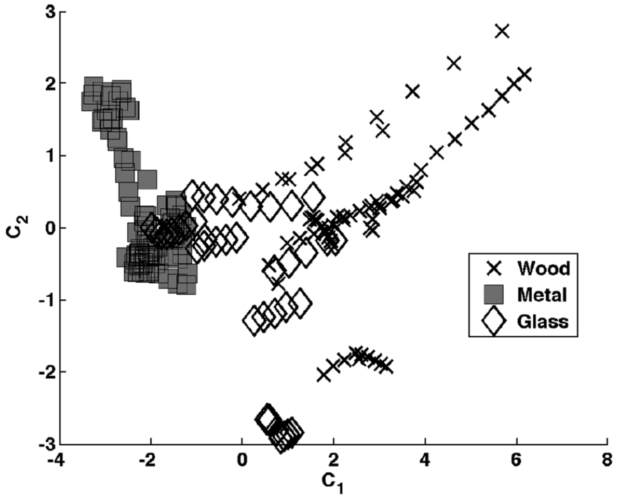


Fig. 5. Scatter plot of the canonical variables allowing the clearest separation between typical Wood (×), Metal (□), and Glass (◇) sound categories.

A. Determination of Sound Characteristics

As a starting point, results from acoustic analysis revealed that α (characterizing the damping) was the most relevant descriptor to discriminate material categories, therefore confirming several findings in the literature on the relevance of damping for material identification (see the Introduction). Thus, damping was kept as a relevant sound characteristic to control in the synthesizer.

Furthermore, acoustic analysis showed that in addition to the damping, a second dimension related to spectral characteristics of sounds was significant, in particular for the distinction between Glass and Metal sounds. Interestingly, this result is supported by electrophysiological data that revealed ERP differences between Metal on one side and both Glass and Wood sounds on the other side. Since these differences were interpreted as reflecting processing of sound duration (related to damping) and spectral content, ERP data showed the relevance of both these aspects for material perception. Thus, from a general point of view, it is relevant to assume that material perception seems to be guided by additional cues (other than damping) that are most likely linked to the spectral content of sounds. This assumption is in line with several studies showing the material categorization can be affected by spectral characteristics, in particular, Glass being associated with higher frequencies than Metal sounds [20], [55].

In line with this assumption, synthesis applications confirmed that damping was relevant but was not in some cases sufficient to achieve material categories. For instance, it was not possible to transform a given Wood or Glass sound into a Metal sound by only applying a Metal damping on a Wood or Glass spectrum. The resulting sound did not sound metallic enough. It was therefore necessary to also modify the spectral content, and in particular the spectral distribution, to emphasize the metallic aspect of the sounds (examples in [38]). We found another limitation of damping in the case of Glass synthesis. Indeed, Glass sounds match a wide range of damping values (from highly damped jar sounds to highly resonant crystal glass sounds) and are most often characterized by a sparse distribution of spectral components (i.e., few distinct components). These observations indicated that the perceptual distinction between Glass and Metal may be due to the typical dissonant aspect of Metal and can be accurately reflected by the roughness that was highlighted as the most relevant descriptor in the acoustic analysis after the damping. Thus, we concluded on the necessity to take into account a control of the spectral shape, in addition to the control of damping, for a more accurate evocation of the perceived material.

Besides, electrophysiological data provided complementary information regarding the temporal dynamics of the brain processes associated with the perception and categorization of typical sounds. First, it is known that P100 and N100 components are influenced by variations in sound onset parameters [56]. The lack of differences on these ERP components is taken to indicate similar brain processes for all typical sounds, showing that the information of the perceived material does not lie in the sound onset. As a synthesis outcome, it means that a modification of sound onset does not affect the nature of the perceived material and consequently, that AT is not a relevant parameter for the control of the perceived material. Second, it is known that N100 component is also influenced by pitch variations [40]. While octave differences were largest between Glass and the other two categories, the lack of differences on N100 component is taken to indicate that pitch may not affect sound categorization. Thus, electrophysiological data support our previous assumption concerning the weak influence linked to the octave differences on sound categorization (Section II-B2).

Based on these considerations, damping and spectral shaping were determined as relevant sound characteristics for material perception. Control strategies associated with these characteristics are proposed and detailed in the following sections.

B. Control of damping

The control of damping was designed by acting on parameters α_G and α_R of the damping law (4). This control gave an accurate manipulation of sound dynamics (time evolution) with a reduced number of control parameters. In particular, the parameter α_G governed the global sound decay (quantified by the descriptor D, Section III) and the parameter α_R allowed controlling the damping differently for high- and low-frequency components. This control made it possible to synthesize a wide variety of realistic sounds. Since from a physical point of view, high frequency components generally are more heavily damped than low frequency ones, we expected both parameters α_G and α_R to be positive in the case of natural sounds.

C. Control of Spectral Shaping

We here propose two control strategies. The first one relied on spectral dilation and was based on physical considerations, in particular on the dispersion phenomena. The second control was based on amplitude and frequency modulations and relied on adding components to the original spectrum. This latter control had a smaller influence on pitch compared with the first control since the original spectrum is not distorted. It also has interesting perceptual consequences. For instance, by creating components within a specific frequency band (i.e., critical band of hearing), this control specifically influenced the perception of roughness that was highlighted as a relevant acoustic descriptor in the acoustic analysis (Section III). These two control strategies are detailed in the following sections.

1) Control by Spectral Dilation: From the analysis of natural sounds, and from physical models describing wave propagation in various media (i.e., various physical materials), two important phenomena can be observed: dispersion and dissipation [32], [57]. Dissipation is due to various loss mechanisms and is directly linked to the damping parameters (α_G and α_R) as described above. Dispersion is linked to the fact that the wave propagation speed varies with respect to frequency. This phenomenon occurs when the phase velocity of a wave is not constant and introduces inharmonicity in the spectrum of the corresponding sound. An example of dispersive medium is the stiff string for which the mth partial is not located at $m\omega_1$ but at $m\omega_1\sqrt{1+\beta m^2}$ where ω_1 is the fundamental frequency and β the coefficient of inharmonicity depending on the physical parameters of the string [58]. We based our first spectral shaping strategy on the spectral dilation defined by

$$\tilde{\omega}_m = W(\breve{\omega}_{\min}, \breve{\omega}_{\max}, \omega)\breve{\omega}_m + (1 - W(\omega_{\min}, \omega_{\max}, \omega))\,\omega_m \tag{15}$$

where W is a window function (defined later in the text) and

$$\breve{\omega}_m = S_G\omega_m\sqrt{1 + S_R\left(\frac{\omega_m}{\omega_1}\right)^2} \tag{16}$$

with ω_m and $\tilde{\omega}_m$ that correspond to the frequency of the initial and shifted component of rank m, respectively. Equation (16) is a generalization of the inharmonicity law previously defined for stiff strings so that the expression is not limited to harmonic sounds but can be applied to any set of frequencies. S_G and S_R are defined as the global and relative shaping parameters, respectively. Ranges of S_G and S_R are constrained so that $\breve{\omega}_m$ are real-valued and $\breve{\omega}_m > \omega_1$ for all $m = 1, \ldots, M$ with M the number of components. Thus, S_R should be lower bounded

$$\min S_R = -\frac{1}{M^2} \tag{17}$$

and S_G should satisfy

$$S_G\frac{\omega_m}{\omega_1}\sqrt{1 + S_R\left(\frac{\omega_m}{\omega_1}\right)^2} > 1 \quad \text{for all } m = 1, \ldots, M. \tag{18}$$

A window function $W(\omega_{\min}, \omega_{\max}, \omega)$ provided a local control of spectral shaping within a given frequency range $[\omega_{\min}; \omega_{\max}]$. In particular, it was of interest to keep the first components unchanged during spectral control to reduce pitch variations. For instance, a window function $W(\omega_3, F_s/2, \omega)$ where F_s is the sampling frequency can be applied to only act on frequencies higher than ω_2. In practice, we chose a Tukey (tapered cosine) window defined between $\omega_{\min}$ and $\omega_{\max}$. The window is parameterized by a ratio ρ (between 0 and 1) allowing the user to choose intermediate profiles from rectangular ($\rho = 0$) to Hann ($\rho = 1$) windows. Consequently, the user is able to act on the weight of the local control.

From an acoustic point of view, the control acts on the spectral descriptors SB and R in a global way. For example, a decrease of the S_G value leads to a decrease of the SB value and at the same time to an increase of the R value.

2) Control by Amplitude and Frequency Modulations: Amplitude modulation creates two components on both sides of the original one and the modulated output waveform is expressed by

$$\begin{aligned} d_m^{\mathrm{AM}}(t) &= A_m\left(1 + I\cos(\omega_n t)\right)\cos(\omega_m t) \\ &= A_m\cos(\omega_m t) + \frac{A_m I}{2}\cos\left((\omega_m + \omega_n)t\right) \\ &\quad + \frac{A_m I}{2}\cos\left((\omega_m - \omega_n)t\right) \end{aligned}$$

where $I \in [0, 1]$ is the modulation index, ω_n the modulating frequency, A_m the amplitude, and ω_m the frequency of the mth component.

Frequency modulation creates a set of components on both sides of the original one and the modulated output waveform is expressed by

$$\begin{aligned} d_m^{\mathrm{FM}}(t) &= A_m\cos\left(\omega_m t + I\sin(\omega_n t)\right) \\ &= A_m \sum_{k=-\infty}^{\infty} J_k(I)\cos\left((\omega_m + k\omega_n)t\right) \end{aligned}$$

where $k \in N$ and $J_k(I)$ is the Bessel function of order k. The amplitude of these additional components are given by the amplitude of the original partial A_m and the values of $J_k(I)$ for a given modulation index I.

For both amplitude and frequency modulations, synthesis applications showed that applying the same value of the modulating frequency ω_n to all components led to synthetic sounds perceived as too artificial. To avoid this effect, we proposed a definition of the modulating frequency $\omega_{n,m}$ for each spectral component m based on perceptual considerations. Thus, $\omega_{n,m}$ was expressed as a percentage of the critical bandwidth Δf_m associated with each component m [59]

$$\Delta f_m = 25 + 75\left(1 + 1.4 f_m^2\right)^{0.69} \tag{19}$$

where f_m is expressed in kHz. Since Δf_m increases with respect to frequency, components created at high frequencies are more distant (in frequency) on both sides of the central component than components created at low frequencies. This provided an efficient way to control roughness since the addition of components within a critical bandwidth increases the perception of roughness. In particular, it is known that the maximum sensory dissonance corresponds to an interval between spectral components of about 25% of the critical bandwidth [51], [60].

Synthesis applications showed that both spectral shaping controls allowed for morphing particularly between Glass and Metal sounds while keeping the damping unchanged. In this case, the damping coefficients of the modified frequencies were recalculated according to the damping law. Both controls provided a local control since modifications can be applied on each original component independently. The control based on amplitude and frequency modulations allowed subtle spectral modifications compared with the control based on spectral dilation and in particular, led to interesting fine timbre effects such as cracked glass sounds (sound examples can be found in [38]).

D. Control of the Perceived Material

A global control strategy of the perceived material that integrates the previous damping and spectral shaping controls is proposed in this section. This strategy is hierarchically built on three layers: the "Material space" (accessible to the user), the "Damping and Spectral shaping parameters" and the "Signal parameters" (related to the signal model). Note that the mapping strategy does not depend on the synthesis technique. As a consequence, the proposed control can be applied to any sound generation process.[3] Note also that the proposed strategy is not unique and represents one among several other possibilities [24], [25].

Fig. 6 illustrates the mapping between these three layers based on the first spectral shaping control (using S_G and S_R). The Material space is designed as a unit disk of center C with three fixed points corresponding to the three reference sounds (Wood, Metal, and Glass) equally distributed along the external circle. The Glass sound position is arbitrarily considered as the angle's origin ($\theta = 0$) and consequently, the Metal sound is positioned at $\theta = 2\pi/3$ and the Wood sound at $\theta = 4\pi/3$. The three reference sounds were synthesized from the same initial set of harmonic components (fundamental frequency of 500 Hz and 40 components) so that Wood, Metal, and Glass sounds were obtained by only modifying the damping and spectral shaping parameters (values given in Table II and sound positions shown in Fig. 6). These parameters were chosen on the basis of the sound quality of the evoked material. Note that the reference sounds could be replaced by other sounds.

The user navigates in the Material space between Wood, Metal, and Glass sounds by moving a cursor and can synthesize the sound corresponding to any position. When moving along the circumference of the Material space circle, the corresponding sound $S_h(\theta)$ characterized by its angle θ is generated with Damping and Spectral shaping parameters defined by

$$\mathbf{P}_{S_h}(\theta) = T(\theta)\mathbf{P}_G + T\left(\theta - \frac{2\pi}{3}\right)\mathbf{P}_M + T\left(\theta - \frac{4\pi}{3}\right)\mathbf{P}_W \tag{20}$$

where $\mathbf{P}$ represent the parameter vector $\{\alpha_G, \alpha_R, S_G, S_R\}$ of the sound S_h and of the reference sound of Glass (G), Metal (M), and Wood (W). The function $T(\theta)$ was defined so that the interpolation process was exclusively made between two reference sounds at a time

$$T(\theta)\begin{cases} = -\frac{3}{2\pi}\theta + 1, & \text{for } \theta \in \left[0; \frac{2\pi}{3}\right[\\ = 0, & \text{for } \theta \in \left[\frac{2\pi}{3}; \frac{4\pi}{3}\right] \\ = \frac{3}{2\pi}\theta - 2, & \text{for } \theta \in \left]\frac{4\pi}{3}; 2\pi\right[\end{cases}. \tag{21}$$

Inside the circle, a sound $S_h'(r, \theta)$ characterized by its angle θ and its radius r is generated with parameters defined by

$$\mathbf{P}_{S_h'}(r, \theta) = (1 - r)\mathbf{P}_C + r\mathbf{P}_{S_h}(\theta) \tag{22}$$

[3] In practice, we implemented an additive synthesis technique (sinusoids plus noise) in the synthesizer previously developed [23] since it was the most natural one according to the signal model. Other techniques could have been considered as well such as frequency modulation (FM) synthesis, subtractive synthesis, etc.

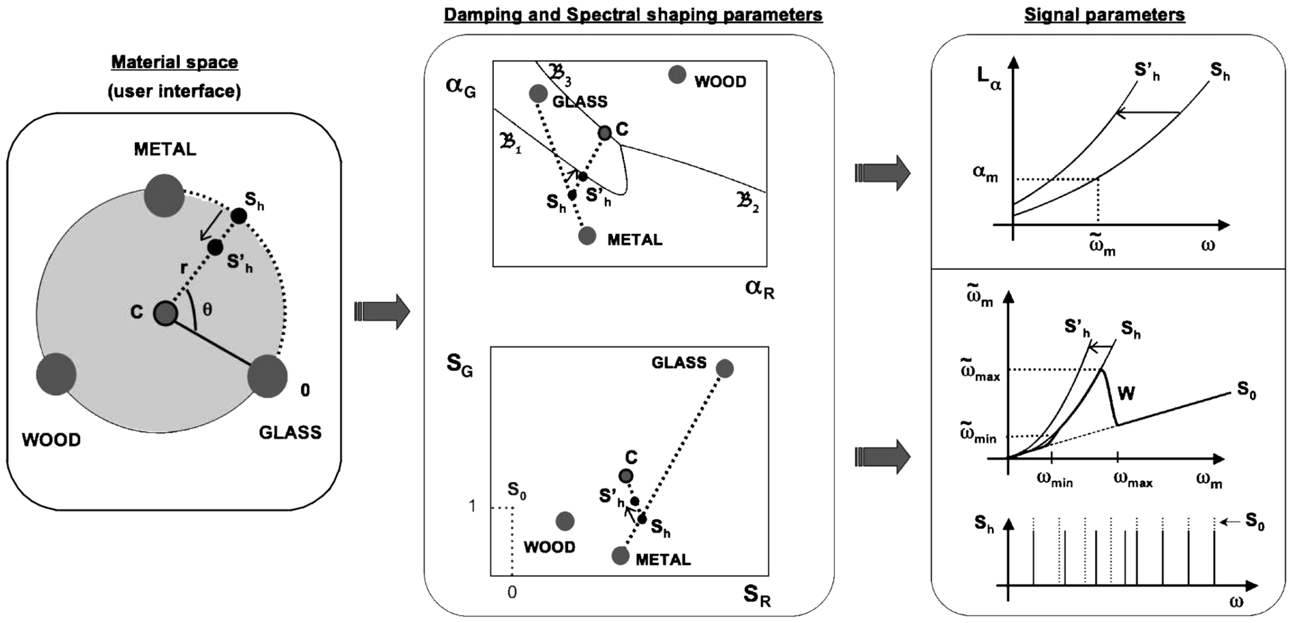


Fig. 6. Control of the perceived material based on three-layer architecture: the "Material space" (user interface), the "Damping and spectral shaping parameters" and "Signal parameters." The navigation in the Material space (e.g., from sound S_h to sound S'_h) involves modifications of both Damping and Spectral shaping parameters. The $\{\alpha_G, \alpha_R\}$ space was calibrated with specific domain for each material category (borders defined in Fig. 7). In this example, we represented the Spectral shaping parameters $\{S_G, S_R\}$ corresponding to the first spectral shaping control. Finally, at signal level, the damping coefficients α_m are computed from (4) with values of $\{\alpha_G, \alpha_R\}$ and the frequencies $\tilde{\omega}_m$ were computed from (15) with values of $\{S_G, S_R\}$. The Metal, Wood, and Glass reference sounds are constructed from the same initial harmonic sound S_0 located at point (0,1) in the $\{S_G, S_R\}$ space. The role of spectral shaping with the window function W is illustrated at the bottom. S_0 is represented in dotted and S_h in bold. The amplitude of the spectrum S_h was arbitrarily reduced for a sake of clarity.

TABLE II
VALUES OF DAMPING (α_G AND α_R) AND SPECTRAL SHAPING (S_G AND S_R) PARAMETERS CORRESPONDING TO THE REFERENCE SOUNDS OF METAL, WOOD, AND GLASS CATEGORY IN THE MATERIAL SPACE

	α_G	α_R ($\times 10^{-4}$)	S_G	S_R
Metal	0.6	2	0.5	0.1
Wood	3	4	0.85	0.05
Glass	2.5	1.5	2.4	0.2

where $\mathbf{P}_C$ represents the parameter vector of the sound C defined by $\{\bar{\alpha}_G, \bar{\alpha}_R, \bar{S}_G, \bar{S}_R\}$ with bar symbol denoting the average of the three values (corresponding to Wood, Metal, and Glass reference sounds) for each parameter and where $\mathbf{P}_{S_h}(\theta)$ is defined in (20). A similar strategy was designed for the mapping based on the second spectral shaping control (amplitude and frequency modulations). In that case, the parameter vector $\mathbf{P}$ corresponded to $\{\alpha_G, \alpha_R, I, \omega_n\}$.

The second layer concerns the controls of Damping and Spectral shaping parameters. For each control, the parameters are represented in two-dimensions to propose an intuitive configuration, called damping (α_G, α_R) and spectral shaping (S_G, S_R) spaces, respectively. The intuitive manipulation of α_G and α_R was achieved by a calibration process that consisted in determining a specific domain for each material category in the (α_G, α_R) space. Borders between material domains were determined based on results from predictive discrimination analysis. In practice, we calibrated the (α_G, α_R) space delimited by extreme values of α_G and α_R for typical sounds (range of $\alpha_G = [0.25; 3.34]$ and range of $\alpha_R = [0.5; 6.64] \times 10^{-4}$; see Fig. 7). This space was sampled in 2500 evenly spaced points and each point was associated with a posterior probability of belonging to a material category. This probability was computed from a Bayesian rule based on the knowledge of the positions of typical sounds in the (α_G, α_R) space. Classification functions were determined between pairs of material categories and were expressed as quadratic combination of α_G and α_R $(\times 10^{-4})$. Boundary curves were materialized from the set of points that have similar classification probabilities δ for both categories

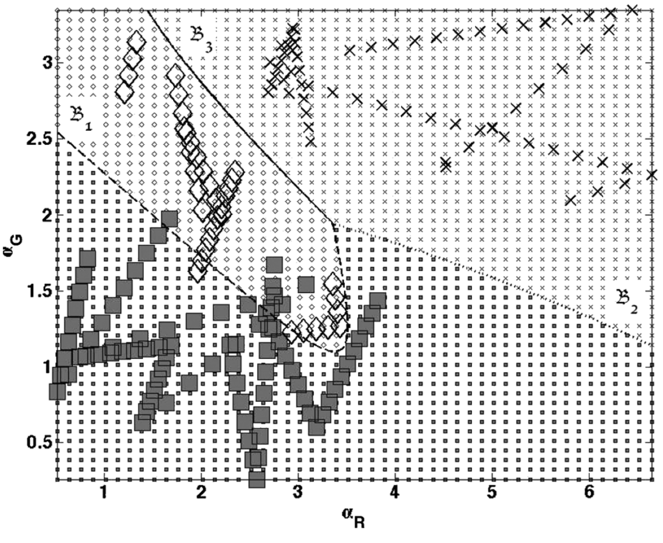


Fig. 7. Calibration of the $\{\alpha_G, \alpha_R\}$ space from border sections $\mathfrak{B}_1$ between Metal and Glass (dashed line), $\mathfrak{B}_2$ between Metal and Wood (dotted line) and $\mathfrak{B}_3$ between Glass and Wood (solid line). The positions of typical sounds for Wood (×), Metal (□), and Glass (◇) used for the classification process are also represented.

$$\{\mathbf{x} : \delta_{G_1}(\mathbf{x}) = \delta_{G_2}(\mathbf{x})\} \tag{23}$$

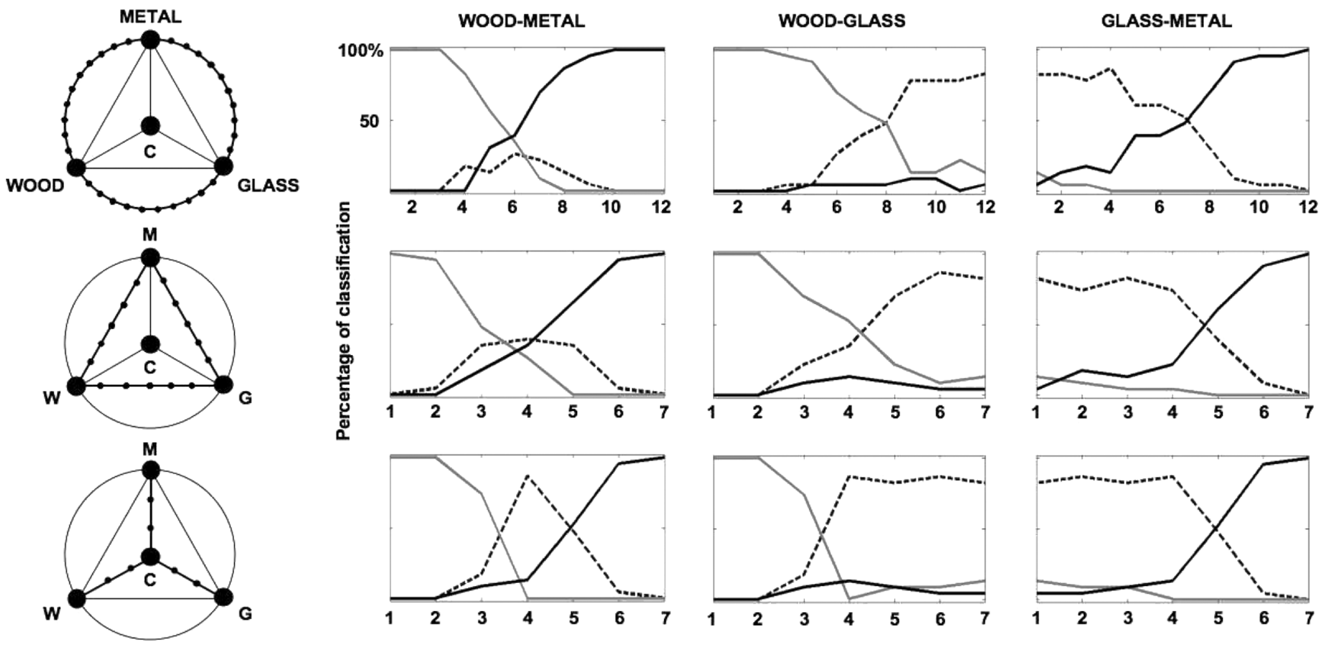


Fig. 8. Perceptual evaluation test of the control strategy. Left: position of sounds (black markers) used for the test in the Material Space as a function of the trajectory: along the external circle (first row), along the chords whose endpoints are the reference sounds (second row) and along the radii of the circle from the center C to the reference sounds (third row). Right: percentage of classification as Wood (gray curves), Metal (black curves), or Glass (dashed curves) corresponding to the three types of trajectory for each sound as a function of its position of the continuum.

or equivalently

$$\{\mathbf{x} : 0 = \delta_{G_1}(\mathbf{x}) - \delta_{G_2}(\mathbf{x})\} \quad (24)$$

where $\mathbf{x} = (\alpha_R, \alpha_G)$ and G_1 and G_2 the categories.

For our concern, the border noted $\mathfrak{B}_1$ between Metal and Glass categories was defined by

$$\mathfrak{B}_1 : 0 = 109.69 - 48.44\alpha_R - 48.30\alpha_G + 5.33(\alpha_R)^2 + 11.33\alpha_R\alpha_G + 3.37(\alpha_G)^2 \quad (25)$$

and all the points for which this function is negative were classified into the Metal category. The border $\mathfrak{B}_2$ between Wood and Metal was defined by

$$\mathfrak{B}_2 : 0 = -41.18 + 1.50\alpha_R + 17.75\alpha_G + 0.27(\alpha_R)^2 - 0.09\alpha_R\alpha_G - 0.20(\alpha_G)^2 \quad (26)$$

and all the points for which this function is negative were classified into the Wood category. Finally, the border $\mathfrak{B}_3$ between Wood and Glass regions was defined by

$$\mathfrak{B}_3 : 0 = 68.51 - 46.94\alpha_R - 30.55\alpha_G + 5.60(\alpha_R)^2 + 11.24\alpha_R\alpha_G + 3.17(\alpha_G)^2 \quad (27)$$

and all the points for which this function is negative were classified into the Wood category. The calibration of the (α_G, α_R) space was completed by keeping the section of the borders that directly separate the two sound categories: as shown in Fig. 7, the border section that was kept for $\mathfrak{B}_1$ is represented by a dashed line, the section kept for $\mathfrak{B}_2$ is represented by a dotted line and the section kept for $\mathfrak{B}_3$ is represented by a solid line. These borders were reported in the Middle layer (Fig. 6) allowing an intuitive manipulation of damping parameters. Note that these borders do not represent a strict delimitation between sound categories and a narrow transition zone may be taken into account on both sides of the borders. In particular, sounds belonging to this transition zone may be perceived as ambiguous sounds such as sounds created at intermediate positions of the continua.

Finally, the bottom layer concerns the signal parameters determined as follows: the damping coefficients α_m are computed from (4) with $\{\alpha_G, \alpha_R\}$ values and frequencies $\tilde{\omega}_m$ from (15) with $\{S_G, S_R\}$ values. The amplitudes A_m are assumed to be equal to one.

V. Perceptual Evaluation of the Control Strategy

The proposed control strategy for the perceived material was evaluated with a formal perceptual test. Twenty-three participants (9 women, 14 men) participated in the experiment. Sounds were selected in the Material Space as shown in Fig. 8 (left). Three types of trajectory between two reference sounds were investigated: along the external circle (by a 12-step continuum), along chords whose endpoints are the reference sounds (7-step continuum) and along the radii of the circle from the center C to the reference sounds (7-step continuum). Sounds were presented once randomly through headphones. The whole set of sounds are available at [38]. Participants were asked to categorize each sound as Wood, Metal, or Glass, as fast as possible, by selecting with a mouse on a computer screen the corresponding label. The order of labels displayed on the screen was balanced across participants. The next sound was presented after a 2-seconds silence. Participants' responses were collected and averaged for each category (Wood, Metal, and Glass) and for each sound.

Fig. 8 (right) shows results as a function of sound position along the continua. Sounds at extreme positions were classified by more than 70% of participants in the correct category, leading to the validation of the reference sounds as typical exemplars of their respective material category. In Wood–Metal transition, sounds at intermediate positions were classified as Glass with highest percentages for those along the trajectory via the center C. By contrast, in both Wood–Glass and Glass–Metal transitions, intermediate sounds were most often classified in one of the two categories corresponding to the extreme sounds. From an acoustic point of view, this reflects the fact that, the interpolation of Damping parameters between Metal and Wood sounds crosses the Glass category while this is not the case for the other two transitions (see Fig. 6).

These results were in line with the ones obtained from the behavioral data in the first categorization experiment (Fig. 2) and consequently, allowed us to validate the proposed control strategy as an efficient way to navigate in the Material space. Note that the interpolation process was computed on a linear scale between parameters of the reference sounds [cf. (21) and (22)]. The next step will consist in taking into account these results and modify the interpolation rules so that the metric distance between a given sound and a reference sound in the Material space closely reflects perceptual distance.

VI. Conclusion

In this paper, we proposed a control strategy for the perceived material in an impact sound synthesizer. To this end, we investigated the sound characteristics relevant for an accurate evocation of material by conducting a sound categorization experiment. To design the stimuli, sounds produced by impacting three different materials, i.e., wood, metal, and glass, were recorded and synthesized by using analysis–synthesis techniques. After tuning, sound continua simulating progressive transitions between material categories were built and used in the categorization experiment. Both behavioral data and electrical brain activity were collected. From behavioral data, a set of typical sounds for each material category was determined and an acoustic analysis including descriptors known to be relevant for timbre perception and material identification was conducted. The most relevant descriptors that allow discrimination between material categories were identified: the normalized sound decay (related to the damping) and the roughness. Electrophysiological data provided complementary information regarding the perceptual/cognitive aspects related to the sound categorization and were discussed in the context of synthesis. Based on acoustic and ERP data, results confirmed the importance of damping and highlighted the relevance of spectral descriptors for material perception. Control strategies for damping and spectral shaping were proposed and tested in synthesis applications. These strategies were further integrated in a three-layer control architecture allowing the user to navigate in a "Material Space." A formal perceptual evaluation confirmed the validity of the proposed control strategy. Such a control offered an intuitive manipulation of parameters and allowed defining realistic impact sounds directly from the material label (i.e., Wood, Metal, or Glass).

References

[1] M. V. Mathews, "The digital computer as a musical instrument," *Science*, vol. 142, no. 3592, pp. 553–557, 1963.

[2] R. Moog, "Position and force sensors and their application to keyboards and related controllers," in *Proc. AES 5th Int. Conf.: Music Digital Technol.*, 1987, pp. 179–181, A. E. S. New York, Ed..

[3] M. Battier, "L'approche gestuelle dans l'histoire de la lutherie électronique. Etude d'un cas: Le theremin," *Proc. Colloque International* , ser. Collection Eupalinos, Editions Parenthèses, 1999, Les nouveaux gestes de la musique.

[4] J. Tenney, "Sound-generation by means of a digital computer," *J. Music Theory*, vol. 7, no. 1, Spring, 1963.

[5] A. Camurri, M. Ricchetti, M. Di Stefano, and A. Stroscio, "Eyesweb—Toward gesture and affect recognition in dance/music interactive systems," in *Proc. Colloquio di Informatica Musicale*, 1998.

[6] P. Gobin, R. Kronland-Martinet, G. A. Lagesse, T. Voinier, and S. Ystad*, From Sounds to Music: Different Approaches to Event Piloted Instruments*, ser. Lecture Notes in Computer Science. : Springer-Verlag, 2003, vol. 2771, pp. 225–246.

[7] M. Wanderley and M. Battier, "Trends in gestural control of music," IRCAM-Centre Pompidou, 2000.

[8] J.-C. Risset and D. L. Wessel, "Exploration of timbre by analysis and synthesis," in *The Psychology of Music*, ser. Cognition and Perception, 2nd ed. New York: Academic, 1999, pp. 113–169.

[9] D. L. Wessel, "Timbre space as a musical control structure," *Comput. Music J.*, vol. 3, no. 2, pp. 45–52, 1979.

[10] S. Ystad and T. Voinier, "A virtually-real flute," *Comput. Music J.*, vol. 25, no. 2, pp. 13–24, Summer, 2001.

[11] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Amer.*, vol. 61, no. 5, pp. 1270–1277, 1977.

[12] C. L. Krumhansl, "Why is musical timbre so hard to understand," in *Structure and Perception of Electroacoustic Sound and Music*. Amsterdam, The Netherlands: Elsevier, 1989.

[13] J. Krimphoff, S. McAdams, and S. Winsberg, "Caractérisation du timbre des sons complexes. II: Analyses acoustiques et quantification psychophysique [characterization of timbre of complex sounds. II: Acoustical analyses and psychophysical quantification]," *J. Phys.*, vol. 4, no. C5, pp. 625–628, 1994.

[14] S. McAdams, S. Winsberg, S. Donnadieu, G. D. Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychol. Res.*, vol. 58, pp. 177–192, 1995.

[15] W. A. Sethares, "Local consonance and the relationship between timbre and scale," *J. Acoust. Soc. Amer.* , vol. 94, no. 3, pp. 1218–1228, 1993.

[16] P. N. Vassilakis, "Auditory roughness as a means of musical expression," Dept. of Ethnomusicology, Univ. of California, Selected reports in ethnomusicology (perspectives in systematic musicology), 2005, vol. 12, pp. 119–144.

[17] R. P. Wildes and W. A. Richards*, Recovering Material Properties From Sound*, W. A. Richards, Ed. Cambridge, MA: MIT Press, 1988, ch. 25, pp. 356–363.

[18] W. W. Gaver, "How do we hear in the world ? explorations of ecological acoustics," *Ecol. Psychol.*, vol. 5, no. 4, pp. 285–313, 1993.

[19] R. Lutfi and E. Oh, "Auditory discrimination of material changes in a struck-clamped bar," *J. Acoust. Soc. Amer.*, vol. 102, no. 6, pp. 3647–3656, 1997.

[20] R. L. Klatzky, D. K. Pai, and E. P. Krotkov, "Perception of material from contact sounds," *Presence: Teleoperators and Virtual Environments*, vol. 9, no. 4, pp. 399–410, 2000.

[21] B. L. Giordano and S. McAdams, "Material identification of real impact sounds: Effects of size variation in steel, wood, and Plexiglas plates," *J. Acoust. Soc. Amer.*, vol. 119, no. 2, pp. 1171–1181, 2006.

[22] M. Aramaki, M. Besson, R. Kronland-Martinet, and S. Ystad, "Timbre perception of sounds from impacted materials: Behavioral, electrophysiological and acoustic approaches," in *Computer Music Modeling and Retrieval—Genesis of Meaning of Sound and Music*, ser. LNCS, S. Ystad, R. Kronland-Martinet, and K. Jensen, Eds. Berlin, Heidelberg, Germany: Springer-Verlag, 2009, vol. 5493, pp. 1–17.

[23] M. Aramaki, R. Kronland-Martinet, T. Voinier, and S. Ystad, "A percussive sound synthesizer based on physical and perceptual attributes," *Comput. Music J.*, vol. 30, no. 2, pp. 32–41, 2006.

[24] M. Aramaki, R. Kronland-Martinet, T. Voinier, and S. Ystad, "Timbre control of a real-time percussive synthesizer," in *Proc. 19th Int. Congr. Acoust. (CD-ROM)*, 2007, 84-87985-12-2.

[25] M. Aramaki, C. Gondre, R. Kronland-Martinet, T. Voinier, and S. Ystad, "Thinking the sounds: An intuitive control of an impact sound synthesizer," in *Proc. 15th Int. Conf. Auditory Display (ICAD 2009)*, 2009.

[26] K. Steiglitz and L. E. McBride, "A technique for the identification of linear systems," *IEEE Trans. Autom. Control*, vol. AC-10, no. 10, pp. 461–464, Oct. 1965.

[27] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[28] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.

[29] R. Badeau, B. David, and G. Richard, "High-resolution spectral analysis of mixtures of complex exponentials modulated by polynomials," *IEEE Trans. Signal Process.*, vol. 54, no. 4, pp. 1341–1350, Apr. 2006.

[30] L.-M. Reissell and D. K. Pai, "High resolution analysis of impact sounds and forces," in *Proc. WHC '07: 2nd Joint EuroHaptics Conf. Symp. Haptic Interfaces for Virtual Environment and Teleoperator Syst.*, Washington, DC, 2007, pp. 255–260, IEEE Computer Society.

[31] M. Aramaki and R. Kronland-Martinet, "Analysis-synthesis of impact sounds by real-time dynamic filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 695–705, Mar. 2006.

[32] R. Kronland-Martinet, P. Guillemain, and S. Ystad, "Modelling of natural sounds by time-frequency and wavelet representations," *Organised Sound*, vol. 2, no. 3, pp. 179–191, 1997.

[33] R. Parncutt, *Harmony—A Psychoacoustical Approach*. Berlin/Heidelberg, Germany: Springer, 1989.

[34] A. Chaigne and C. Lambourg, "Time-domain simulation of damped impacted plates: I. Theory and experiments," *J. Acoust. Soc. Amer.*, vol. 109, no. 4, pp. 1422–1432, 2001.

[35] T. Ono and M. Norimoto, "Anisotropy of dynamic young's modulus and internal friction in wood," *Jpn. J. Appl. Phys.*, vol. 24, no. 8, pp. 960–964, 1985.

[36] S. McAdams, A. Chaigne, and V. Roussarie, "The psychomechanics of simulated sound sources: Material properties of impacted bars," *J. Acoust. Soc. Amer.*, vol. 115, no. 3, pp. 1306–1320, 2004.

[37] M. Aramaki, H. Baillères, L. Brancheriau, R. Kronland-Martinet, and S. Ystad, "Sound quality assessment of wood for xylophone bars," *J. Acoust. Soc. Amer.*, vol. 121, no. 4, pp. 2407–2420, 2007.

[38] 2010 [Online]. Available: http://www.lma.cnrs-mrs.fr/~kronland/Categorization/sounds.html, last checked: Oct. 2009

[39] M. Aramaki, L. Brancheriau, R. Kronland-Martinet, and S. Ystad, "Perception of impacted materials: Sound retrieval and synthesis control perspectives," in *Computer Music Modeling and Retrieval—Genesis of Meaning of Sound and Music*, ser. LNCS, S. Ystad, R. Kronland-Martinet, and K. Jensen, Eds. Berlin, Heidelberg, Germany: Springer-Verlag, 2009, vol. 5493, pp. 134–146.

[40] M. D. Rugg and M. G. H. Coles, "The ERP and cognitive psychology: Conceptual Issues," in *Electrophysiology of Mind. Event-Related Brain Potentials and Cognition*, ser. Oxford Psychology. New York: Oxford Univ. Press, 1995, pp. 27–39, no. 25.

[41] J. Eggermont and C. Ponton, "The neurophysiology of auditory perception: From single-units to evoked potentials," *Audiol. Neuro-Otol.*, vol. 7, pp. 71–99, 2002.

[42] A. Shahin, L. E. Roberts, C. Pantev, L. J. Trainor, and B. Ross, "Modulation of p2 auditory-evoked responses by the spectral complexity of musical sounds," *NeuroReport*, vol. 16, no. 16, pp. 1781–1785, 2005.

[43] S. Kuriki, S. Kanda, and Y. Hirata, "Effects of musical experience on different components of meg responses elicited by sequential pianotones and chords," *J. Neurosci.*, vol. 26, no. 15, pp. 4046–4053, 2006.

[44] E. Kushnerenko, R. Ceponiene, V. Fellman, M. Huotilainen, and I. Winkler, "Event-related potential correlates of sound duration: Similar pattern from birth to adulthood," *NeuroReport*, vol. 12, no. 17, pp. 3777–2781, 2001.

[45] C. Alain, B. M. Schuler, and K. L. McDonald, "Neural activity associated with distinguishing concurrent auditory objects," *J. Acoust. Soc. Amer.*, vol. 111, no. 2, pp. 990–995, 2002.

[46] S. McAdams, "Perspectives on the contribution of timbre to musical structure," *Comput. Music J.*, vol. 23, no. 3, pp. 85–102, 1999.

[47] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. New York: Wiley, 2005.

[48] G. Peeters, "A Large set of audio features for sound description (similarity and description) in the Cuidado Project," IRCAM, Paris, France, 2004, Tech. Rep..

[49] J. Marozeau, A. de Cheveigné, S. McAdams, and S. Winsberg, "The dependency of timbre on fundamental frequency," *J. Acoust. Soc. Amer.*, vol. 114, pp. 2946–2957, 2003.

[50] J. W. Beauchamp, "Synthesis by spectral amplitude and "brightness" matching of analyzed musical instrument tones," *J. Audio Eng. Soc.*, vol. 30, no. 6, pp. 396–406, 1982.

[51] R. Plomp and W. J. M. Levelt, "Tonal consonance and critical bandwidth," *J. Acoust. Soc. Amer.*, vol. 38, pp. 548–560, 1965.

[52] E. Terhardt, "On the perception of periodic sound fluctuations (roughness)," *Acustica*, vol. 30, no. 4, pp. 201–213, 1974.

[53] P. N. Vassilakis, "SRA: A web-based research tool for spectral and roughness analysis of sound signals," in *Proc. 4th Sound Music Comput. (SMC) Conf.*, 2007, pp. 319–325.

[54] 2009 [Online]. Available: http://www.cycling74.com/, last checked: Octob. 2009

[55] D. Rocchesso and F. Fontana, 2003, "The Sounding Object," [Online]. Available: http://www.soundobject.org/SobBook/SobBook_JUL03.pdf last checked: Oct. 2009

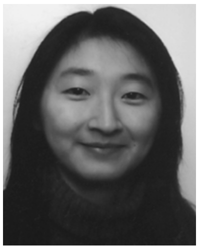
[56] M. Hyde, "The N1 response and its applications," *Audiol. Neuro-Otol.*, vol. 2, pp. 281–307, 1997.

[57] C. Valette and C. Cuesta, *Mécanique de la corde vibrante (Mechanics of vibrating string)*, ser. Traité des Nouvelles Technologies, série Mécanique. London, U.K.: Hermès, 1993.

[58] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments (Second Edition)*. Berlin, Germany: Springer-Verlag, 1998.

[59] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*. Berlin, Germany: Springer-Verlag, 1990.

[60] H. L. F. von Helmholtz, *On the Sensations of Tone as the Physiological Basis for the Theory of Music*, 2nd ed. New York: Dover, 1877, reprinted 1954.



Mitsuko Aramaki (M'09) received the M.S. degree in mechanics (speciality in acoustic and dynamics of vibrations) from the University of Aix-Marseille II, Marseille, France, and the Ph.D. degree for her work at the Laboratoire de Mécanique et d'Acoustique, Marseille, France, in 2003, on analysis and synthesis of impact sounds using physical and perceptual approaches.

She is currently a Researcher at the Institut de Neurosciences Cognitives de la Méditerranée, Marseille, where she works on a pluridisciplinary project combining sound modeling, perceptual and cognitive aspects of timbre, and neuroscience methods in the context of virtual reality.

Mireille Besson received the Ph.D. degree in neurosciences from the University of Aix-Marseille II, Marseille, France, in 1984.

After four years of post-doctorate studies at the Department of Cognitive Science, University of California at San Diego, La Jolla, working with Prof. M. Kutas and at the Department of Psychology, University of Florida, Gainesville, working with Prof. I. Fischler, she obtained a permanent position at the National Center for Scientific Research (CNRS), Marseille, France. She is currently Director of Research at the CNRS, Institut de Neurosciences Cognitives de la Méditerranée (INCM), where she is the head of the "Language, Music, and Motor" team. Her primary research interests are centered on brain imaging of linguistic and non linguistic sound perception and on brain plasticity mainly using event-related brain potentials. She is currently conducting a large research project on the influence of musical training on linguistic sound perception in normal reading and dyslexic children.

Richard Kronland-Martinet (M'09–SM'10) received the M.S. degree in theoretical physics in 1980, the Ph.D. degree in acoustics from the University of Aix-Marseille II, Marseille, France, in 1983, and the "Doctorat d'Etat es Sciences" degree from the University of Aix-Marseille II in 1989 for his work on analysis and synthesis of sounds using time–frequency and time–scale (wavelets) representations.

He is currently Director of Research at the National Center for Scientific Research (CNRS), Laboratoire de Mécanique et d'Acoustique, Marseille, where he is the Head of the group "Modeling, Synthesis and Control of Sound and Musical Signals." His primary research interests are in analysis and synthesis of sounds with a particular emphasis on high-level control of synthesis processes. He recently addressed applications linked to musical interpretation and semantic description of sounds using a pluridisciplinary approach associating signal processing, physics, perception, and cognition.

Sølvi Ystad received the Ph.D. degree in acoustics from the University of Aix-Marseille II, Marseille, France, in 1998.

She is currently a Researcher at the National French Research Center (CNRS) in the research team S2M—Synthesis and Control of Sounds and Musical Signals—in Marseille, France. Her research activities are related to sound modeling with a special emphasis on the identification of perceptually relevant sound structures to develop efficient synthesis models. She was in charge of the research project "Towards the sense of sounds," financed by the French National Agency (ANR—http://www.sensons.cnrs-mrs.fr) from 2006–2009.

Acoustical correlates of timbre and expressiveness in clarinet performance

Mathieu Barthet[1], Philippe Depalle[2], Richard Kronland-Martinet, Sølvi Ystad
CNRS Laboratoire de Mécanique et d'Acoustique
31 chemin Joseph-Aiguier
13402 Marseille Cedex 20
France

This study deals with the acoustical factors liable to account for expressiveness in clarinet performances. Mechanical and expressive performances of excerpts from Bach's *Suite no. II* and Mozart's *Quintet for Clarinet and Strings* were recorded. Timbre, timing, dynamics and pitch descriptors were extracted from the recorded performances. The data were processed with a two-way analysis of variance, where the musician's expressive intentions and the note factors were defined as the independent variables. In both musical excerpts, a strong effect of the expressive intention was observed on the timbre (Attack Time, Spectral Centroid, Odd/Even Ratio), temporal (Intertone Onset Intervals) and dynamics (Root Mean Square envelope) descriptors. The changes in the timbre descriptors were found to depend on the position of the notes in the musical phrases. These results suggest that timbre, as well as temporal and dynamics variations, may mediate expressiveness in the musical messages transmitted from performers to listeners.

Since the beginning of the twentieth century, the authors of studies on musical performance have been attempting to analyze, understand and model the processes underlying the act of musical interpretation, namely "the act of performance with the implication that in this act the performer's judgment and personality necessarily have their share" (Scholes, 1960). Although many studies have focused on timing (e.g. note durations, tempo, chord asynchronization) and dynamics (see e.g. Repp, 1992; Todd, 1992; Kendall & Carterette, 1990), intonation, phrasing and articulation (see e.g. Gabrielsson & Lindstrom, 1995), less attention has been paid so far to timbre (see Juslin & Laukka, 2003, for a review). Here we present the first part of a study on the role of timbre in the musical message transmitted from performers to listeners. For this purpose, mechanical and expressive clarinet performances were recorded and analyzed in order to determine which (if any) acoustical correlates of timbre change when a performer plays more expressively.

The notion of expressive deviations

Early studies carried out around 1930 by C. E. Seashore's group at Iowa University led to the conjecture that "the artistic expression of feeling in music consists in aesthetic deviation from the regular - from pure tone, true pitch, even dynamics, metronomic time, rigid rhythms, etc." (Seashore, 1938). This suggests that expressiveness in music can be characterized by measuring acoustical features of the musical instrument's tones related to time, energy, frequency, and/or the performer's instrumental gestures. In line with these early studies, artistic expression has often been approached by measuring the deviations of time and frequency parameters with respect to fixed and regular values corresponding to a strictly "mechanical" rendering of the score (Seashore, 1938; Gabrielsson, 1999).

The measurement of musical performance

Detailed reviews of studies taking a psychological approach to the measurement of musical performance have been published by Gabrielsson (1999) and Palmer (1997). Repp (1992) investigated the timing differences and commonalities between several famous pianists' interpretations of Schumann's *"Träumerei"*. Statistical analyses of the Intertone Onset Intervals (IOIs) showed the existence of recurrent patterns, such as the ritardandi observed at the end of musical phrases, corresponding to "how most pianists transmit musical structure and expression through timing variations", as well as other patterns reflecting the individuality and eccentricity of some performers, such as Vladimir Horowitz and Alfred Cortot, in particular. In a study on the interpretation of a Mozart piano sonata, Palmer (1996) established that expert pianists consistently repeated the same prosodic timing and intensity patterns. These results tend to prove that timing and dynamic deviations are not random, but are linked to musicians' expressive intentions. The important role of timing and dynamics regarding piano expressivity, as observed by Repp and Palmer, may be related to the fact that piano is impoverished in terms of its ability to

[1] Any comments can be sent to the first author at barthet@lma.cnrs-mrs.fr
[2] Present address: Sound Processing and Control Laboratory, The Schulich School of Music, McGill University, 555 Sherbrooke Street West, Montreal (Quebec), H3A 1E3 Canada

manipulate timbre and that performers naturally use the degrees of freedoms at their disposal in a performance (note that, due to the sympathetic resonances of the strings, the piano still allows to perform subtle timbre modifications by varying the key playing technique). In the case of timbre, Födermayr and Deutsch (1993) based on their analysis of spectrograms of several interpretations of an aria by Verdi (*"Parmi veder le lagrime"* in the piece *"Rigoletto"*), noted that one of the singers applied a subtle change of timbre to a vowel for expressive effect. The present study focuses on whether changes of timbre of this kind occur arbitrarily or whether, on the contrary, they are dictated by the performer's expressive intentions.

On the definition of timbre

Back in 1938, Seashore (1938) was already convinced that timbre contributes importantly to musical aesthetics but no appropriate means of measurement were available for examining this parameter more closely: "We should here recognize that timbre as a fourth attribute of tone is by far the most important aspect of tone and introduces the largest number of problems and variables". More than seventy years later, there still exists no widely-accepted definition of timbre on which researchers can base general models for timbre. In the psychoacoustical context, timbre is defined as the attribute of the auditory sensation which allows to distinguish different sounds equal in pitch, loudness, and duration, and has shown to be related to the temporal and spectral characteristics of the sounds (ANSI, 1960). Timbre is hence closely related to the identity of the sound source. However, as remarked by Schaeffer (1966), this facet of timbre is paradoxical: how can we speak of an instrument's timbre when each of its tones also possesses a specific timbre ? In his description of timbre, Schaeffer combines the causal invariants which may be partly responsible for the instrument's identity (e.g. the hammered strings in the case of the piano), with the sources of variations, some of which are linked to the instrument's register (e.g. the low registers are generally richer than the high registers), and others which are due to the performers' control gestures. A description of timbre cannot therefore be limited to the typological aspects mentioned above but should also include the morphological aspects. Timbre can therefore be regarded as an elementary perceptual property of sound which can vary in a single instrument with time. This twofold nature of timbre (identity/quality) can be explained in terms of cognitive categorization's theories: musical sounds can be categorized either in terms of the sources from which they are generated, or simply as sounds, in terms of the properties which characterize them (Handel, 1995; Castellengo & Dubois, 2005).

The timbre descriptors: the acoustical correlates of timbre

Previous research on timbre has mostly consisted in quantifying the acoustical correlates, which are also known as *timbre descriptors* (see Hajda, Kendall, Carterette, & Harshberger, 1997; McAdams, 1994, for detailed historical reviews). The methods used to address this issue are mostly based on multidimensional scaling (MDS) techniques, with which various timbres can be mapped in a low-dimensional space (the so-called *timbre space*), where the relative positions reflect the degree of perceived proximity. The structure of the perceptual representation of timbre is sensitive to the choice and number of stimuli used in these studies. However, the differences in timbre between orchestral instruments' tones are usually modeled in a three-dimensional perceptual space (see e.g. Grey, 1977; Wessel, 1979; Krumhansl, 1989; Kendall & Carterette, 1991; McAdams, Winsberg, Donnadieu, De Soete, & Krimphoff, 1995). The dimensions of this space are generally well correlated with descriptors based on the temporal (e.g. Attack Time), spectral (e.g. Spectral Centroid) and spectro-temporal (e.g. Spectral Flux) aspects of sounds (see Krimphoff, McAdams, & Winsberg, 1994, for the definitions of these descriptors).

Methods

Procedure

We focused here on a monodic instrument, the clarinet, because its mechanical and acoustic properties make it possible for the player to control the timbre very closely while playing: in self-sustained instruments of this kind, the coupling between the exciter and the resonator is maintained practically throughout the duration of the sound. In order to test whether timbre plays an expressive role, we developed a method with which performances played with different expressive intentions can be compared. Contrary to what occurs with pitch and rhythm, which can be compared with the composer's indications on the score, it is difficult to define a control level in the case of timbre. In this study, mechanical or inexpressive performances were used as a reference against which deviations in acoustical correlates of timbre are quantified.

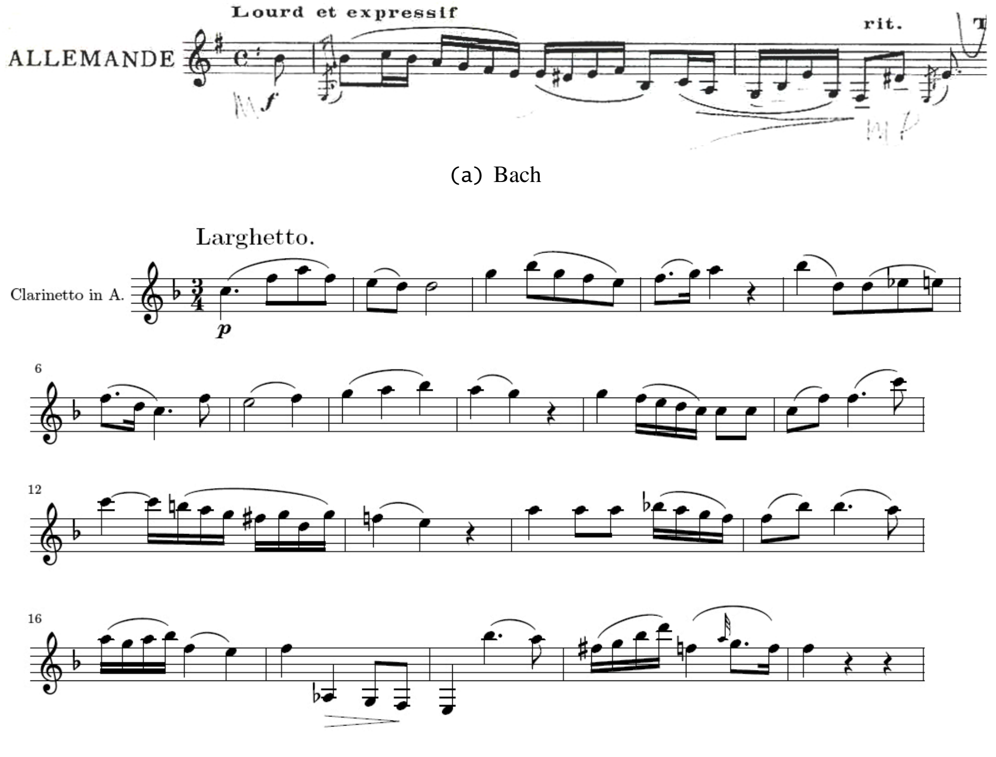
Sound corpus

Musical excerpts were recorded with the same professional clarinet player during two different recording sessions. The scores and a description of the musical excerpts are shown in Figure 1 and Table 1, respectively.

(Insert Figure 1)

(Insert Table 1)

Bach's *Allemande* The first musical excerpt was the opening phrase from the *Allemande* movement of Bach's *Suite no. II* (BWV 1008). Note that this Suite was written by Bach for the cello. An adaptation for the clarinet by U. Delécluse was used here by the performer. The musical phrase is in quadruple time and is played in the lowest register of the clarinet, the *chalumeau*. A ritardando is indicated at the end of the phrase. The clarinetist was asked to repeat the phrase 20 times with two different levels of expression. The first



(a) Bach

(b) Mozart

Figure 1. Scores of the Bach (a) and Mozart (b) musical excerpts.

Table 1
Description of the sound corpus

Musical excerpt	*Allemande - Bach's Suite*	*Larghetto - Mozart's Quintet*
Duration (bars/notes)	1.8 bars (27 notes, N1 to N27)	17 bars (75 notes, N1 to N75)
Nb. of phrases	1	4
Nb. of mechanical performances	20 (P1 to P20)	2 (P1 to P2)
Nb. of expressive performances	20 (P21 to P40)	4 (P3 to P6)
Reference tempo (*bpm*)	48	44

level corresponded to a mechanical or inexpressive rendering (keeping strictly to the indications on the score), whereas the second level corresponded to an expressive interpretation. A reference tempo of 48 bpm was chosen by the performer. During the mechanical performances, the metronome beats were delivered to the performer via an earphones. During the expressive performances, the reference tempo was given only just before the recordings, and was then turned off.

Mozart's *Larghetto* The second musical excerpt was the first 17 bars of the *Larghetto* movement of Mozart's *Quintet for Clarinet and Strings* (KV 581). This piece, written for the middle register of the clarinet, the *clarino*, is in triple time. The *Larghetto* movement was chosen because it seemed likely that a slow movement would probably give the performer more time to modulate the timbre while playing. The clarinetist was asked to give 2 performances in a mechanical way and 4 in an expressive way, at the self-selected reference tempo of 44 *bpm*.

Recordings All the recordings were carried out in an anechoic chamber to prevent any room reflections from affecting the sound. As timbre is known to be influenced by the characteristics of the recording equipment and the settings, two different types of microphones and setting positions were used. The first was a system of microphones attached to the body and the bell of the instrument (*SD Systems LCM 82 for Clarinet*). The second was a *Neumann KM 183* omnidirectional microphone, placed approximately 1.50 m from the instrument, at right angles to its body. The recordings of the Bach excerpt obtained with the *SD Systems* microphones were used for further analyses, as they were judged to reproduce more accurately the timbre of the instrument played in the lower register (*chalumeau*), in comparison to the recordings obtained with the *Neumann* microphone which presented too much bass components. Conversely, the recordings obtained with the *Neumann* microphone were selected in the case of the Mozart sequence, because they were judged to be more faithful to the timbre of the instrument in the *clarino* register than those obtained with the *SD Systems* microphones, which were very bright and short of bass components. Note that the use of different microphones for the Bach and Mozart excerpts did not affect the results since the analyses consisted in evaluating the differences between the mechanical and expressive performances within a same musical excerpt. All the performances were digitized at a sampling rate of 44.1 kHz.

Segmentation of the performances

The performances were first segmented in order to analyze the timing of the notes (i.e., the onset and offset times). We previously developed a semi-automatic note segmentation procedure (some errors had to be corrected manually) based on the detection of the instabilities of the fundamental frequencies (F0) occurring at each transition from one note to another. The instabilities are due to the fact that the F0s are not clearly defined when the self-sustained oscillations have not started and when they have ended. Further details about this procedure are given in (Barthet, Kronland-Martinet, & Ystad, 2006).

Acoustical analyses of the performances

The next step consists in using acoustical analysis techniques to extract temporal, timbre, dynamics and pitch descriptors from the recorded performances.

Temporal descriptors.

The Intertone Onset Interval deviation (ΔIOI) The duration of tones, which is classically quantified in terms of the Intertone Onset Interval (IOI) (Repp, 1992), is used by performers as a means of expression (Palmer, 1996). In order to characterize the local changes of tempo during a performance, we computed the IOI deviation descriptor ΔIOI, defined as the difference between the measured IOIs (called the effective IOIs) and the IOIs obtained by directly transcribing the notations on the score (called the nominal IOIs):

$$\Delta IOI = IOI_{eff} - IOI_{nom} \quad (1)$$

where IOI_{eff} and IOI_{nom} are the effective and nominal IOIs, respectively.

The tempo (TEMPO) The mean tempo of a performance, denoted TEMPO, was defined as the ratio between the total number of beats in the musical excerpt and the sum of the effective IOIs.

Timbre descriptors.

In a previous study, we observed that the perceptual dissimilarities between synthesized clarinet tones could be accurately represented in a three-dimensional timbre space (Barthet, Guillemain, Kronland-Martinet, & Ystad, 2010), whose dimensions were well correlated with the Attack Time (AT), the Spectral Centroid (SC) and the Odd/Even Ratio (OER). These three timbre descriptors were therefore used here to quantify clarinet timbre variations.

The Attack Time (AT) The attack time is correlated with the rate of energy increase in the onset of a sound. Results presented in (Barthet, Guillemain, et al., 2010) have shown that the attack time of clarinet tones depends on two main control parameters, the player's blowing pressure and the force he imposes on the reed with the lower lip, which modulates the reed channel aperture.

There exists no computation methods so far which could be used to explain the perception of attack times in a large range of tone dynamics. As pointed out by Schaeffer, the perception of attack is a complex phenomenon which paradoxically seems to depend not only on the physical attack transient but also on the shape of the dynamics during the successive phases of sounds (Schaeffer, 1966, p. 226 & 229). Gordon (1987) tested various models for attack times based on the amplitude envelope. The most successful model in comparison to perceptual measurements defined the Attack Time as the time the amplitude envelope takes to go beyond a certain threshold relative to its maximum value. Krimphoff et al. (1994) also used a threshold to account for the perception of the beginning of a sound. The expression used here for the Attack Time descriptor (AT) takes both thresholds into account:

$$AT = t_{e_{AT}} - t_{s_{AT}} \quad (2)$$

where $t_{s_{AT}}$ and $t_{e_{AT}}$ are the start and the end of the attack times, corresponding to the times at which the amplitude envelope reaches 10% and 90% of the maximum amplitude, respectively (the thresholds have been defined as in Peeters, 2004). This descriptor was found to be strongly correlated with the first dimension of the clarinet timbre space obtained in (Barthet, Guillemain, et al., 2010).

The Spectral Centroid (SC) In order to find an acoustical descriptor predicting the distribution of sounds along one of the dimension of a perceptual timbre space, Grey and Gordon (1978) proposed to characterize numerically the spectral energy distribution by its mean (centroid or "balance point"). This parameter, later called Spectral Centroid (SC), is a good index to our ability to distinguish broad differences in timbre between various musical instruments (Grey, 1977; Krumhansl, 1989; McAdams et al., 1995), as well as the finer differences in timbre produced by the same instrument (Traube, 2004; Loureiro, Paula, & Yehia, 2004; Barthet, Guillemain, et al., 2010). Lichte (1941) probably made the first reference to the perceptual sensation associated with the Spectral Centroid (*brightness*) in a study on verbal descriptions of timbre[1]. Kendall and Carterette (1996) indeed observed that the Spectral Centroid accurately maps a perceptual quality which the authors called *nasality* (also referred as *brightness* or *sharpness*). Independently of the hearing level, the brightness may explain how the ear recognize tones played *piano*, which are generally perceived as dull and mellow, or on the contrary *forte*, which are generally perceived as bright and harsh (Risset, 1991).

In the context of musical performance, the Spectral Centroid might be a relevant means of accounting for the timbre variations produced by performers because it has often been found to be closely related to variations in the musician's control gestures. For example, Traube (2004) observed that the point at which a guitar string is plucked strongly affects the Spectral Centroid of the resulting sound: the closer to the middle of the string the pluck occurs, the lower the Spectral Centroid of the guitar tone, and the less bright the tone will be. In the case of the clarinet, the playing technique chosen by the performer, which is linked to choices about the instrument itself (such as the bending and length of the mouthpiece table and the strength of the reed) is known to affect the brightness of the tones. The "French" technique, where the clarinetist takes only a small part of the mouthpiece into his/her mouth tends to brighten the tones, whereas the "German" technique where a much larger part of the mouthpiece is held inside the mouth, generally yields less bright tones (Fritz, 2004, p. 175). As shown in previous studies, even when a same playing technique is used, modulations of Spectral Centroids can be obtained by varying the control parameters of the instrument: a monotonous increase in the Spectral Centroid was observed as the mouth pressure and reed aperture increased, in the case of clarinet tones synthesized with a physical model (Helland, 2004; Barthet, Guillemain, Kronland-Martinet, & Ystad, 2005; Barthet, 2008). Listeners use changes in the Spectral Centroid to discriminate between different clarinet timbres, both in the cases of synthetic (Barthet, Guillemain, et al., 2010) and natural (Loureiro et al., 2004) tones.

We have assumed the existence of a link between brightness and the perception of tension. As the latter is known to be an important aspect of musical composition, which usually consists of series of tensions and releases, this suggests that brightness variations may provide performers with a means of communicating the musical structure to the listeners. In the context of music sequential integration, Wessel (1979) put forward the idea that differences in brightness could surprisingly induce melodic segregation in much the same way as differences in pitch (see also Bregman, 1994).

Different ways of defining the Spectral Centroid are given in the literature, depending on the amplitude and frequency scales adopted, and whether physiological auditory models are used. Although the Spectral Centroid descriptors based on physiological auditory data (e.g. the sharpness defined by Zwicker and Fastl (1990), or the descriptor proposed by Marozeau, Cheveigné, McAdams, and Winsberg (2003) based on the partial loudness measurements, and an Equivalent Rectangular Band-rate scale) have increased the correlations between perceptual and acoustical data in timbre studies, these improvements are rather small in comparison with the predictions obtained using methods lowering the cost of computation based on the Fourier analysis (see Grey & Gordon, 1978; Marozeau et al., 2003, for comparisons). As a means of analysis and synthesis, the Spectral Centroid has been efficiently used by Beauchamp (1982) to determine the parameters of a nonlinear/filter synthesis model via an automatic analysis procedure. The latter author's definition was used here to compute the short-term Spectral Centroid (SC):

$$SC(n) = \frac{\sum_{k=1}^{K} f(k) A_n(k)^2}{b_0 + \sum_{k=1}^{K} A_n(k)^2} \tag{3}$$

where $A_n(k)$ is the magnitude of the k^{th} coefficient of the Discrete Fourier Transform (DFT) associated with the frame centered at time n, $f(k)$ is the frequency associated with the k^{th} spectral component, K denotes the last frequency bin to be processed, and b_0 is a positive amplitude threshold forcing the descriptor to decrease at very low amplitudes when noise predominates. As clarinet tones include both a deterministic part (corresponding to the harmonic signal resulting from the self-sustained oscillations) and a stochastic broadband part (resulting for instance from breath and key noises), the Spectral Centroid was calculated using a frequency scaling method that takes all the spectral components into account, and not only the harmonic partials. We also used a power amplitude scale in order to assign a greater weight to the dominant harmonics. Note that due to the stabilization term b_0, the values of SC can be smaller than the fundamental frequency of the tone.

In order to characterize the Spectral Centroid variations at note level, we calculated the mean value and the range of variations of the Spectral Centroid within the duration of a tone, which were denoted the Spectral Centroid Mean (SCM) and the Spectral Centroid Range (SCR), respectively. These parameters are defined by the following equations:

$$\begin{cases} SCM &= \dfrac{1}{IOI} \displaystyle\sum_{n=n_{on}}^{n=n_{off}} SC(n) \\ SCR &= \underset{n_{on} \leq n \leq n_{off}}{Max} (SC(n)) - \underset{n_{on} \leq t \leq n_{off}}{Min} (SC(n)) \end{cases} \tag{4}$$

where n_{on} and n_{off} are the onset and offset times of the note, respectively, and IOI is its duration. Note that these descriptors are independent: notes can potentially have the same SCMs but different SCRs, and vice versa.

[1] An effort has been made in this article to distinguish between the Spectral Centroid as a measure of the spectral distribution and the brightness as a perceptual attribute of sound.

The Odd/Even Ratio (OER) The Odd/Even Ratio, which is used to analyze harmonic or quasi-harmonic sounds, accounts for the difference in the relative energy between odd and even harmonics (see e.g. Peeters, 2004). The Odd/Even Ratio is particularly suitable for characterizing clarinet tones, since this instrument's "closed/open" cylindrical resonator is known to favor the odd harmonics at the expense of the even ones, which are very weak in this case (Benade & Kouzoupis, 1988). In (Barthet, Guillemain, et al., 2010), the Odd/Even Ratio was found to be strongly correlated with one of the dimensions of the perceptual timbre space of synthetic clarinet tones.

We have defined the time-varying Odd/Even Ratio (OER) by the following equation:

$$OER(t) = \frac{b_0 + \sum_{h=0}^{\frac{H}{2}-1} A_{2h+1}(t)^2}{b_0 + \sum_{h=1}^{\frac{H}{2}} A_{2h}(t)^2} \tag{5}$$

where $A_h(t)$ denotes the instantaneous amplitude of the h^{th} harmonic component. H is the total number of harmonics under consideration, which is assumed to be even in equation 5, so that an equal number of odd and even harmonics are compared. As with SC, an amplitude threshold b_0 was added to the classical definition of the OER to prevent the descriptor from tending to infinity when noise predominates. Note that OER is dimensionless. $OER < 1$ indicates that the even harmonics predominate, whereas $OER > 1$ indicates that the odd harmonics predominate. As with the Spectral Centroid, the following two note level timbre descriptors were defined, based on the Odd/Even Ratio: the Odd/Even Ratio Mean (OERM) and the Odd/Even Ratio Range (OERR). These parameters were computed in the same way as in equation 4.

Dynamics descriptor.

The Root Mean Square (RMS) envelope was used to characterize the changes in the acoustical energy. This parameter has been classically defined as follows:

$$ENV(n) = \sqrt{\frac{\sum_{k=1}^{K} A_n(k)^2}{N}} \tag{6}$$

where the various quantities are defined as in equation 3 and N is the number of points used to calculate the Discrete Fourier Transform. As with SC and OER, we computed the mean value and the range of variation of the envelope during each of the tones, which were denoted ENVM and ENVR, respectively.

Pitch descriptor.

The pitch of complex harmonic tones is closely linked to the fundamental frequency (Terhardt, Stoll, & Seewann, 1982). The latter was used as a first approximation to characterize the pitch of clarinet tones. The instantaneous fundamental frequency F0 was obtained using the method developed in (Jaillet, 2005), which involves detecting spectral peaks in the time-frequency plane, using a global optimization process. This method is implemented in the LEA software program produced by Genesis (2008). The mean value and the range of variation of the fundamental frequency during a tone will be denoted F0M and F0R, respectively.

Computation of the descriptors.

The Short Time Discrete Fourier Transform was computed using a 1024-point Hann window (approximately 20 ms at the sampling frequency of 44.1 kHz) with a 50% overlap. b_0 was set at a value giving a spectral dynamic of 60 dB. In order to compute the Odd/Even Ratio, each tone was analyzed using a bank of bandpass filters, the frequencies of which matched the frequencies of the tone components (which correspond to a harmonic series in the case of sustained clarinet sounds). This provided us with short-band analytic signals associated with the frequency components of the tone. The instantaneous amplitude and phase of the tone components were then obtained from the short-band analytic signals (see for example Picinbono, 1997).

Synchronization of the descriptors

As changes in IOI occurred between the various performances, a time synchronization procedure had to be performed at note level, to be able to compare the descriptors SC, OER, ENV and F0. For this purpose, a time-warping procedure was carried out on the descriptors. The temporal profiles of the descriptors associated with each tone were shortened or lengthened using cubic spline interpolation methods, so that the new durations corresponded to the mean IOI based on the repeated performances. Note that this time-warping procedure was used by Wanderley (Wanderley, 2002), for instance, to examine the regularity of clarinetists' spatial movements.

Statistical analyses

Reproducibility of the expressive deviations.

In order to determine the level of similarity between the expressive deviations observed during performances played with the same expressive intentions (i.e. between all mechanical to mechanical and all expressive to expressive performances), Pearson product-moment correlations (r) were computed on the various time, frequency and energy descriptors across the repeated performances.

Comparison between mechanical and expressive performances.

The analyses described above show the consistency of the acoustical parameters observed with a given musical intention, but they cannot be used to test whether any differences occur when the player's intentions change. In order to test whether the descriptors change depending on the performer's expressive intentions, two-way analyses of variance (ANOVA) were conducted with the player's expressive intentions and the note factors as independent variables. The dependent variables were the note level values of the descriptors (ΔIOI, AT, SCM, SCR, OERM, OERR, ENVM, ENVR, F0M, F0R). These statistical analyses show for all the descriptors the one-way effects of the player's expressive intentions and the note factors, and the two-way effect of interaction effect between the main factors. The magnitudes of the effects were estimated by using the partial eta squared (η^2) index of effect size. The definitions in (Cohen, 1977, p. 285) have been adopted to discuss the effect sizes (small effect size: $\eta^2 = .01$, medium effect size: $\eta^2 = .06$, large effect size: $\eta^2 = .14$). When interactions were observed, a multiple comparison procedure (MCP) based on the Holm-Sidak sequential procedure (Holm, 1979) was conducted to identify which tones in the musical sequence differed significantly between the mechanical and expressive performances. The Holm-Sidak procedure was used here as it is more powerful than non sequential multiple comparison tests, such as Bonferroni's, or Sidak's tests (Ludbrook, 1998).

An alpha level of .05 was used for all statistical tests.

Results

Reproducibility of the expressive deviations

The mean correlations (r) within the mechanical and expressive performances were computed for each descriptor (ΔIOI, AT, SC, OER, ENV and F0). For the Bach excerpt, these correlations were on average .93 for the mechanical performances (minimum: $r = .81$, $p < .001$) and .91 for the expressive performances (minimum: $r = .80$, $p < .001$). For the Mozart excerpt, the correlations were on average .93 for the mechanical performances (minimum: $r = .83$, $p < .001$) and .90 for the expressive performances (minimum: $r = .78$, $p < .001$). Hence, for both excerpts, ΔIOI, AT, SC, OER, ENV and F0 were highly correlated across performances played with the same musical intention. These results show that the performer consistently repeated the patterns linked to time, frequency, and energy, whenever the same interpretative strategy was used.

Influence of the note factor

(Insert Table 2)

The results of the two-way analyses of variance conducted on the various note-level descriptors for the Bach and Mozart performances are presented in Table 2. It can be seen from this table that the note factor had a highly significant effect on all the descriptors, both for the Bach and Mozart performances. The effect sizes were found to be very large ($M = .92$, $.80 \leq \eta^2 \leq 1.00$). These results show that the values of the time, frequency and energy descriptors varied according to the musical characteristics of the notes (such as pitch and duration) and/or their location in the musical structure. The influence of the note factor can be explained straight-forwardly for the descriptors which are by definition correlated to the notes' characteristics, i.e. duration (ΔIOI) and pitch (F0M, SCM). The intrinsic mechanical and acoustical properties of the clarinet (for example, the visco-thermal losses increase with frequency) also explain why the timbre descriptors can vary with pitch. According to the model of musical communication proposed in (Kendall & Carterette, 1990), the variability of the time, frequency and energy descriptors according to the note factor could also be related to the choice of controls made by the performer in order to communicate the musical structure to the listener.

Comparison between mechanical and expressive performances

Temporal descriptors.

ΔIOI The ANOVA showed a strong effect of the player's expressive intention on the IOI deviation descriptor both for the Bach and Mozart performances, $F(26, 1026) = 257.36$, $p < .001$, and $F(74, 300) = 54.49$, $p < .001$, respectively. The interaction between the expressive intention and the note factors was found to be highly significant for both excerpts (see Table 2) with large effect sizes ($\eta^2 = .30$ and .40 for the Bach and Mozart excerpts, respectively). As significant interactions were found between the main factors, Multiple Comparison Procedures were performed. The results of the MCPs indicate that many tones (13 out of 27 and 13 out of 75, for the Bach and Mozart excerpts, respectively) lasted significantly longer than nominal in the expressive performances (see Figure 2, and Tables 3 and 4, in the Appendix). Indeed, for both excerpts, the expressive performances lasted longer than the mechanical performances, on average (Bach: $M = 10.75$ s, $SD = 0.21$ and $M = 9.79$ s, $SD = 0.13$, respectively ; Mozart: $M = 72.46$ s, $SD = 0.89$ and $M = 67.80$ s, $SD = 0.64$, respectively). Consequently, the average tempi of the expressive performances were slower than the ones of the mechanical performances (Bach: $M = 40.47$ bpm, $SD = 0.78$ and $M = 44.45$ bpm, $SD = 0.57$, respectively ; Mozart: $M = 42.23$ bpm, $SD = 0.43$ and $M = 45.13$ bpm, $SD = 0.43$, respectively), which were closer to the reference tempo (Bach: 48 bpm; Mozart: 44 bpm). This is not surprising as the mechanical performances were played in keeping with the metronome. In the case of the Bach excerpt, the shape of the IOI deviation pattern (ΔIOI) was very similar between the two interpretations (Figure 2). This shows that in both mechanical and expressive interpretations, the durations of the notes were lengthened or shortened with respect to the theoretical score indications, but the changes were more pronounced when the piece was played expressively. This pattern has often been reported to occur in studies on timing variations in musical performance (see e.g. Penel & Drake, 2004). For instance, the final ritardando in the Bach excerpt (N23 to N27) occurred in both interpretations, but it was more pronounced in the expressive performances.

(Insert Figure 2)

(Insert Figure 3)

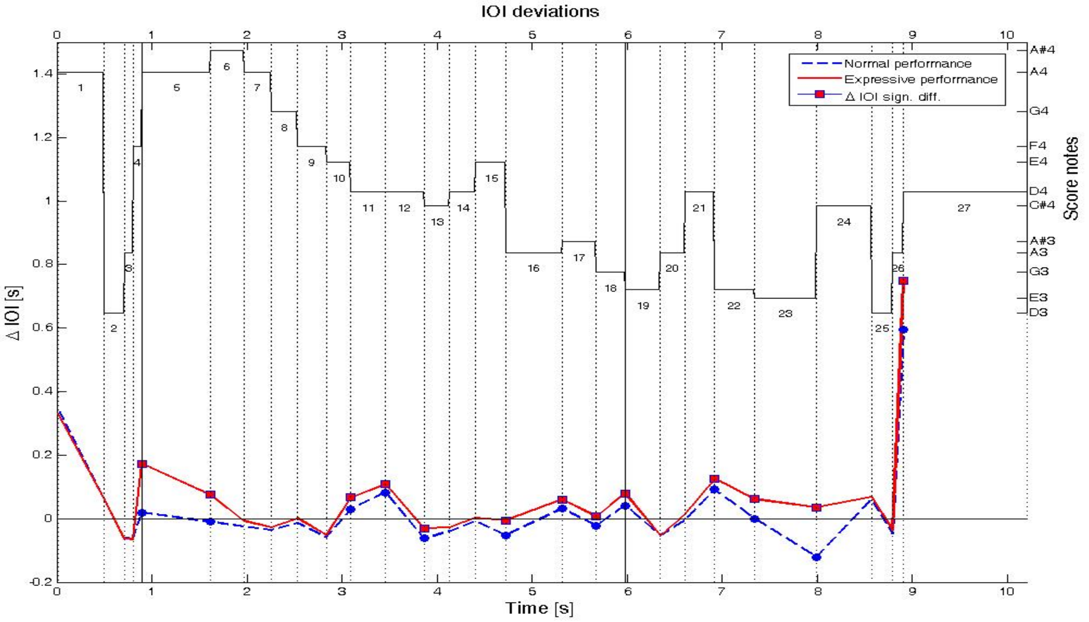


Figure 2. Comparison between the mean Intertone Onset Interval deviations (ΔIOI) measured in the mechanical (dashed line) and expressive (solid line) performances of the excerpt from Bach's *Suite no. II*. The notes with which the multiple comparisons showed the existence of significant differences are indicated by circular and square markers. The dashed vertical lines correspond to the onsets and offsets of the tones. The thin vertical lines indicate the bars. The thick vertical lines indicate the beginnings and ends of the various musical phrases. The notes on the score are displayed at the top of the figure, along with their ranks in the musical sequence.

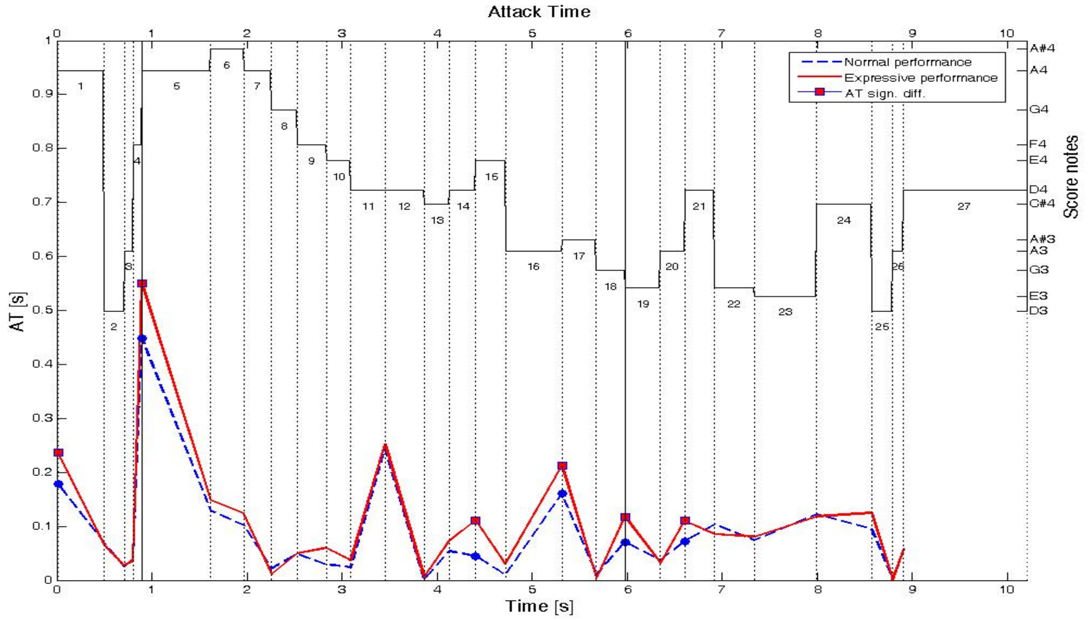


Figure 3. Average Attack Times in the mechanical (dashed line) and expressive (solid line) interpretations of the Bach excerpt. The notes and groups of notes with which the statistical analyses showed the existence of significant effects are indicated by circular and square markers. For other explanations, see the legend to Figure 2.

Table 2
Results of the two-way analyses of variance for the various note-level descriptors (ΔIOI, AT, SCM, SCR, OERM, OERR, ENVM, ENVR, F0M, F0R) for the Bach and Mozart performances.

		Bach				Mozart			
	Source	*df*	*F*	η^2	*p*	*df*	*F*	η^2	*p*
ΔIOI	Exp.	1	257.36^{***}	.20	<.001	1	54.49^{***}	.15	<.001
	Note	26	663.22^{***}	.94	<.001	74	70.28^{***}	.95	<.001
	Exp. x Note	26	16.80^{***}	.30	<.001	74	2.68^{***}	.40	<.001
	Error	1026	(0.001)			300	(0.006)		
AT	Exp.	1	33.47^{***}	.03	<.001	1	28.23^{***}	.09	<.001
	Note	26	159.31^{***}	.80	<.001	74	29.40^{***}	.88	<.001
	Exp. x Note	26	2.87^{***}	.07	<.001	74	1.30	.24	.07
	Error	1026	(0.003)			300	(0.027)		
SCM	Exp.	1	24.87^{***}	.02	<.001	1	96.75^{***}	.24	<.001
	Note	26	914.67^{***}	.96	<.001	74	128.43^{***}	.97	<.001
	Exp. x Note	26	6.33^{***}	.14	<.001	74	3.45^{***}	.46	<.001
	Error	1026	(1450.7)			300	(2320.25)		
SCR	Exp.	1	0.44	.00	.51	1	24.15^{***}	.08	<.001
	Note	26	364.55^{***}	.90	<.001	74	34.74^{***}	.90	<.001
	Exp. x Note	26	6.51^{***}	.14	<.001	74	1.98^{***}	.33	<.001
	Error	1026	(2173.42)			300	(6932.37)		
OERM	Exp.	1	54.86^{***}	.05	<.001	1	190.84^{***}	.39	<.001
	Note	26	554.83^{***}	.93	<.001	74	33.07^{***}	.89	<.001
	Exp. x Note	26	6.47^{***}	.14	<.001	74	5.00^{***}	.55	<.001
	Error	1026	(0.41)			300	(0.48)		
OERR	Exp.	1	2.84	.00	.09	1	5.22^{*}	.02	.02
	Note	26	197.68^{***}	.83	<.001	74	30.96^{***}	.88	<.001
	Exp. x Note	26	3.97^{***}	.09	<.001	74	3.21^{***}	.44	<.001
	Error	1026	(1.06)			300	(0.94)		
ENVM	Exp.	1	125.02^{***}	.11	<.001	1	308.16^{***}	.51	<.001
	Note	26	522.56^{***}	.93	<.001	74	49.25^{***}	.92	<.001
	Exp. x Note	26	5.62^{***}	.13	<.001	74	2.10^{***}	.34	<.001
	Error	1026	(.001)			300	(0.001)		
ENVR	Exp.	1	41.38^{***}	.04	<.001	1	0.31	.00	.58
	Note	26	702.57^{***}	.95	<.001	74	27.23^{***}	.87	<.001
	Exp. x Note	26	1.64^{*}	.04	.02	74	1.85^{***}	.31	<.001
	Error	1026	(.001)			300	(0.002)		
F0M	Exp.	1	5.22^{*}	.01	.02	1	50.53^{***}	.14	<.001
	Note	26	72978.64^{***}	.99	<.001	74	192457.77^{***}	1.00	<.001
	Exp. x Note	26	2.43^{***}	.06	<.001	74	0.59	.13	.99
	Error	1026	(4.73)			300	(0.973)		
F0R	Exp.	1	1.16	.00	.28	1	0.002	.00	.96
	Note	26	320.17^{***}	.89	<.001	74	517.87^{***}	.99	<.001
	Exp. x Note	26	2.62^{***}	.06	<.001	74	0.64	.14	.99
	Error	1026	(825.62)			300	(486.38)		

Note. The expressive intention and note factors are denoted Exp. and Note, respectively. Results enclosed in parentheses represent mean square errors. η^2 *is the partial eta squared measure of effect size.* $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Pitch descriptors.

F0M and F0R The player's expressive intention significantly affected the mean fundamental frequency F0M both for the Bach and Mozart excerpts ($F(26, 1026) = 5.22$, $p = .02$ and $F(74, 300) = 50.53$, $p < .001$, respectively), but not the fundamental frequency range F0R (Table 2). The effect of interaction between the expressive intention and the note factors was found to be significant only in the case of the Bach performances with a medium effect size both for F0M and F0R ($\eta^2 = .06$). However, the results of the MCP on the Bach excerpt (see Table 3) showed that only one tone in the case of F0M (N4) and two tones in that of F0R (N1 and N24) showed significant differences when the player's intentions changed. In addition, the differences in F0M and F0R were due either to instabilities in the descriptor F0 at the onset and/or offset of the tones (a phenomenon induced by the method of analysis), or were very small. The maximum F0 difference in the sustained part of the tones was approximately 1 Hz (note N1), which was close to the frequency discrimination threshold measured in the case of pure tones (1 Hz in the case of a pure 200-Hz tone presented at 40 dB according to Wier, Jesteadtt, & Green, 1977). However, informal listening made by the authors did not reveal noticeable changes of pitch between the sequences. Based on these results, it appears that the contribution of pitch to the expression of the playing intention was weak at best in the case of the Bach excerpt, and non significant in the case of the Mozart excerpt.

Timbre descriptors.

AT Highly significant effects of the player's expressive intentions on the tones' Attack Time (AT) were found for both the Bach and Mozart excerpts, $F(26, 1026) = 33.47$, $p < .001$ and $F(74, 300) = 28.23$, $p < .001$, respectively. However, the interaction between the expressive intention and the note factors was only significant for the Bach performances, $F(26, 1026) = 2.87$, $p < .001$, with a medium effect size ($\eta^2 = .07$). The *post hoc* analyses conducted for the Bach excerpt showed that 6 tones had significantly higher ATs in the expressive performances than in the mechanical ones (cf. Figure 3 and Table 3).

SCM and SCR The ANOVA showed that the effect of the player's expressive intention on the Spectral Centroid Mean SCM was highly significant for both the Bach and Mozart excerpts, $F(26, 1026) = 24.87$, $p < .001$ and $F(74, 300) = 96.75$, $p < .001$, respectively. For the Spectral Centroid Range SCR, the one-way effect of the expressive intention was only highly significant for the Mozart excerpt, $F(74, 300) = 24.15$, $p < .001$. However, for both excerpts, strong interactions between the expressive intention and the note factors occurred for SCM and SCR (Table 3). The effect sizes of the interaction between the expressive intention and the note factors were found to be large for both the Bach and the Mozart excerpts, although larger for the latter (Bach: $\eta^2 = .14$ for both SCM and SCR; Mozart: $\eta^2 = .46$ for SCM and $\eta^2 = .33$ for SCR). The results of the MCPs show that the mean and/or the range of the Spectral Centroid values were significantly different between the expressive and the mechanical performances in a large number of notes for both excerpts (Bach: 14 out of 27, as shown in Figure 4(a) and Table 3, and Mozart: 33 out of 75, as shown in Figure 4(b) and Table 4). In order to evaluate if such Spectral Centroid changes would be noticeable from the perceptual point view, we used the mean difference threshold (Just Noticeable Difference) in SC reported in (Kendall & Carterette, 1996) as a reference since it was obtained from perceptual experiments with human listeners. To address this issue, a F0-normalized Spectral Centroid ($\overline{SC}$) was computed as in (Kendall & Carterette, 1996). This was done by using a linear amplitude scale, $b_0 = 0$ and dividing by F0 in equation 3. For the Bach excerpt, the $\overline{SC}$ differences were higher than the JND threshold (0.117) for 13 out of the 14 tones for which significant differences of SCM and SCR were reported ($0.04 \leq \Delta\overline{SC} \leq 2.9$). For the Mozart excerpt, the $\overline{SC}$ differences were higher than the JND threshold for 27 out of the 33 tones for which significant differences of SCM and SCR were reported ($0.01 \leq \Delta\overline{SC} \leq 1.95$).

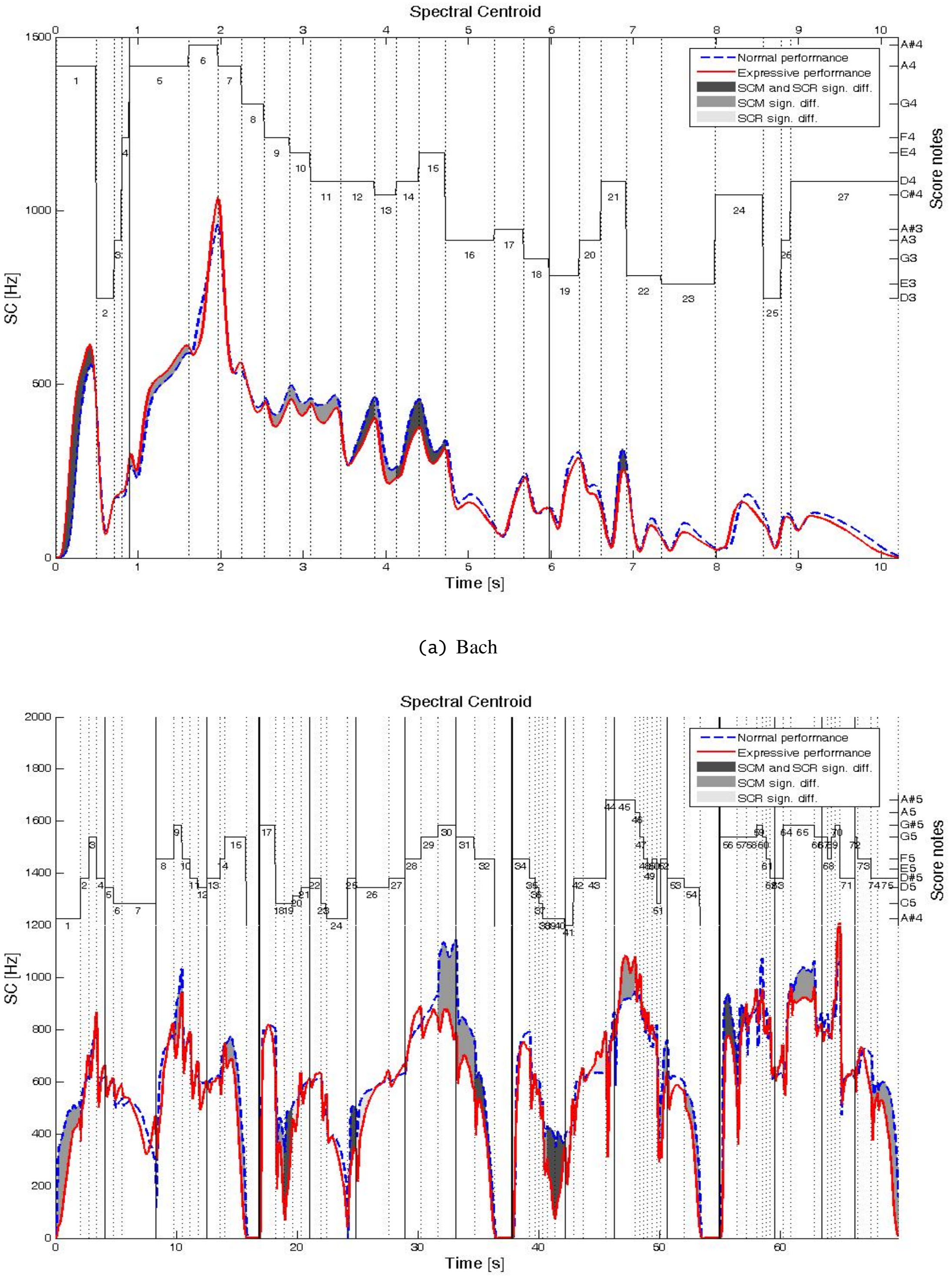
In both excerpts, the changes observed in the Spectral Centroid with the expressive performances depended on the musical structure. In some parts of the musical phrases, the SCM was higher in the expressive than in the mechanical performance: for instance, at the beginning of the phrase, from notes N1 to N5, for the Bach excerpt, or in the third musical phrase, at the 12^{th} bar, for the Mozart excerpt. In other parts, the opposite pattern occurred: for instance, in the middle of the phrase, from notes N9 to N15, for the Bach excerpt, and from the 7^{th} to 9^{th} bar, for the Mozart excerpt. Upon listening informally to the Bach performances, the authors noted that the sequence from notes N9 to N15 sounded mellower in the expressive performances than in the mechanical ones (see Sound Examples *1a* and *1b*).

It is worth noting that the significant changes in SC detected in the expressive rendering were not due to changes in F0 since no significant effects of the expressive intentions were observed on the fundamental frequency F0 (except possibly for only one tone in the Bach excerpt (N1), see the discussion on the pitch descriptors, above). Furthermore, within a musical phrase, some tones with the same F0 had very different Spectral Centroid values. For instance, in the mechanical performances of the Bach excerpt, notes N11, N12 and N27, which all corresponded to a D4 ($\overline{F0} \approx 293.67$ Hz), had significantly different SCM values (448.93, 355.72 and 82.07 Hz[2], respectively): N11 vs N12, $t(351) = 21.45$, $p < .001$, N11 vs N27, $t(351) = 30.46$, $p < .001$, and N12 vs N27, $t(351) = 21.45$, $p < .001$. These findings confirm that the Spectral Centroid variations depend on the position of the notes in the musical structure.

(Insert Figure 4)

OERM and OERR As for the other timbre descriptors, the results of the ANOVA showed that the effect of expressive intention on the Odd/Even Ratio Mean OERM were highly significant for both the Bach and the Mozart excerpts, $F(26, 1026) = 54.86$, $p < .001$, and $F(74, 300) = 190.84$, $p < .001$, respectively. For the Odd/Even Ratio Range OERR, the one-way effect of expressive intention was non significant for the Bach excerpt, $F(26, 1026) = 2.84$, $p = .09$, and weakly significant for the Mozart excerpt,

[2] The fact that SCM can be smaller than F0 is due to the stabilization term b_0 in equation 3.



(a) Bach

(b) Mozart

Figure 4. Average short-term Spectral Centroid in the mechanical (dashed line) and expressive (solid line) interpretations of the Bach (a) and Mozart (b) excerpts. The notes and groups of notes with which statistical analyses showed the existence of significant effects are shown in gray. The dark gray areas indicate significant differences in both the Spectral Centroid Mean (SCM) and Spectral Centroid Range (SCR). The pale gray areas indicate significant differences in SCM only. The light gray areas indicate significant differences in SCR only. For other explanations, see the legend to Figure 2.

$F(74, 300) = 5.22$, $p = .02$ ($\eta^2 = .02$), which is probably due to the strong one-way effect of the note factor (Table 2). Indeed, the two-ways effects of interaction between the expressive intention and the note factors were highly significant both for OERM and OERR, in both excerpts. Again, the effect size of this interaction was larger for the Mozart excerpt ($\eta^2 = .55$ for OERM and $\eta^2 = .44$ for OERR) than for the Bach excerpt ($\eta^2 = .14$ for OERM and $\eta^2 = .09$ for OERR). The results of the MCPs showed that significant differences in OERM and/or OERR were observed with 12 notes for the Bach excerpt (see Figure 5(a) and Table 3), and with more than half of the notes (48 notes out of 75) for the Mozart excerpt (see Figure 5(b) and Table 4). Note that the Odd/Even Ratio was mostly greater than one, except during the attack parts of a few notes (see Figure 5), which points out, as was to be expected, the dominance of odd harmonics compared to even harmonics for clarinet tones.

(Insert Figure 5)

Dynamics descriptor.

ENVM and ENVR The effect of the performer's expressive intention on the mean value of the RMS envelope ENVM was found to be highly significant for both the Bach and the Mozart excerpts, $F(26, 1026) = 125.02$, $p < .001$, and $F(74, 300) = 308.16$, $p < .001$, respectively. Interactions between the expressive intention and the note factors were highly significant for ENVM in both the Bach and Mozart excerpts, with medium ($\eta^2 = .13$) and large effect sizes ($\eta^2 = .34$), respectively (Table 2). Regarding the range of variations of the RMS envelope ENVR, only a weakly significant effect of the expressive intention was found for the Bach excerpt, $F(26, 1026) = 41.38$, $p < .05$, with a small effect size ($\eta^2 = .04$). However, the two-way effects of interaction between the expressive intention and the note factors were significant for both the Bach and Mozart excerpts, with small ($\eta^2 = .04$) and large ($\eta^2 = .31$) effect sizes, respectively. The multiple comparison procedure (see Table 3) showed the existence of significant differences in the ENVM and ENVR values with most of the notes (20 notes in all for the Bach excerpt, and 43 notes for the Bach excerpt).

Relationships between the descriptors.

Is is worth noting that for the Bach excerpt, the short notes (such as the grace notes N2, N3, N26) did not generally show any significant differences in ΔIOI, AT, SC, or OER, possibly due to the fact that when playing short notes, the performer did not have enough time to make expressive timing or timbre variations.

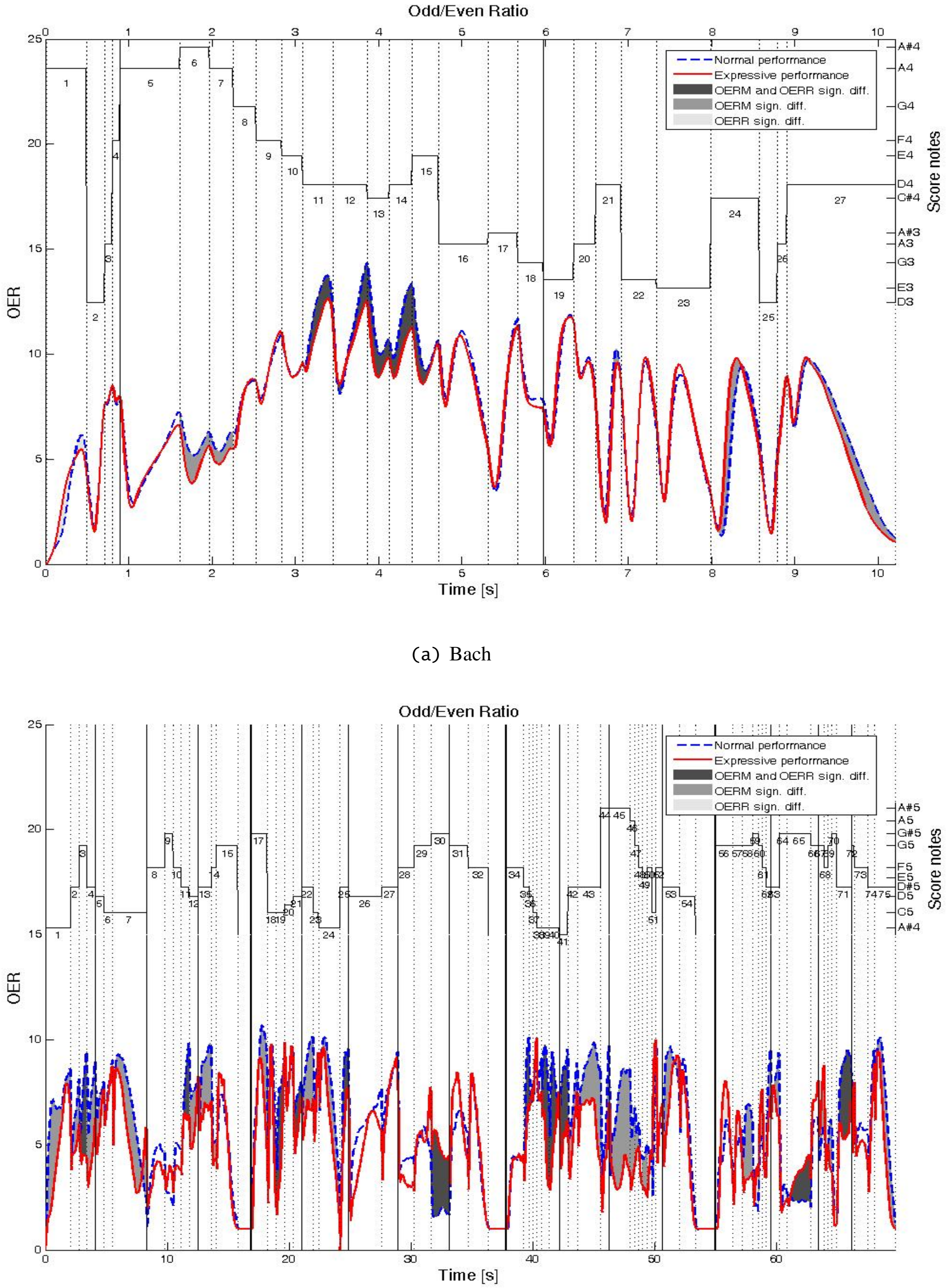
Some notes systematically showed disparities in terms of both the timbre and the temporal descriptors. For instance, for the Bach excerpt, the first note in the first bar (N5) after the anacrusis (N1) and the grace notes (N2 to N4) lasted longer, and had a longer Attack Time, and a higher Spectral Centroid in the more expressive interpretation. It is worth noting that this note plays an important role in the musical structure because it is the first in the musical phrase. Timbre-related changes and timing variations of this kind may be used by the performers to emphasize the importance of a specific tone by increasing the musical tension locally.

The correlations (Pearson) between the spectral timbre descriptors and the acoustical energy were also analyzed, as the perceptual effects of these features have been studied in (Barthet, Depalle, Kronland-Martinet, & Ystad, 2010). Figure 6 shows comparisons between the average Spectral Centroid, Root Mean Square envelope and Odd/Even Ratio patterns, computed from the expressive performances of the the Bach and Mozart excerpts. The Spectral Centroid was closely correlated in a linear way with the RMS envelope for both the Bach ($r = .82$, $p < .0001$) and Mozart ($r = .92$, $p < .0001$) excerpts. As explained by the Worman-Benade laws, the spectral richness of acoustical wind instrument tones is highly sensitive to the blowing pressure, which is correlated with the resulting acoustical energy output (Benade & Kouzoupis, 1988). Conversely, the relationship between the Odd/Even Ratio and the RMS envelope was more complex. For some notes, the differences in the Odd/Even Ratio depending on the levels of expression were in line with those in the Spectral Centroid (for instance, with OERM and SCM in the passage from N11 to N15 of the Bach excerpt, see Table 3). However, this was not systematic (for instance, notes N6, N7, N24, and N27 of the Bach excerpt showed significant differences in the OERM but not in the SCM values, see Table 3). The linear correlation between OER and ENV was very weak for the Bach ($r = .08$, $p < .05$) and Mozart ($r = .41$, $p < .0001$) excerpts. For the Bach excerpt, OER and ENV were linearly correlated at the beginning of the phrase, when the acoustical energy was high. Then, from note N8 onwards, a non-linear jump in the OER occurred towards higher values when the acoustical energy decreased. The overall increase in OER observed during the decrescendo phase was linked to the fact that when the playing level is weak, clarinet tones are nearly sinusoidal and contain very few even harmonics (this also explains why the Spectral Centroid decreased during the decrescendo). For the Mozart excerpt, OER and ENV were sometimes highly correlated (as in the case of note N27 in the second phrase), and they sometimes showed opposite variations (as in the case of note N45 in the third musical phrase).

(Insert Figure 6)

Discussion

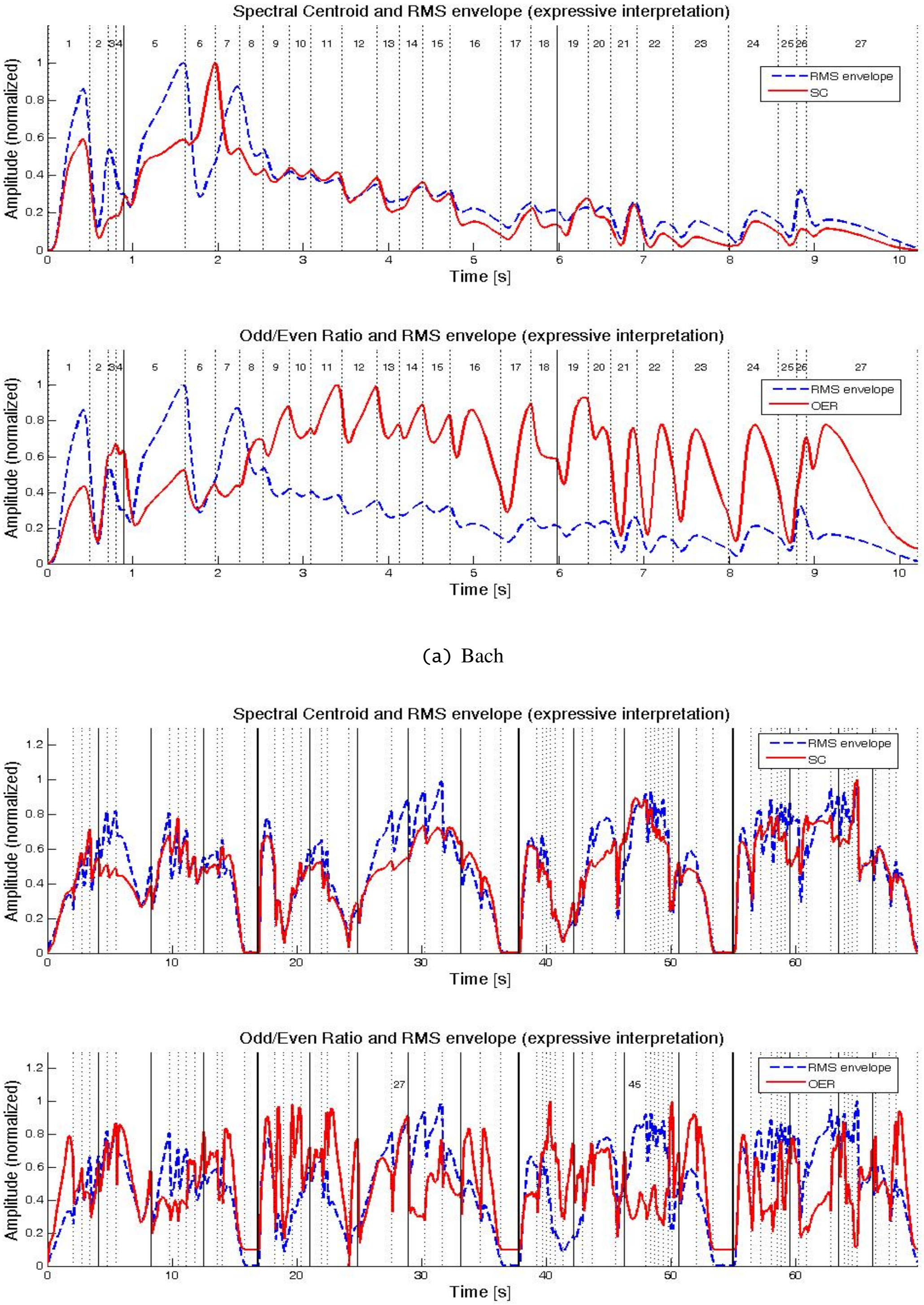
Several conclusions can be drawn from the results of these experiments. First, the changes in the temporal, timbre, dynamics and pitch descriptors were found to be highly consistent when the performer's expressive intentions were



(a) Bach

(b) Mozart

Figure 5. Average short-term Odd/Even Ratio in the mechanical (dashed line) and expressive (solid line) interpretations of the Bach (a) and Mozart (b) excerpts. The notes and groups of notes with which statistical analyses showed the existence of significant effects are shown in gray. The dark gray areas indicate significant differences in both the Odd/Even Ratio Mean (OERM) and Odd/Even Ratio Range (OERR). The pale gray areas indicate significant differences in OERM only. The light gray areas indicate significant differences in OERR only. For other explanations, see the legend to Figure 2.



(a) Bach

(b) Mozart

Figure 6. Comparison between the averaged Spectral Centroid (SC) and Odd/Even Ratio (OER) variations and the Root Mean Square envelope (ENV) in the case of the expressive interpretation of the Bach (a) and Mozart (b) excerpts. The values of SC, OER and ENV have been normalized with respect to their maximal values, for display purposes.

the same. These findings confirm that expert performers are able to faithfully reproduce their expressive deviations at a given level of expressiveness. The behavior of the timbre descriptors is therefore not random but seems on the contrary to follow rules of interpretation. This suggests that the performer used the acoustical parameters associated with timbre to vary his interpretation. This would certainly be true if changes in the timbre descriptors occurred when the level of expression changed. The above statistical analyses show in fact that most of the time, strong interactions occured between the performer's expressive intentions and note factors, for the timbre descriptors. Although the timbre-related changes observed in the case of the Bach excerpt were quite subtle, they were found to be significant (medium to large effect sizes) in the case of 18 notes out of 27 (upon combining the AT, SCM, SCR, OERM, and OERR data). The changes in the timbre descriptors made for expressive purposes were higher in the performances of the Mozart excerpt (58 notes out of 75 showed significant differences with large effect sizes, when the SCM, SCR, OERM, and OERR data were combined). This piece contains many more long notes (half notes, quarter notes, dotted quarter notes) than in the Bach excerpt, which mainly contains semi-quavers notes. In other words, a close control of the tone quality may be possible and worthwhile only when the duration of the notes is sufficiently long.

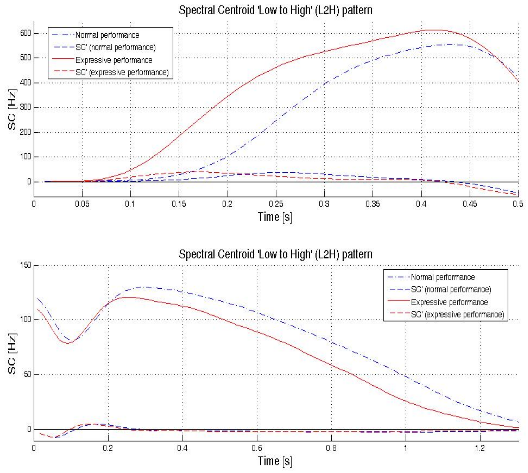


Figure 7. Examples of Spectral Centroid 'Low to High' (L2H) (top) and 'High to Low' (H2L) (bottom) patterns. The SC patterns and their derivatives SC$'$ correspond to tones in the mechanical and expressive performances of the Bach excerpt.

Several kinds of Spectral Centroid patterns were found to occur recurrently. For instance, notes N1, N6, N12, N14 in the Bach excerpt showed an increasing SC trend, which we have called the 'Low to High' (L2H) pattern, whereas notes N16 and N27 showed a decreasing SC trend, which we have called the 'High to Low' (H2L) pattern. During notes N22 to N24, the SC successively increased and decreased, corresponding to what we have called the 'Low to High to Low' (L2H2L) pattern. Figure 7 gives some examples of typical SC patterns at each of the two expressive levels in the case of the Bach excerpt. The L2H SC pattern corresponding to the first note in the sequence (N1) can be seen at the top of the figure, and the H2L SC pattern corresponding to the last note in the sequence (N27) is shown at the bottom. It is worth pointing out that the shape of the L2H pattern depended on the player's expressive intentions. The slope was higher at the beginning of the tone in the case of the more expressive performance (see the derivative SC$'$, which was higher during the first part of the tone in the expressive performance than in the mechanical one). The shape of the H2L pattern given in this example was roughly the same in both interpretations, as shown by the fact that the derivatives were almost the same (the curves differed in their means, but not in their range, see Table 3). In order to determine whether these variations in SC corresponded to audible changes of brightness, we used a transformation relying on an additive-based synthesis model for harmonic tones to freeze the Spectral Centroid of the tone at a given time. This was done by eliminating the change in shape of the spectral envelope over time as in the amplitude-envelope coherence transformation defined in (McAdams, Beauchamp, & Meneguzzi, 1999). The Spectral Centroids of the tones were frozen to their respective values at the beginning of the tone, after the attack part, and at the end of the tone, before the release part (cf. Sound Examples *3a* to *3c* corresponding to note N1, and *4a* to *4c* corresponding to note N27). These synthesized tones give the net changes of brightness between the beginning and the end of the tones. It is worth noting the corresponding changes in the RMS level which occurred, i.e. the L2H pattern was concurrent with an increase in the overall amplitude, and the H2L pattern was concurrent with a decrease in the overall amplitude. The question arises as to whether the timbre- or dynamic-related variations are more important from the perceptual point of view as means of conveying expressiveness. This point is addressed in our companion article (Barthet, Depalle, et al., 2010).

(Insert Figure 7)

Summary and conclusions

It was proposed in this study to investigate whether timbre, timing, dynamics and pitch descriptors show any changes when performers play expressively. To answer this question, mechanical and expressive clarinet performances of musical excerpts from the Baroque and Classical repertoire were compared. When the clarinetist repeated the performances with the same musical intentions, the timbre descriptors showed a high level of consistency. The performer also reproduced the Intertone Onset Interval deviations and Root Mean Square envelope highly consistently, as previously reported to occur in studies on timing and dynamics variations (see e.g. Palmer, 1996; Repp, 1992; Kendall & Carterette, 1990). The temporal and spectral timbre descriptors studied here (Attack Time, Spectral Centroid and Odd/Even Ratio), as well as the timing and dynamics descriptors, differed significantly between mechanical and expressive performances. The timbre-related changes across expressive levels did not occur at every note, but were specific to some notes, or groups of notes in the musical phrases (such as the first note in a phrase, or specific passages). The most conspicuous changes were in the mean Spectral Centroid and Odd/Even Ratio values and the range of variations in the duration of the tones. These changes seemed to be made more frequently in the case of long notes (such as half and quarter notes), possibly because a performer needs to have sufficient time to control the timbre while playing. According to the communication theory of musical expression proposed by Kendall and Carterette (1990) in the context of traditional Western art music, a musical message is first recoded from ideas to notation by the composer, then recoded from notation to acoustical signal by a performer, and finally recoded from acoustical signal to ideas by the listener. The findings reported in this study support this model of musical communication, since the distinct signatures across expressive levels showed by the timbre, timing and dynamics descriptors afford a basis for perceptual discrimination between the performances.

In line with the "musical expression" hypothesis put forward by Clarke (1988) to explain the role of expressive timing, it seems likely that timbre as well as timing may be used by performers to communicate the musical structure to the listeners, in addition to mediating expressiveness. For instance, by increasing the brightness of a tone while it is being held (as reflected in an increase in the Spectral Centroid), a performer may insist on the role of this specific tone in the musical structure. This might be so in the case of some notes in particular, which all showed specific SC patterns across the repetitions of mechanical and expressive performances, such as the 'Low to High' (L2H) and 'High to Low' (H2L) SC patterns.

These results suggest that, in the case of a self-sustained instrument such as the clarinet, there may exist a link between the process of musical interpretation and the way the acoustical correlates of timbre vary with time. Some timbre-related variations seem to be intrinsically linked to intensity variations due to the physical principles involved in the way the instrument works (cf. the strong correlation between Spectral Centroid and Root Mean Square envelope), whereas others are not linearly correlated (cf. the weak linear correlation between the Odd/Even Ratio and the RMS envelope). Further research is now required to determine whether the relationships between Spectral Centroid, Odd/Even Ratio and acoustical energy simply result from the intrinsic acoustical characteristics of the instrument, or whether they are deliberately induced by performers. The question as to whether performers intended to produce intensity or timbre variations is hard to answer. It is likely that these variations are both concomitantly sought by performers at the same time, but a specific experimental procedure would have to be designed to be able to address this issue.

Therefore, variations in the temporal and spectral parameters contributing to timbre are likely to be factors which account for the disparities and commonalities between performers along with timing variations. The next step will consist in comparing the variations in the timbre descriptors observed between performances of the same musical excerpt by several performers using the same musical instrument. This issue could also be investigated with performers playing different instruments, in order to establish whether there exist any similarities in the variations in the timbre descriptors between the performances of a clarinetist and a cellist, for example. In the present study, which was the first step towards elucidating the role of timbre in musical interpretation, we focused on the timbre-related differences occurring at single note level and in small groups of notes. As music usually involves larger phrases and movements, it might also be interesting to explore the variations in the timbre descriptors occurring at a higher level in the musical structure.

This study has shed light on some of the acoustical parameters which mediate a performer's expressive intentions, but the perceptual effects of these parameters were not investigated. In our companion article (Barthet, Depalle, et al., 2010), we have compared the perceptual effects of Spectral Centroid, Intertone Onset Interval, and acoustical energy variations on the musical appreciation of listeners. To address this issue, an analysis-by-synthesis approach was adopted in order to assess the effects of the expressive deviations detected in recorded clarinet performances.

Acknowledgments

We would like to thank the clarinetist Claude Crousier for participating in this project and for many fruitful discussions. We are grateful to Mitsuko Aramaki and Henri Burnet from the Institut des Neurosciences Cognitives de la Méditerranée and Jean Pierre Durbec from the Centre d'Océanologie de Marseille for their precious help. We also wish to thank the reviewers for their helpful comments and corrections. This project was partly supported by the French National Research Agency (ANR JC05-41996 "senSons", `http://www.sensons.cnrs-mrs.fr/`).

References

ANSI. (1960). USA Standard Acoustical Terminology. New York: American National Standards Institute.

Barthet, M. (2008). *De l'interprète à l'auditeur: une analyse acoustique et perceptive du timbre musical*. Unpublished doctoral dissertation, Université Aix-Marseille II.

Barthet, M., Depalle, P., Kronland-Martinet, R., & Ystad, S. (2010). Exploration of timbre, timing and dynamics in expressive clarinet performance by analysis and synthesis. *Music Perception (accepted for publication)*.

Barthet, M., Guillemain, P., Kronland-Martinet, R., & Ystad, S. (2005). On the relative influence of even and odd harmonics in clarinet timbre. In *Proc. Int. Comp. Music Conf. (ICMC'05)* (pp. 351–354). Barcelona, Spain.

Barthet, M., Guillemain, P., Kronland-Martinet, R., & Ystad, S. (2010). From clarinet control to timbre perception. *Acta Acustica (accepted for publication).*

Barthet, M., Kronland-Martinet, R., & Ystad, S. (2006). Consistency of timbre patterns in expressive music performance. In *Proc. 9th Int. Conf. on Digital Audio Effects (DAFx06)* (pp. 19–24). Montreal, Quebec, Canada.

Beauchamp, J. W. (1982). Synthesis by spectral amplitude and brightness matching of analyzed musical instrument tones. *Journal of the Audio Engineering Society*, *30*(6), 396–406.

Benade, A. H., & Kouzoupis, S. N. (1988, January). The clarinet spectrum: Theory and experiment. *Journal of the Acoustical Society of America*, *83*(1), 292–304.

Bregman, A. (1994). *Auditory scene analysis - the perceptual organization of sound.* Cambridge: MIT Press.

Castellengo, M., & Dubois, D. (2005). Timbre ou timbres ? Propriété du signal, de l'instrument, ou construction cognitive ? (Timbre or timbres ? Property of the signal, the instrument, or cognitive construction ?). In *Proc. of the Conf. on Interdisciplinary Musicology (CIM05).* Montréal, Québec, Canada.

Clarke, E. F. (1988). Generative principles in music performance. In J. A. Sloboda (Ed.), *Generative processes in music. The psychology of performance, improvisation and composition* (p. 1-26). Oxford: Oxford University Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Födermayr, F., & Deutsch, W. A. (1993). Parmi veder le lagrime. One aria, three interpretations. In *Proc. of the Stockholm Music Acoustics Conf.* (p. 96-107).

Fritz, C. (2004). *La clarinette et le clarinettiste: Influence du conduit vocal sur la production du son (The clarinet and the clarinetist: Influence of the vocal tract on the sound production).* Unpublished doctoral dissertation, Université Paris 6 et University of New South Wales.

Gabrielsson, A. (1999). The performance of music. In *Psychology of music* (2nd ed.). Academic Press.

Gabrielsson, A., & Lindstrom, B. (1995). Emotional expression in synthesizer and sentograph performance. *Psychomusicology*, *14*, 94-116.

Genesis. (2008). *Lea software.* `http://www.genesis-acoustics.com/`.

Gordon, J. W. (1987). The perceptual attack time of musical tones. *J. Acoust. Soc. Am.*, *82*(1), 88-105.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am.*, *61*, 1270-1277.

Grey, J. M., & Gordon, J. W. (1978). Perception of spectral modifications on orchestral instrument tones. *Computer Music Journal*, *11*(1), 24-31.

Hajda, J. M., Kendall, R. A., Carterette, E. C., & Harshberger, M. L. (1997). Methodological issues in timbre research. In I. Deliége & J. Sloboda (Eds.), *Perception and Cognition of Music* (2nd ed., p. 253-306). New York: Psychology Press.

Handel, S. (1995). Timbre perception and auditory object identification. In B. C. J. Moore (Ed.), *Handbook of perception and cognition* (2nd ed., p. 425-461). San Diego, California: Academic Press.

Helland, R. T. (2004). *Synthesis models as a tool for timbre studies.* Unpublished master's thesis, Norwegian University of Science and Technology, Trondheim, Norway.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, *6*, 65-70.

Jaillet, F. (2005). *Représentation et traitement temps-fréquence des signaux audionumériques pour des applications de design sonore.* Unpublished doctoral dissertation, Université de la Méditerranée - Aix-Marseille II.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770-814.

Kendall, R. A., & Carterette, E. C. (1990). The communication of musical expression. *Music Perception*, *8*(2), 129–164.

Kendall, R. A., & Carterette, E. C. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Perception*, *8*, 369-404.

Kendall, R. A., & Carterette, E. C. (1996). Difference thresholds for timbre related to spectral centroid. In B. Pennycook & E. Costa-Giomi (Eds.), *Proc. of the 4th international conference on music perception and cognition (ICMPC)* (p. 91-95). Montreal, Canada: Faculty of Music, McGill University.

Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caractérisation du timbre des sons complexes, II Analyses acoustiques et quantification psychophysique (Characterization of complex sounds timbre, II Acoustical analyses and psychophysical quantification). *Journal de Physique IV, Colloque C5*, *4*, 625-628.

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand ? In S. Nielzén & O. Olsson (Eds.), *Proc. of the Marcus Wallenberg Symposium held in Lund, Sweden* (p. 43-53). Amsterdam: Excerpta Medica.

Lichte, W. H. (1941). Attributes of complex tones. *Journal of Experimental Psychology*, *28*(6), 455-480.

Loureiro, M. A., Paula, H. B. de, & Yehia, H. C. (2004). Timbre classification of a single instrument. In *ISMIR 2004 5th International Conference on Music Information Retrieval.* Barcelona, Spain.

Ludbrook, J. (1998). Multiple comparison procedures updated. *Clin. Exp. Pharmacol. Physiol.*, *25*(12), 1032-1037.

Marozeau, J., Cheveigné, A. de, McAdams, S., & Winsberg, S. (2003, November). The dependency of timbre on fundamental frequency. *Journal of the Acoustical Society of America*, *114*(5), 2946-2957.

McAdams, S. (1994). La reconnaissance de sources et d'événements sonores (The recognition of sources and sound events). In S. McAdams & E. Bigand (Eds.), *Penser les sons* (1st ed., p. 157-213). Paris: Presses Universitaires de France.

McAdams, S., Beauchamp, J. W., & Meneguzzi, S. (1999). Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *Journal of the Acoustical Society of America*, *105*(2), 882-897.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychological Research*, *58*, 177-192.

Palmer, C. (1996). Anatomy of a performance: Sources of musical expression. *Music Perception*, *13*, 433-454.

Palmer, C. (1997). Music performance. *Annual Review of Psychology*, *48*, 115–138.

Peeters, G. (2004). *A large set of audio features for sound description (similarity and description) in the cuidado project* (Tech. Rep.). Paris: I.R.C.A.M.

Penel, A., & Drake, C. (2004). Timing variations in music performance: Musical communication, perceptual compensation, and/or motor control ? *Perception & Psychophysics*, *66*(4), 545-562.

Picinbono, B. (1997). On instantaneous amplitude and phase of signals. *IEEE Transactions on Signal Processing*, *45*(3), 552-560.

Repp, B. H. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's Träumerei. *Journal of the Acoustical Society of America*, *92*(5), 2546–2568.

Risset, J.-C. (1991). Timbre et synthèse des sons. In Christian Bourgois (Ed.), *Le timbre, métaphore pour la composition* (p. 239-260). Paris: I.R.C.A.M.

Schaeffer, P. (1966). *Traité des objets musicaux (Treaty of musical objects)*. Éditions du seuil.

Scholes, P. A. (1960). The Oxford Companion to Music. In (2nd ed.). Oxford: Oxford University Press.

Seashore, C. E. (1938). *Psychology of music*. New York: McGraw-Hill - Reprinted 1967 by Dover Publications.

Terhardt, E., Stoll, G., & Seewann, M. (1982). Pitch of complex signals according to virtual-pitch theory: Tests, examples, and predictions. *Journal of the Acoustical Society of America*, *71*(3), 671-678.

Todd, N. P. M. (1992). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America*, *91*(6), 3540-3550.

Traube, C. (2004). *An interdisciplinary study of the timbre of the classical guitar*. Unpublished doctoral dissertation, Music Technology, Department of Theory, Faculty of Music, McGill University, Montreal, Canada.

Wanderley, M. (2002). Quantitative analysis of non-obvious performer gestures. In *Gesture and Sign Language in Human-Computer Interaction: International Gesture Workshop* (p. 241). Berlin / Heidelberg: Springer.

Wessel, D. L. (1979). Timbre space as a musical control structure. *Computer Music Journal*, *3*(2), 45-52.

Wier, C. C., Jesteadtt, W., & Green, D. M. (1977). Frequency discrimination as a function of frequency and sensation level. *J. Acoust. Soc. Am.*, *6*(1), 178-184.

Zwicker, E., & Fastl, H. (1990). *Psychoacoustics: Facts and Models*. New York: Springer-Verlag.

(Insert Table 3)

(Insert Table 4)

Table 3

Results of multiple comparisons on the Bach excerpt.

Notes	F0M	F0R	ΔIOI	AT	SCM	SCR	OERM	OERR	ENVM	ENVR
N1	-	7.55**	-	3.51**	7.93**	3.90**	-	2.15*	6.73**	-
N2	-	-	-	-	-	-	-	-	2.85**	-
N3	-	-	-	-	-	-	-	-	-	-
N4	7.66**	-	-	-	-	-	-	-	-	-
N5	-	-	13.25**	6.27**	2.83**	-	-	-	-	2.65**
N6	-	-	7.23**	-	-	7.93**	5.08**	-	6.27**	2.22*
N7	-	-	-	-	-	4.91**	3.31**	-	3.52**	3.50**
N8	-	-	-	-	-	2.61**	-	2.60**	4.41**	-
N9	-	-	-	-	2.44*	-	-	-	4.49**	-
N10	-	-	-	-	2.79**	-	-	-	3.78**	-
N11	-	-	3.09**	-	3.16**	-	4.66**	2.82**	3.40**	-
N12	-	-	2.25*	-	2.41*	3.44**	2.30*	6.46**	2.52**	2.40*
N13	-	-	2.63**	-	4.52**	-	6.65**	2.23*	3.47**	-
N14	-	-	-	-	4.94**	3.37**	7.83**	2.87**	3.70**	2.19*
N15	-	-	-	4.05**	3.94**	2.80**	3.96**	3.34**	3.57**	-
N16	-	-	4.11**	-	-	-	-	-	2.42*	-
N17	-	-	2.44*	3.19**	-	-	-	-	-	-
N18	-	-	2.49*	-	-	-	-	-	-	-
N19	-	-	3.31*	2.87**	-	-	-	-	-	-
N20	-	-	-	-	-	-	-	-	1.98*	-
N21	-	-	-	2.33*	2.81**	3.30**	2.38*	-	2.36*	2.08*
N22	-	-	2.88**	-	-	2.05*	-	-	-	-
N23	-	-	5.26**	-	-	-	-	-	-	-
N24	-	2.03*	13.53**	-	-	-	3.72**	-	-	2.11*
N25	-	-	-	-	-	-	-	-	-	3.67**
N26	-	-	-	-	-	-	-	-	4.70**	-
N27	-	-	13.15**	-	-	-	2.32*	-	-	2.11*

Note. Comparisons were made on each of the 27 tones in the excerpt between the 20 mechanical performances (control group) and the 20 expressive performances. T-tests corrected for multiple comparisons (Holm-Sidak correction); $^{*}p < .05$, $^{**}p < .01$. Non-significant values are not shown.

Table 4
Results of multiple comparisons on the Mozart excerpt.

Notes	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13
ΔIOI	3.06**	-	-	-	-	-	7.40**	-	-	-	-	-	-
SCM	3.97**	-	-	-	-	-	-	-	3.05**	-	-	-	-
SCR	-	2.49*	-	2.17*	-	-	-	2.73**	-	-	-	2.14*	-
OERM	3.40**	2.57*	5.00**	5.24**	2.18*	-	2.13*	-	-	-	2.72**	3.17**	3.53**
OERR	-	-	2.23*	-	-	-	-	-	-	-	2.70**	-	-
ENVM	2.09*	2.83**	2.92**	2.53*	-	-	-	2.33*	2.02*	-	3.28**	4.30**	3.40**
ENVR	-	-	-	-	-	-	-	4.35**	-	-	-	-	-
Notes	N14	N15	N16	N17	N18	N19	N20	N21	N22	N23	N24	N25	N26
ΔIOI	-	4.76**	3.44**	-	-	-	-	-	-	-	2.43*	-	-
SCM	-	3.39**	-	-	4.04**	5.43**	-	-	-	-	-	2.78**	-
SCR	-	-	-	-	-	2.85**	2.54*	-	-	-	-	3.20**	2.86**
OERM	2.00*	-	-	2.64**	2.89**	4.73**	-	2.43*	4.13**	-	2.33*	2.99**	-
OERR	-	-	-	-	2.87**	3.04**	-	1.99*	-	-	-	-	-
ENVM	2.90**	2.12*	-	-	2.85**	4.42**	3.19**	3.44**	2.12*	-	-	-	-
ENVR	-	-	-	-	-	-	-	-	-	-	-	2.83**	-
Notes	N27	N28	N29	N30	N31	N32	N33	N34	N35	N36	N37	N38	N39
ΔIOI	-	-	-	2.27*	2.71**	4.45**	2.83**	-	-	-	-	-	-
SCM	-	-	-	6.50**	4.43**	2.74**	-	-	-	-	-	2.73**	5.71**
SCR	-	-	-	-	-	2.40*	-	-	-	-	-	-	2.52*
OERM	-	-	-	5.11**	-	-	-	-	4.18**	-	-	-	5.95**
OERR	-	-	3.74**	3.44**	-	2.28*	-	-	2.95**	-	-	-	2.34*
ENVM	-	-	-	5.04**	5.17**	2.01*	-	-	3.56**	2.71**	3.25**	3.06**	4.35**
ENVR	-	-	-	3.40**	-	3.27**	-	2.01*	-	-	-	-	-
Notes	N40	N41	N42	N43	N44	N45	N46	N47	N48	N49	N50	N51	N52
ΔIOI	-	-	-	-	-	-	-	-	-	-	-	-	-
SCM	4.45**	-	-	-	-	2.62**	-	-	-	-	-	2.61*	3.90**
SCR	3.19**	3.42**	-	2.59*	2.13*	-	2.05*	-	-	-	-	-	-
OERM	4.33**	4.94**	2.90**	3.86**	3.04**	5.89**	4.54**	2.12*	-	3.02**	-	3.12**	-
OERR	2.60**	3.28**	-	-	-	-	-	-	-	-	-	-	-
ENVM	3.08**	2.31*	2.63**	2.35*	-	-	2.17*	-	-	1.98*	2.84**	3.90**	5.09**
ENVR	-	-	-	-	-	-	2.87**	-	-	-	-	-	-
Notes	N53	N54	N55	N56	N57	N58	N59	N60	N61	N62	N63	N64	N65
ΔIOI	-	3.84**	-	-	-	-	-	-	-	-	-	-	-
SCM	-	2.85**	-	3.75**	2.18*	-	-	-	-	-	-	-	2.39*
SCR	-	-	-	2.25*	2.09*	-	2.62**	-	-	-	-	-	-
OERM	2.02*	2.34*	-	-	-	4.81**	-	2.84**	-	2.69**	2.12*	-	2.04*
OERR	-	-	-	2.34**	2.58*	-	2.21*	-	-	-	-	-	-
ENVM	3.61**	2.34*	-	3.31**	-	-	-	-	-	-	-	-	-
ENVR	-	-	-	3.41**	-	-	2.08*	-	-	-	-	-	-
Notes	N66	N67	N68	N69	N70	N71	N72	N73	N74	N75			
ΔIOI	-	-	-	-	-	-	-	2.61**	3.44**	5.22**			
SCM	-	-	-	-	2.70*	-	-	-	2.67**	4.24**			
SCR	-	-	-	-	-	-	-	-	-	-			
OERM	-	3.09**	-	2.64**	-	4.45**	2.21*	-	-	2.89**			
OERR	-	-	4.45**	-	-	-	-	-	-	-			
ENVM	-	-	-	-	-	2.95**	3.03**	-	3.00**	3.57**			
ENVR	-	-	-	-	-	-	-	-	-	-			

Note. Comparisons were made on each of the 75 tones in the excerpt between the 2 mechanical performances (control group) and the 4 expressive performances. T-tests corrected for multiple comparisons (Holm-Sidak correction); $^{*}p < .05$, $^{**}p < .01$. Non-significant values are not shown.

Analysis-by-synthesis of timbre, timing and dynamics in expressive clarinet performance

Mathieu Barthet[1], Philippe Depalle[2], Richard Kronland-Martinet, Sølvi Ystad
CNRS Laboratoire de Mécanique et d'Acoustique
31 chemin Joseph-Aiguier
13402 Marseille Cedex 20
France

In a previous study, mechanical and expressive clarinet performances of Bach's *Suite no. II* and Mozart's *Quintet for Clarinet and Strings* were analysed to determine whether some acoustical correlates of timbre (Spectral Centroid), timing (Intertone Onset Interval) and dynamics (Root Mean Square envelope) showed significant differences depending on the expressive intention of the performer. In the present companion study, we investigate the effects of these acoustical parameters on listeners' preferences. An analysis-by-synthesis approach was used to transform previously recorded clarinet performances by reducing the expressive deviations from the Spectral Centroid, the Intertone Onset Interval and the acoustical energy. Twenty skilled musicians were asked to select which version they preferred in a paired-comparison task. The results of satisitcal analyses show that the removal of the Spectral Centroid variations resulted in the greatest loss of musical preference.

The model for musical communication proposed by Kendall and Carterette (1990) in the context of traditional Western art music describes a musical performance as an act during which a performer transforms the composer's notational signals into acoustical signals which have to be subsequently decoded by a listener. Whether this act of performance can be said to constitute an (aesthetic) interpretation depends on how the performer translates the musical notations into sounds. As a matter of fact, in The Oxford Companion to Music, the notion of interpretation is defined as "the act of performance with the implication that in this act, the performer's judgment and personality necessarily have their share" (Scholes, 1960). Previous studies on musical performance have shown that performers use timing and intensity variations to play expressively (see Gabrielsson, 1999, for a review). However, less attention has been paid so far to timbre (see e.g. Juslin & Laukka, 2003, for a review), the perceptual attribute of sound which was described by Seashore (1938) as "the most important aspect of tone", and which "introduces the largest number of problems and variables". This article is the second part of a study designed to test whether timbre variations are linked to the expressive intentions of performers and to the musical preferences of listeners. Previous analyses of recorded clarinet performances helped to determine the acoustical correlates of a performer's expressive variations in timing, timbre and dynamics (Barthet, Depalle, Kronland-Martinet, & Ystad, in press). The perceptual effects of these acoustical parameters were investigated in the present study using an analysis-by-synthesis approach (Risset & Wessel, 1999).

Modeling musical interpretation

In addition to measuring musical performances, many authors have focused on modeling musical expressiveness. Detailed historical reviews of models for musical interpretation have been published by Widmer and Goebl (2004) and De Poli (2006). The strategies most commonly used for this purpose are analysis-by-measurement (see for example Todd, 1992, 1995; Windsor, Desain, Penel, & Borkent, 2006), analysis-by-synthesis and music theory knowledge (see for example Friberg, 1995), machine learning (see for example Tobudic & Widmer, 2003; Goebl, Pampalk, & Widmer, 2004), and combinations of these methods. Some of these models are predictive and account for the act of performance at its source, based on the assumption that interpretation can be described as a set of generative rules, while others are based on descriptive parameters and account for the effects of interpretation, i.e. the expressive deviations. All the above models focus mainly on timing, dynamics, articulation, and intonation, whereas timbre has often been neglected. It seems necessary to model changes of timbre in order to efficiently simulate expressive performances, especially in the case of self-sustained instruments such as the clarinet or the violin, with which the sound continues to be controlled after the onset of a note (Kergomard, 1991). In their model, Canazza, Rodá, and Orio (1999) and Canazza, De Poli, Drioli, Rodá, and Vidolin (2004) took timbre variations into account in addition to the rhythmic and dynamic

[1] Corresponding author. Electronic mail: barthet@lma.cnrs-mrs.fr
[2] Present address: Sound Processing and Control Laboratory, The Schulich School of Music, McGill University, 555 Sherbrooke Street West, Montreal (Quebec), H3A 1E3 Canada

aspects of musical performance.

Timbre, a multidimensional perceptual attribute of complex tones

Timbre is by definition the perceptual attribute that allows one to distinguish tones of equal pitch, loudness and duration (ANSI, 1960). According to Handel (1995), the identification of timbre depends on on our ability to recognize various physical factors which determine the acoustic signal produced by musical instruments (called "source" mode of timbre perception in Hajda, Kendall, Carterette, & Harshberger, 1997), as well as to analyze the acoustic properties of sound objects perceived by the ear, which has traditionally been modelled as a time-evolving frequency analyser (called "interpretative" mode of timbre perception in Hajda et al., 1997). Timbre therefore involves two complementary facets, as it relates to both the identity and the quality of sound sources. No general models for timbre have been developed so far. However, the seminal research by Grey (1977), Wessel (1979), Krumhansl (1989), Kendall and Carterette (1991) and McAdams, Winsberg, Donnadieu, De Soete, and Krimphoff (1995) developed geometrical models for timbre which represent the organization of perceptual distances (the so-called timbre space), measured on the basis of dissimilarity judgments between tones with equal pitch, loudness, and perceived durations. Two- to four-dimensional timbre spaces have often been found in dissimilarity studies between various natural or synthetic tones corresponding to orchestral instruments (Caclin, McAdams, Smith, & Winsberg, 2005). The main acoustic correlates of timbre-space dimensions are the Attack Time (correlated with the rate of energy increase in a sound, see Krimphoff, McAdams, & Winsberg, 1994) and the Spectral Centroid (the mean of the spectral energy distribution, see Grey & Gordon, 1978). The Spectral Flux (measure of the fluctuation of the spectrum over time, see McAdams et al., 1995), and the Spectral Irregularity (an index to the disparities between the harmonic components, see Krimphoff et al., 1994) are examples of other timbre descriptors which have often been proposed as correlates of timbre space dimensions.

On the role of timbre in musical performance

In Western traditional music - which is the type of music with which most studies on musical performance have dealt - the role of timbre seems to be understimated compared to those of rhythm, pitch and dynamics. This is notably revealed by the Western traditional notation system which almost omits timbre except in the references to the instrument (typological/identity timbre aspect), or in certain musical terms (e.g., *dolce*, *duramente*, *con brio*, *legato*), which can have direct or indirect consequences on morphological/quality timbre aspects (Risset, 1994). It is worth noting that some highly complex timbre notation systems including more than one hundred symbols have been developed for some traditional Chinese and Japanese instruments such as the Chin, an ancient Chinese seven-string lute (see Traube, 2004), and the Shakuhachi, a Japanese bamboo flutes. In the music played with these instruments, timbre therefore seems to be given as much importance as rhythm or pitch. It was not until the contemporary period and the new possibilities provided by digital sounds that timbre was placed at the foreground of Occidental music by composers such as Varèse, who proposed not only to compose with sounds but to compose the sounds themselves (Risset, 1994). The lack of notational information relating to timbre in traditional Western tonal music certainly does not mean that performers do not use timbre as a parameter to express feelings and emotions, however.

In a previous experiment (Barthet et al., in press), we investigated the acoustical parameters accounting for expressiveness in clarinet playing. To address this issue, mechanical and expressive performances of excerpts from the classical and baroque repertoire were recorded and analyzed. Several temporal and spectral parameters were computed to characterize the acoustical features of the performer's interpretations. Statistical analysis of the data showed significant effects of the expressive intention on the Intertone Onset Interval (IOI), quantifying the durations of the tones (see Repp, 1992), three timbre descriptors adapted to the clarinet (the Attack Time (AT), the Spectral Centroid (SC), and the Odd/Even Ratio (OER), see Barthet, 2008), and the Root Mean Square envelope (ENV), characterizing the variations of the acoustical energy. These results showed that changes in the timing descriptor (IOI), the timbre descriptors (AT, SC, OER) and the dynamics descriptor (ENV) occurred when the performer played in a more expressive way. In the present study, an analysis-by-synthesis approach was used to determine the perceptual effects of these parameters. Among the three timbre descriptors of clarinet tones (Attack Time, Spectral Centroid, Odd/Even Ratio) that were analysed in (Barthet et al., in press), the Spectral Centroid was found to be the main predictor of the performer's expressive intentions (among all descriptors, SC was the one that most frequently differentiated between mechanical and expressive performances). We therefore decided to focus on this timbre descriptor in the present experiments. For this purpose, the acoustical features charcterizing recorded clarinet performances were transformed using signal processing techniques in order to further assess the perceptual effects of these changes on the musical preferences of listeners.

Methods

General methodology

We developed a general methodology to explore the role of timbre in musical performance, based on the analysis-by-synthesis paradigm (Risset & Wessel, 1999). The method, which is summarized in Figure 1, comprises four steps: the extraction of a representation from the signal (analysis), the transformation of this representation in the frequency domain, the conversion of the modified representation back to the time domain (synthesis), and the analysis of the perceptual effects induced by the transformation.

(Insert Figure 1)

In order to identify the acoustical parameters that contribute most to aesthetic judgments, we investigated the effects of reducing the performer's original expressive deviations on musical preferences expressed by listeners in a paired-comparison task.

Stimuli

Sound corpus.

Expressive clarinet performances of the *Allemande* movement of Bach's *Suite no. II* (BWV 1008) and the *Larghetto* movement from Mozart's *Quintet for Clarinet and Strings* (KV 581) were recorded in an anechoic chamber (see Barthet et al., in press, for further details). The scores of the excerpts and the sound examples related to the study are available at: `http://www.lma.cnrs-mrs.fr/~kronland/Interpretation_perceptual`. The first musical phrases in these pieces were selected to generate the stimuli. These excerpts were chosen so that they would be sufficiently long to have a musical meaning (a musical phrase), but sufficiently short for a paired-comparison task (10- to 15 s excerpts; see (Gabrielsson & Lindstrom, 1985) for a discussion on the stimuli durations in paired comparisons).

Analysis-synthesis model.

The musical sequences were resynthesized using an additive-based synthesis model. The sound model decomposes an audio signal into a deterministic part, consisting of a sum of quasi-sinusoidal components plus a residual part. A tone $s(t)$ can thus be written as follows:

$$s(t) = \sum_{h=1}^{H} A_h(t) cos[\phi_h(t)] + r(t)$$
$$\phi_h(t) = 2\pi \int_0^t f_h(\tau)d\tau + \phi_h(0) \qquad (1)$$

where $A_h(t)$, $\phi_h(t)$, and $f_h(t)$ are the instantaneous amplitude, phase, and frequency, respectively, of the h^{th} among H sinusoids, $\phi_h(0)$ is the initial phase, and $r(t)$ is the residual. This method is particularly suitable for changing the characteristics of the tones related to timbre and timing, for example, since sounds are reconstructed as a superposition of partial components, the frequency and amplitude of which can be individually controlled.

In addition to their harmonic structure, clarinet tones contain ancillary noises (for instance breath and key noises) which also contribute to the identity of the instrument. The latter noises are partly contained in the residual. The residual was obtained by performing a time-domain subtraction between the original sequence and the resynthesized deterministic part with no transformations. Resynthesized sequences were then obtained by juxtaposing the resynthesized tones. This procedure gives a high-quality resynthesis of the clarinet performances (cf. Sound Examples *1* to *3*).

Transformations.

Recordings of a players' original performances can be transformed by appropriately manipulating the control parameters in the model. In order to assess the perceptual effects of Spectral Centroid (SC) variations, Intertone Onset Interval (IOI) deviations, and acoustical energy (ENV) variations, three transformations were carried out to independently modify these parameters.

Spectral Centroid freezing (T_T) A transformation acting on the Spectral Centroid was designed to control the shape of the tones' Spectral Centroids without affecting their acoustical energy. The method used for this purpose was based on the transformation described by McAdams, Beauchamp, and Meneguzzi (1999), which consists in eliminating the Spectral Flux (measure of the fluctuation of the spectrum over time) while leaving the RMS envelope intact. As a matter of fact, the removal of the Spectral Flux corresponds to canceling the time-dependent variations in the spectral envelope's shape, thus cancelling the Spectral Centroid variations. As the drastic elimination of all Spectral Centroid variations may cause tones to sound too artificial, the microfluctuations in the instantaneous amplitudes were preserved by separating the low-frequency variations in the instantaneous amplitudes $L_h(t)$, defined as the amplitude variations of frequency below 10 Hz, from the high-frequency variations $H_h(t)$, defined as the amplitude variations of frequency above 10 Hz. The modified instantaneous amplitude $\widetilde{A_h}(t)$ of the h^{th} component of a given tone is obtained as follows:

$$\widetilde{A_h}(t) = \beta_h(t) ENV(t)$$
$$\beta_h(t) = \frac{\overline{A_h} + H_h(t)}{\sqrt{\sum_{h=1}^{H} [\overline{A_h} + H_h(t)]^2}} \qquad (2)$$

where $ENV(t)$ is the RMS envelope of the tone, and $\beta_h(t)$ is a term computed to fluctuate around the time-averaged amplitude $\overline{A_h}$ of the h^{th} harmonic. $\overline{A_h}$ was determined during the sustained part of the tone. Since the high-frequency fluctuations $H_h(t)$ are small in comparison with $\overline{A_h}$, the term $\beta_h(t)$ is almost constant with time. The shape of the tones' new instantaneous amplitudes $\widetilde{A_h}(t)$ is therefore very similar to that of the RMS envelope of the tone $ENV(t)$ with various scale factors. Figure 2 shows the instantaneous amplitudes of a clarinet tone before and after the transformation. The modified Spectral Centroid is almost frozen over time although it varies quickly around the mean value of the initial Spectral Centroid calculated during the sustained part of the tone.

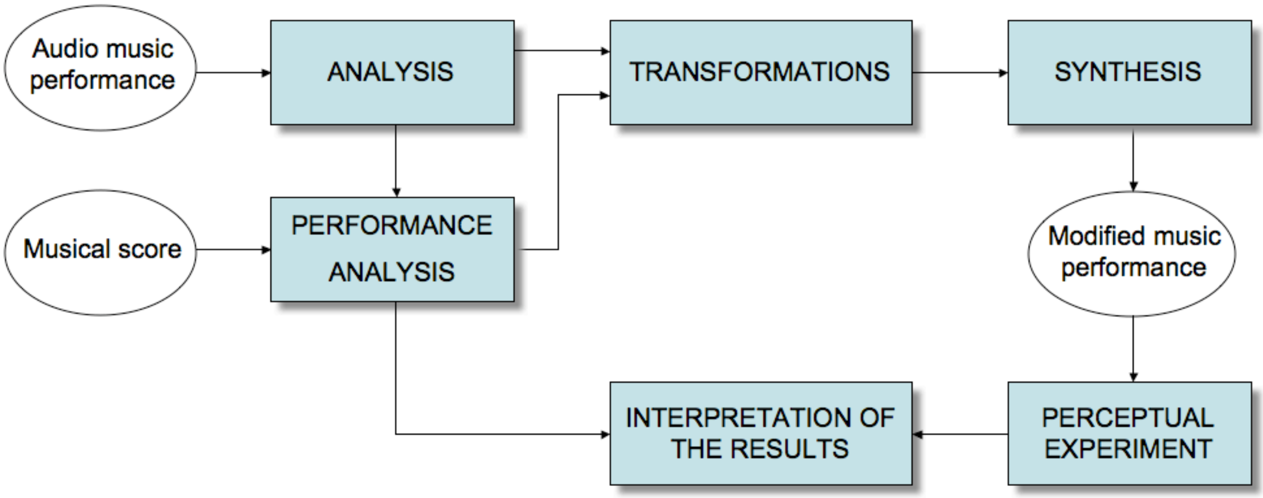


Figure 1. Exploration of the acoustical correlates of musical expressiveness based on the analysis-by-synthesis approach.

(Insert Figure 2)

The application of this transformation to one of the clarinet performances is given, as example, in Figure 3. The original Spectral Centroid variations are presented in Figure 3(a), and the modified ones in Figure 3(b).

Intertone Onset Interval deviation cancellation (T_R) We designed a transformation for replacing the effective Intertone Onset Intervals (as played by the performer) with the nominal IOIs given by the transcription of the score notations. This was done by applying time-scale modifications to the instantaneous frequencies and amplitudes of the tone's components. These changes were applied only to the sustained and release portions of the tones, sparing the original attack, which is known to be an important attribute of timbre. The time dilation/contraction coefficient α was computed for each tone as follows:

$$\alpha = \frac{IOI_{nom} - AT}{IOI_{eff} - AT} \tag{3}$$

where IOI_{nom} and IOI_{eff} denote the tone's nominal and effective IOIs, and AT is the attack time. For each tone, after the attack, the instantaneous frequencies $\widetilde{f_h}(t)$ and amplitudes $\widetilde{A_h}(t)$ were transformed as follows:

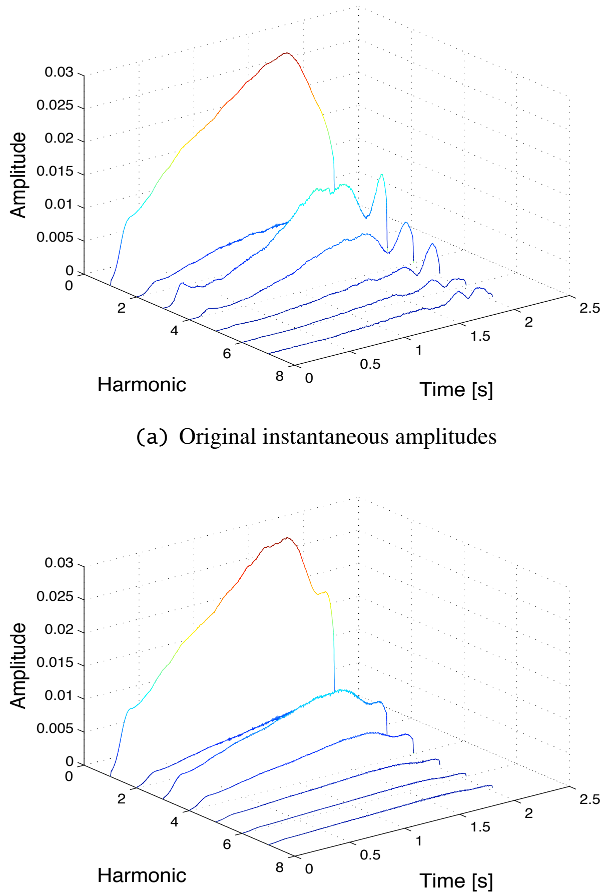
$$\begin{aligned} \widetilde{f_h}(t) &= f_h(\frac{t}{\alpha}) \\ \widetilde{A_h}(t) &= A_h(\frac{t}{\alpha}) \end{aligned} \tag{4}$$

Figures 3(c) and 3(d) show the effect of the transformation on the IOI deviation descriptor ΔIOI. This transformation yields "mechanical" performances which are exactly in line with the timing indications given on the score.

Compression of the dynamics (T_D) In order to reduce the variations in the acoustical energy, we used a dynamic range controller serving as a compressor and limiter (see Zölzer, 1997). The signal's input level was determined via an envelope follower based on peak measurements. A gain factor was then used to adjust the amplitude of the input signals. This compressor/limiter was controlled by the parameters generally used with devices of this kind, i.e. by the thresholds and slope of the limiter, the thresholds and slope of the compressor, and the attack and release times. Although this method is based on nonlinear processing procedures liable to cause harmonic distortions, the control parameters of the dynamic range controller were carefully selected to prevent the occurrence of any audible changes in the timbre. The attack and release times were both set at 10 *ms*. As compression and limiting procedures lead to changes in the signal levels, the loudness of each modified sequence was equalized to a constant value.

Examples of original and transformed RMS envelopes are given in Figures 3(e) and 3(f), respectively. As was to be expected, the range of variation of the acoustical energy is much smaller in the modified version.

To avoid effects of pitch on the preference of the listeners, for each transformation, the fundamental frequencies of the tones were set at their mean values computed in the sustained parts. The instantaneous frequencies of the h tones' components were therefore set at $h\overline{f_0}$, where $\overline{f_0}$ denotes the mean fundamental frequency of the tone. We checked that the frequency changes were only weakly perceptible in the musical excerpts selected (cf. Sound Example *4*).



(a) Original instantaneous amplitudes

(b) Transformed instantaneous amplitudes

Figure 2. Transformations of the instantaneous amplitudes (harmonics 1 to 7): (a) original clarinet tone - (b) after the Spectral Centroid freezing.

Table 1
Description of the stimuli.

Stimuli	*Transformation description*
M_0	No transformation
M_T	Freezing of the Spectral Centroid (T_T)
M_R	Canceling of the IOI deviations (T_R)
M_D	Compression of the dynamics (T_D)
M_{TR}	Combination of T_T and T_R
M_{TD}	Combination of T_T and T_D
M_{RD}	Combination of T_R and T_D
M_{TRD}	Combination of T_T, T_R and T_D

(Insert Figure 3)

Design of the stimuli.
(Insert Table 1)

The three basic transformations (T_T, T_R, T_D) and their four combinations (T_{TR}, T_{TD}, T_{RD}, T_{TRD}) were applied to the expressive clarinet performances. As the original performances were recorded in an anechoic chamber, a slight reverberation was added to the resynthesized versions to make them sound more natural. The 8 stimuli listed in Table 1 were therefore generated.

Participants

Given the relatively demanding requirements of the auditory discrimination task, in which the participants had to rate the interpretations with various levels of expression, the experiment was carried out with skilled musicians (14 males, 6 females; age=19-50 years) as listeners. Most of the participants were students in musicology practising various musical instruments (clarinet, guitar, piano, violin, etc.), who were participating in improvisation workshops at the GRIM (Groupe de Recherche et d'Improvisations Musicales, Marseille).

Apparatus

The experiment was carried out in an audiometric cabin. The user interface was implemented in the *Matlab* environment[1]. The sound files were stored on the hard drive of an *Apple iMac G4* computer and delivered to the participants via a *STAX SRM-310* headphone system.

Procedure

Participants were asked to select which stimuli they preferred in a paired comparisons task. They first underwent a training phase in order to become familiar with the task. In order to assess the influence of the musical excerpt which was used, each participant attended two sessions, one with the stimuli of the Bach sequences, and the other with the stimuli of the Mozart sequences. The order of the two sessions (denoted 'Bach' and 'Mozart' in the following) was counterbalanced across participants. At the end of the test, the participants were asked to complete a questionnaire specifying what criteria they had used to assess the recordings.

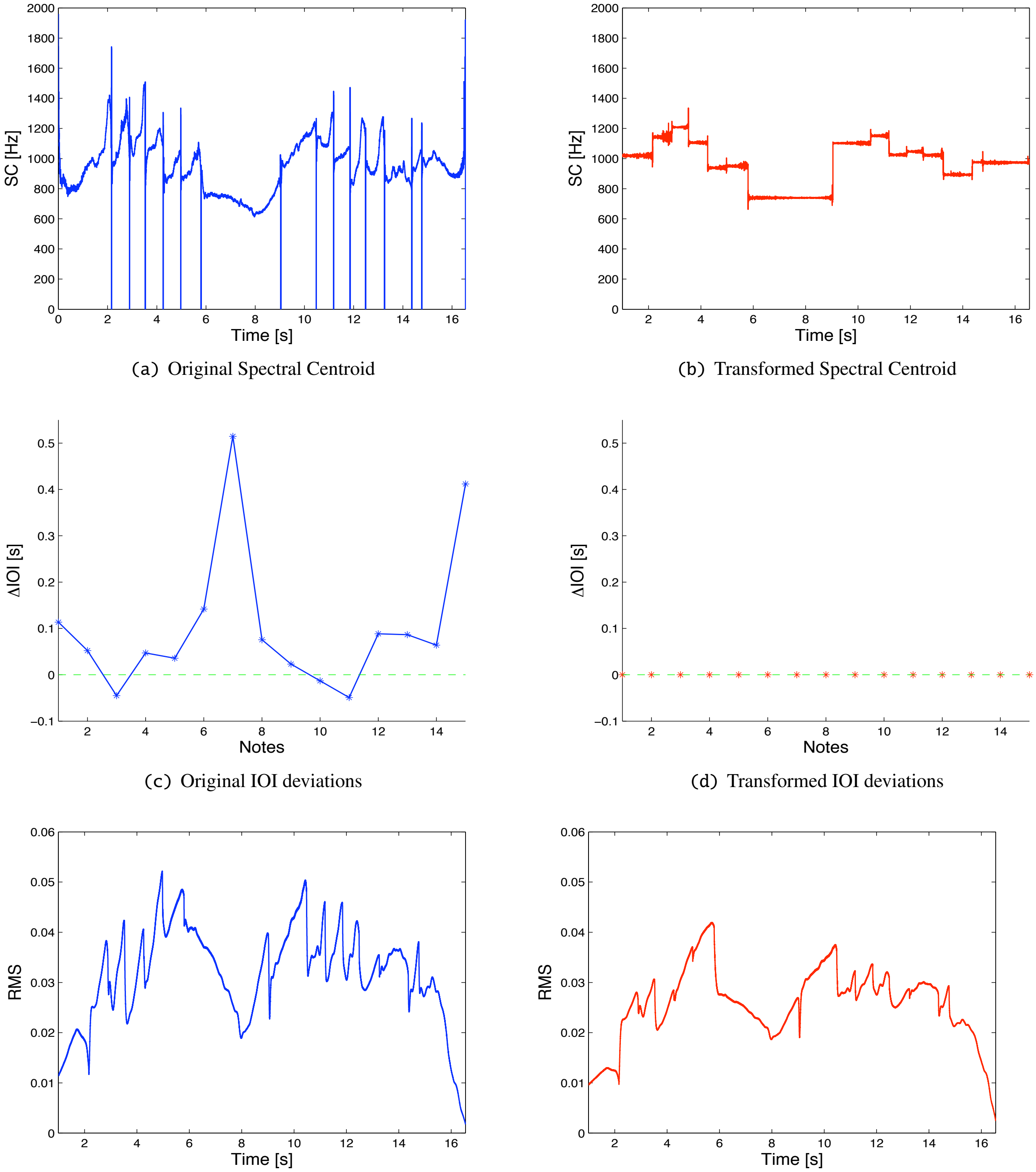
During the experiment, participants listened to several successive pairs of clarinet performances separated by a 1-s interval. They could listen to each performance as many times as they wished. At each trial, participants were asked to indicate which version they preferred. All the possible combinations (28) of the 8 stimuli (the 3 basic transformations, their 4 combinations and the original performance) were presented to each participant. The within-pair order and the order of the pairs were randomized. Each session lasted approximately 20 minutes.

Results and discussion

Presentation of the perceptual data

The responses of each participant m are presented in the form of a preference matrix P_m. The elements of P_m

[1] `http://www.mathworks.com/products/matlab/`



(a) Original Spectral Centroid

(b) Transformed Spectral Centroid

(c) Original IOI deviations

(d) Transformed IOI deviations

(e) Original RMS envelope

(f) Transformed RMS envelope

Figure 3. Original (left) and transformed (right) expressive patterns for a clarinet performance of the Mozart excerpt.

designate whether stimulus i was preferred to stimulus j. The global preference matrix P of the sample is defined as the sum of the individual preference matrices P_m.

With each participant, the various performances were given preference scores $S_m(i)$, depending on the number of times they were preferred to the others (with each performance i, this corresponds to summing the preference matrix elements $P_m(i, j)$ across the columns). The scores range from 0 to 7 (times preferred). The mean preference scores, denoted $S(i)$, were computed on the basis of the preference scores $S_m(i)$, associated to the ratings of the participants.

An alpha level of .05 was used for all statistical tests.

Sample homogeneity

The degree of agreement among the participants was computed using the Kendall coefficient of agreement u for paired comparisons, as described by Siegel and John Castellan (1988). This nonparametric measure of association can be written as follows:

$$u = \frac{2\sum_{i=1}^{N_s}\sum_{j=1}^{N_s} C_2^{P(i,j)}}{C_2^{N_m} C_2^{N_s}} - 1 \tag{5}$$

where C_n^k denotes the binomial coefficient, N_s is the total number of stimuli, and N_m is the total number of participants. Here, $N_s = 8$ and $N_m = 20$. When N_m is even, u can range from $\frac{-1}{N_m - 1}$ to 1, which means that there is complete agreement among the participants. Siegel and John Castellan (1988) defined an index W_T based on u, which can range from 0 to 1. The statistic u can be taken to be an estimate for a population parameter ν, which stands for the true degree of agreement in the population. We tested the null hypothesis ($H_0 : \nu = 0$) that there was no agreement among the participants against the alternative ($H_1 : \nu \neq 0$) that the degree of agreement was greater than what one would have expected had the paired comparisons been done at random. As the total number of participants was large ($N_m > 6$), we used a large-sample approximation of the sampling distribution, asymptotically distributed as a χ^2 distribution with 28 degrees of freedom (*df*). The values of u and W_T calculated with both the Bach and Mozart excerpts can be found in Table 2. These results show that the agreement among the participants' preferences was significantly higher than chance with both the Bach ($u = 0.58$, $p < .001$), and Mozart ($u = 0.52$, $p < .001$) sequences.

(Insert Table 2)

Analyses of variance (ANOVA)

In order to assess the influences of the musical excerpts (2 modalities) and the transformations (8 modalities) on the participants' preferences, the preference scores $S_m(i)$ were subjected to a two-way repeated measures analysis of variance (ANOVA), since each participant attended both the 'Bach' and 'Mozart' sessions. The results of the ANOVA, which are presented in Table 3, show that the differences of preference scores between the two excerpts were not sufficiently large to exclude the possibility that they might be due to chance. Indeed, the effect of the musical excerpt on the preference scores was not found to be significant, $F(1, 133) = 0$, $p = 1.00$ (the mean of the preference scores were identical for both excerpts). The interaction between the musical excerpt and the transformations was also non significant, $F(7, 133) = 1.05$, $p = 0.40$. Conversely, the effect of the transformations on the preference scores was highly significant, $F(7, 133) = 118.99$, $p < .001$. In order to compare the influences of the various transformations, multiple comparison tests were conducted for each excerpt (Tukey HSD, Honestly Significant Difference tests). This procedure determined the significant differences existing between the mean preference scores in each 2 by 2 combination between the various transformations. The results of the multiple comparison procedure are presented in Tables 4 and 5. The preference scores associated with the various performances are described by the box-and-whisker diagrams shown in Figure 4.

(Insert Table 4)

(Insert Table 5)

(Insert Figure 4)

For both the Bach and Mozart excerpts, the version M_0 was the most frequently preferred rendering (see Figure 4). This is not surprising, since the expressive deviations associated with timbre, timing and dynamics had not been removed from this version. It is also not surprising that the performance M_{TRD} to which the 3 basic transformations

Table 2
Coefficients of agreement among the participants at the Bach and Mozart sessions.

Session	*df*	*u*	W_t	X^2
Bach	28	0.58 ($p < .001$)	0.60	338.20
Mozart	28	0.52 ($p < .001$)	0.54	303.80

Table 3
Results of the two-way repeated measures analysis of variance of the preference scores.

Source	*df*	*F*	*p*
Excerpt	1	0	1.00
Transformation	7	118.99***	<.001
Excerpt x Transformation	7	1.05	.40
Error	133	(1.11)	

Note. The result enclosed in parentheses represent the mean square error. $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Table 4

Results of the multiple comparison procedure for the various transformations within the 'Bach' session.

	M_0	M_T	M_R	M_D	M_{TR}	M_{TD}	M_{RD}	M_{TRD}
M_0	-	13.21^{***}	3.60 (p=0.18)	6.61^{***}	15.82^{***}	22.62^{***}	8.81^{***}	22.22^{***}
M_T		-	9.61^{***}	6.61^{***}	2.60 (p=0.59)	9.41^{***}	4.41^{*}	9.01^{***}
M_R			-	3.00 (p=0.40)	12.21^{***}	19.02^{***}	5.21^{**}	18.62^{***}
M_D				-	9.21^{***}	16.02^{***}	2.20 (p=0.78)	15.62^{***}
M_{TR}					-	6.81^{***}	7.01^{***}	6.41^{***}
M_{TD}						-	13.82^{***}	0.40 (p=1.00)
M_{RD}							-	13.41^{***}

The table presents the results of the pairwise multiple comparison procedure for the 8 transformations (Tukey HSD tests). The values of the studentized range statistic q are reported; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 5

Results of the multiple comparison procedure for the various transformations within the 'Mozart' session.

	M_0	M_T	M_R	M_D	M_{TR}	M_{TD}	M_{RD}	M_{TRD}
M_0	-	14.62^{***}	1.40 (p=0.98)	6.81^{***}	13.61^{***}	21.02^{***}	7.21^{***}	20.02^{***}
M_T		-	13.21^{***}	7.81^{***}	1.00 (p=1.00)	6.41^{***}	5.37^{**}	5.41^{**}
M_R			-	5.41^{**}	12.21^{***}	19.62^{***}	6.01^{***}	18.62^{***}
M_D				-	6.81^{***}	14.22^{***}	0.60 (p=1.00)	13.21^{***}
M_{TR}					-	7.41^{***}	6.21^{***}	6.41^{***}
M_{TD}						-	13.61^{***}	1.00 (p=1.00)
M_{RD}							-	12.61^{***}

Legend: see Table 4.

were applied was the least preferred on average.

The removal of the IOI deviations (T_R) was the transformation which resulted on average in the least loss of musical preference in the case of both excerpts (see Figure 4): as shown in Tables 4 and 5, there were no significant differences between M_R and M_0. In the case of the Bach excerpt, this finding is probably due to the style of the musical piece, which is an instrumental dance. Listeners might therefore expect the excerpt to be played with smaller IOI deviations from nominal durations (i.e., in a "mechanical" way). In the case of the Mozart excerpt, the fact that the removal of the IOI deviations had little effect on the listeners' preferences is more surprising, as timing deviations are more likely to occur in slow tempo movements (*Larghetto* in this case). However, it is worth noting that the variance of the preference scores attributed to M_R was greater with the Mozart than with the Bach excerpt.

After the removal of the IOI deviations, the compression of the dynamics was the transformation which had the least effect on the musical preference. However, the differences between M_D and the reference M_0 were significant with both excerpts (see Tables 4 and 5).

At both sessions, the performances which were processed by freezing the Spectral Centroid (M_T, M_{TD}, M_{TR}, M_{TRD}) were consistently the least preferred (see Figure 4). As shown in Tables 4 and 5, these versions showed the most significant differences with the reference version M_0. The Spectral Centroid freezing procedure therefore resulted in a greater loss of musical preference than the removal of the IOI deviations, or the compression of the dynamics. It is worth noting that removal of the Spectral Centroid variations had more degrading effects than the removal of the IOI deviation and the dynamic transformations combined (cf. score of M_T versus the one of M_{RD}).

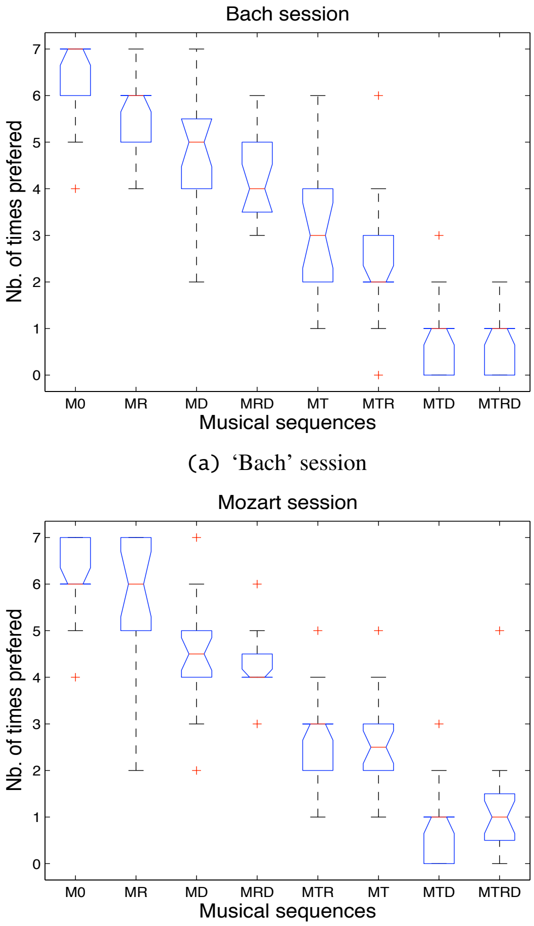
Hierarchical Cluster Analysis (HCA)

As the nature of the stimuli was of the categorical type (the stimuli were either subjected to a transformation or not), we also performed a Hierarchical Cluster Analysis (HCA). Between-transformations distances were obtained by computing the Euclidean distances between the mean preference scores $S(i)$. Two different hierarchical clustering methods were applied to the between-transformations distances, the complete (furthest distance) and the ward (inner squared distance) linkages. Both methods returned similar hierachical cluster trees (dendrograms). As shown in Figure 5, which presents the dendrograms obtained with the complete linkage method, the two main clusters at the 'Bach' and Mozart' sessions were identical. One of them contains all the performances which underwent the Spectral Centroid freezing transformation (M_T, M_{TD}, M_{TR}, M_{TRD}), and the other contains the remaining performances (M_0, M_R, M_D, M_{RD}). These results confirm that the Spectral Centroid transformation had stronger perceptual effects than the other transformations.

(Insert Figure 5)

General discussion

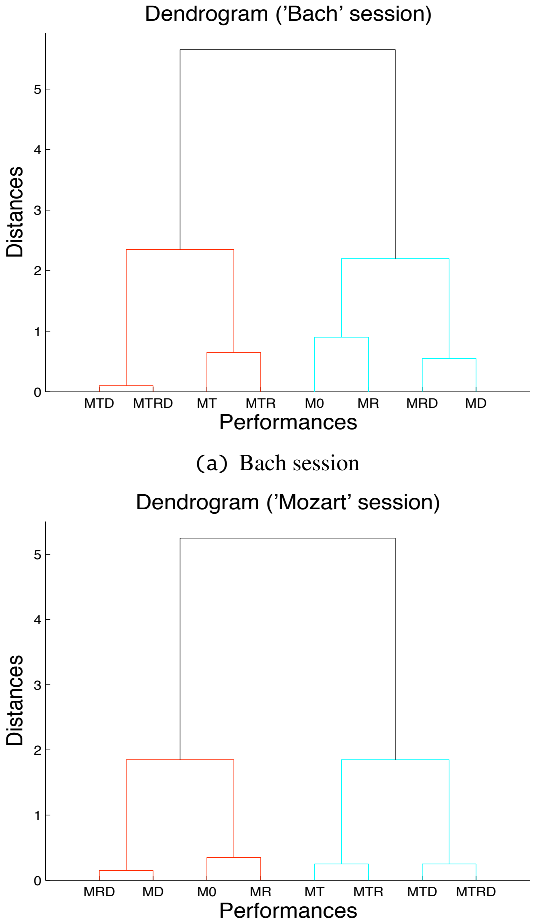
The results show that the preferences of the participants depended on which acoustical parameters had been modified



(a) ‘Bach’ session

(b) ‘Mozart’ session

Figure 4. Box and whisker plots of the listeners’ preference scores at the ‘Bach’ session (a) and the ‘Mozart’ session (b). The various sequences are arranged from left to right in decreasing order, based on the median values of the scores. The lines in the bow are at the lower quartile, median, and upper quartile values. The whiskers show the range of the rest of the data. Their lengths was 1.5 times the interquartile range. Outliers are represented by crosses. The notches on the box give a robust estimate of the uncertainty of the medians. Note that the number of times each sequence was preferred ranged necessarily between 0 and 7.



(a) Bach session

(b) Mozart session

Figure 5. Dendrogram representations of the Hierarchical Cluster Analysis of the between-transformations distances (complete linkage method), for the excerpts from Bach (a) and Mozart (b).

(Spectral Centroid and/or Intertone Onset Interval and/or RMS envelope). The score of a sequence which had undergone a transformation with multiple modifications (e.g., M_{RD}) was lower or at best equal to the lowest of the scores obtained by the sequences which had undergone the basic transformations (e.g., M_D).

No significant effect of the musical excerpt on the preferences was observed. Among the seven transformations, the greatest effects observed were those caused by the freezing of the Spectral Centroid. This transformation results in a greater loss of preference than that caused by removal of the IOI deviations or dynamic compression or these two transformations combined. McAdams et al. (1999) investigated subjects’ ability to discriminate between various isolated instrumental tones in which spectro-temporal simplifications had been made. Among these simplifications, the freezing of the Spectral Centroid (induced by the Spectral Flux freezing process) was found to be most easily discriminated by the listeners. This effect was less marked, however, with clarinet tones than with brass instrument (trumpet) tones, since the spectra of the latter instruments generally undergo much less Spectral Flux than other instruments (Grey, 1977). The strong influence of the removal of Spectral Centroid variations may be due to the fact that this descriptor seems to be correlated to one of the main timbre dimension (Grey, 1977; Krumhansl, 1989; McAdams et al., 1995). Although analysis/synthesis techniques allow one to control the Spectral Centroid independently from other sound dimensions (such as the acoustical energy, in particular), this process is not necessarily acceptable regarding the timbral identity of instrumental tones. However, the participants never preferred the versions in which both the Spectral Centroid and the acoustical energy variations had been transformed. The Spectral Centroid freezing probably alters the original timbre of the instrument so that it becomes unnatural, or at least different. In this case, the transformation would do more than simply remove the expressive deviations of timbre, since it would change the nature of the instrument. It should be noted that the participants did not report that they had relied on timbre (identity) changes. Most of them showed a preference for performances with "lively" rather than "static" tones, which refers directly to the presence or absence of variations in the intensity and/or timbre during the tones. To check the quality of our transformations, we played the timbre-modified sequences to the clarinetist whose performances we had recorded. He perceived that they were different from his original performances but did not think they no longer sounded like a clarinet and mentioned possible causal explanations for the differences. He suggested, for instance, that the instrument might be poorly controlled by the player (e.g. students who move their mouthpiece) or that it might be in poor condition (e.g. an old reed), resulting in a general lack of homogeneity in the interpretation. Abeles (1973) developed a clarinet performance adjudication scale and observed that the timbre of the tones was one of the most important factors used by music teachers to rate clarinet performances. It is therefore not very surprising that the present participants based their assessments mainly on the timbre of the clarinet tones.

In these experiments, the intricate process of interpretation was reduced to a simple linear, additive model (SC variations +/- IOI deviations +/- Energy variations). This is a first step towards reaching a better understanding of the influence of the temporal and spectral parameters related to timbre, timing and intensity. However, these parameters may interact at the level of both performers (in the sound production process) and listeners (in the decoding of the musical signals). For instance, when listening to his own performance of the excerpt from Mozart's *Larghetto*, the clarinetist was convinced that he had deliberately lengthened one of the tones in the sequence, causing the following one to be late. However, the analysis showed that the onset of the second tone was perfectly on the beat. The decrescendo in the first of these two tones, which was associated with a concomitant decrease in brightness, may have induced this feeling of ritardando.

Summary and conclusions

This study focuses on the perceptual effects of variations in acoustical correlates of timbre, timing and intensity on musical preference. To address this issue, an experimental method based on the analysis-by-synthesis approach was developed, which consisted in transforming the expressive content of recorded clarinet performances and assessing the effects of these changes on listeners' musical preferences. Seven transformations were designed to remove or compress the variations in the Spectral Centroid (a timbre parameter), Intertone Onset Interval (a timing parameter), and acoustical energy, either separately or in various combinations. The statistical analyses carried out on the listeners' aesthetic judgments showed that the Spectral Centroid freezing transformation most significantly decreased the musical preference of the performances. This finding seems to be due to the fact that this transformation altered the original timbre of the clarinet tones (the identity of the instrument), as well as drastically affecting the time-evolving spectral shapes, causing the tones to be static and unlively (the quality of the sound).

These results confirm that the performer's choices of timbre, timing and intensity variables (see the companion article by Barthet et al., in press) affect the listener's perception of the musical preference. The variations in the Spectral Centroid during tone production seem to be an important feature of preference in the musical message transmitted from performers to listeners. Indeed, in another study, we established that controlling the time-evolving Spectral Centroid of tones improved the preference of sampler-based generated sequences (Barthet, Kronland-Martinet, & Ystad, 2008).

These findings suggest that it might be worth developing a general set of rules relating to timbre, as previous authors have done in the case of temporal and intensity deviations (see e.g. Mathews, Friberg, Bennett, Sapp, & Sundberg, 2003; Widmer & Goebl, 2004; De Poli, 2006). By taking the variations in the acoustical parameters associated with timbre into account in computational models for music performance, it might then be possible to improve the automatic rendering of musical pieces by computers.

Acknowledgments

We would like to thank the clarinetist Claude Crousier for participating in this project and many fruitful discussions. We are grateful to Mitsuko Aramaki and Henri Burnet from the Institut des Neurosciences Cognitives de la Méditerranée and Jean Pierre Durbec from the Centre d'Océanologie de Marseille for their precious help. We would also like to thank the reviewers for their helpful comments and advices. This project was partly supported by the French National Research Agency[2].

References

Abeles, H. F. (1973). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education*, *21*(3), 246-255.

ANSI. (1960). USA Standard Acoustical Terminology. New York: American National Standards Institute.

Barthet, M. (2008). *De l'interprète à l'auditeur: une analyse acoustique et perceptive du timbre musical.* Unpublished doctoral dissertation, Université Aix-Marseille II.

[2] ANR JC05-41996, "senSons" `http://www.sensons.cnrs-mrs.fr/`

Barthet, M., Depalle, P., Kronland-Martinet, R., & Ystad, S. (in press). Acoustical correlates of timbre and expressiveness in clarinet performance. *Music Perception*.

Barthet, M., Kronland-Martinet, R., & Ystad, S. (2008). Improving musical expressiveness by time-varying brightness shaping. In R. Kronland-Martinet, S. Ystad, & K. Jensen (Eds.), *Sense of Sounds* (p. 313-336). Berlin / Heidelberg: Springer-Verlag.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *J. Acoust. Soc. Am.*, *118*(1), 471-482.

Canazza, S., De Poli, G., Drioli, C., Rodá, A., & Vidolin, A. (2004). Modeling and control of expressiveness in music performance. *Proc. of the IEEE*, *92*(4), 686-701.

Canazza, S., Rodá, A., & Orio, N. (1999). A parametric model of expressiveness in musical performance based on perceptual and acoustical analyses. In *Proc. Int. Comp. Music Conf.* (p. 379-382). Beijing, China.

De Poli, G. (2006). Algorithms for Sound and Music Computing. In (chap. Expressiveness in music performance). Creative Commons.

Friberg, A. (1995). *A quantative rule system for musical performance*. Unpublished doctoral dissertation, Department of Speech, Music and Hearing, Royal Institute of Technology, Stockholm.

Gabrielsson, A. (1999). The performance of music. In *Psychology of music* (2nd ed.). Academic Press.

Gabrielsson, A., & Lindstrom, B. (1985). Perceived sound quality of high-fidelity loudspeakers. *J. Audio Eng. Soc.*, *33*(1), 33-53.

Goebl, W., Pampalk, E., & Widmer, G. (2004). Exploring expressive performance trajectories: six famous pianists play six Chopin pieces. In S. D. Lipscomb, R. Ashley, & P. Gjerdingen R. O. & Webster (Eds.), *Journal of Research in Music Education* (p. 505-509). Evanston, IL, USA: Adelaide, Australia: Causal Productions.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am.*, *61*, 1270-1277.

Grey, J. M., & Gordon, J. W. (1978). Perception of spectral modifications on orchestral instrument tones. *Computer Music Journal*, *11*(1), 24-31.

Hajda, J. M., Kendall, R. A., Carterette, E. C., & Harshberger, M. L. (1997). Methodological issues in timbre research. In I. Deliége & J. Sloboda (Eds.), *Perception and Cognition of Music* (2nd ed., p. 253-306). New York: Psychology Press.

Handel, S. (1995). Timbre perception and auditory object identification. In B. C. J. Moore (Ed.), *Handbook of perception and cognition* (2nd ed., p. 425-461). San Diego, California: Academic Press.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770-814.

Kendall, R. A., & Carterette, E. C. (1990). The communication of musical expression. *Music Perception*, *8*(2), 129–164.

Kendall, R. A., & Carterette, E. C. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Perception*, *8*, 369-404.

Kergomard, J. (1991). Le timbre des instruments à anche. In Christian Bourgois (Ed.), *Le timbre, métaphore pour la composition* (p. 224-235). Paris: I.R.C.A.M.

Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caractérisation du timbre des sons complexes, II Analyses acoustiques et quantification psychophysique (Characterization of complex sounds' timbre, II Acoustical analyses and psychophysical quantification). *Journal de Physique IV, Colloque C5*, *4*, 625-628.

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand ? In S. Nielzén & O. Olsson (Eds.), *Proc. of the Marcus Wallenberg Symposium held in Lund, Sweden* (p. 43-53). Amsterdam: Excerpta Medica.

Mathews, M., Friberg, A., Bennett, G., Sapp, C., & Sundberg, J. (2003). A marriage of the Director Musices program and the conductor program. In *Proc. of the Stockholm Music Acoustics Conference (SMAC 03).* Stockholm, Sweden.

McAdams, S., Beauchamp, J. W., & Meneguzzi, S. (1999). Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *Journal of the Acoustical Society of America*, *105*(2), 882-897.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychological Research*, *58*, 177-192.

Repp, B. H. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's Träumerei. *Journal of the Acoustical Society of America*, *92*(5), 2546–2568.

Risset, J.-C. (1994). Quelques aspects du timbre dans la musique contemporaine (A few aspects of timbre in contemporary music). In A. Zenatti (Ed.), *Psychologie de la musique (Psychology of music)* (1st ed., pp. 87–114). Paris: Presses Universitaires de France.

Risset, J.-C., & Wessel, D. L. (1999). Exploration of timbre by analysis and synthesis. In D. Deutsch (Ed.), *Psychology of music* (2nd ed.). Academic Press.

Scholes, P. A. (1960). The Oxford Companion to Music. In (2nd ed.). Oxford: Oxford University Press.

Seashore, C. E. (1938). *Psychology of music*. New York: McGraw-Hill - Reprinted 1967 by Dover Publications.

Siegel, S., & John Castellan, N., Jr. (1988). Non parametric statistics for the behavioral sciences. In (2nd ed., p. 272). McGraw-Hill International Editions.

Tobudic, A., & Widmer, G. (2003). *Playing Mozart phrase by phrase* (Tech. Rep.). ÖFAI-TR-2003-02.

Todd, N. P. M. (1992). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America*, *91*(6), 3540-3550.

Todd, N. P. M. (1995). The kinematics of musical expression. *Journal of the Acoustical Society of America*, *97*(3), 1940-1949.

Traube, C. (2004). *An interdisciplinary study of the timbre of the classical guitar*. Unpublished doctoral dissertation, Music Technology, Department of Theory, Faculty of Music, McGill University, Montreal, Canada.

Wessel, D. L. (1979). Timbre space as a musical control structure. *Computer Music Journal*, *3*(2), 45-52.

Widmer, G., & Goebl, W. (2004). Computational models of expressive music performance. *Journal of New Music Research*, *33*(3), 203–216.

Windsor, W. L., Desain, P., Penel, A., & Borkent, M. (2006). A structurally guided method for the decomposition of expression in music performance. *Journal of the Acoustical Society of America*, *119*(2), 1182-1193.

Zölzer, U. (1997). Dynamic range control. In *Digital Audio Signal Processing.* John Wiley & Sons.

Acta Acustica united with Acustica
Vol. 96 (2010) 678 – 689

DOI 10.3813/AAA.918322

From Clarinet Control to Timbre Perception

Mathieu Barthet, Philippe Guillemain, Richard Kronland-Martinet, Sølvi Ystad
CNRS Laboratoire de Mécanique et d'Acoustique, 31 chemin Joseph-Aiguier, 13402 Marseille Cedex 20, France.
barthet@lma.cnrs-mrs.fr

Summary

This study investigates the relationships between the control gestures of the clarinet, the generated timbres and their perceptual representation. The understanding of such relationships can provide great interest in several research contexts: synthesis and control (e.g., to improve the quality of current synthesis models), music analysis and perception (e.g., to study music performance), and music information retrieval (e.g., to find relevant acoustical descriptors for automatic instrument and/or performer identification). A physics-based model was used to generate synthetic clarinet tones by varying the main control parameters of the model (related to the blowing pressure and lip pressure on the reed). 16 participants had to rate the dissimilarities between pairs of different tones and describe the factors on which they based their judgments in a questionnaire. The collected data were subjected to various statistical analyses (multidimensional scaling and hierarchical clustering) in order to obtain a low-dimensional spatial configuration (timbre space) which best represents the dissimilarity ratings. The structure of the clarinet timbre space was interpreted both in terms of control parameters and acoustical descriptors. The analyses revealed a 3-dimensional timbre space, whose dimensions were well correlated to the Attack Time, the Spectral Centroid, and the Odd/Even Ratio. Comparisons of natural and synthetic clarinet tones showed that the Odd/Even Ratio appears to be a good predictor of the beating reed situation, specific to single-reed instruments.

PACS no. 43.66.Jh, 43.75.Pq, 43.75.Yy, 43.75.Zz

1. Introduction

Since its conception in the 17^{th} century, the clarinet, a single-reed instrument from the woodwind family, has aroused the interest of composers as a result of its unique timbre [1]. From the acoustical point of view, the clarinet can be described as the association of an exciter (the reed) and a resonator (the air column contained in the bore of the instrument). Under certain control conditions imposed by the performer, the coupling between the exciter and the resonator, can lead to the establishment of self-sustained oscillations which are responsible for the sounds generated by the instrument. By varying their control gestures, performers can produce different timbres (see e.g., [2]). This study aims to increase understanding of the relationships between the control of the instrument, the generated sounds, and how they are perceived. The understanding of such relationships can provide great interest in several research contexts: synthesis and control (e.g., to improve the quality of current synthesis models), music analysis and perception (e.g., to study music performance), and music information retrieval (e.g., to find relevant acoustical descriptors for automatic instrument and/or performer identification). For synthesis purposes, perceptual clues can be used to calibrate the parameters of a synthesis model according to the behavior of an acoustic instrument, or to propose control devices based on a high-level perceptual description (e.g., “play a bright sound"). For music analysis purposes, the study of the links between the control of musical instruments and the perception of generated tones allow to determine the mechanical/acoustical parameters used by performers to alter the timbre of the tones, and find the relevant acoustical descriptors that account for such timbre variations. Such investigations are of importance when studying the role of timbre in musical interpretation (see e.g., [3, 4]) and can give new pedagogical insights. In the music information retrieval context, the results of such studies may be useful to achieve automatic instrument or performer identification based on timbre descriptors.

By varying the two main control parameters (related to the blowing pressure and reed aperture) of a physics-based synthesis model of clarinet [5], isolated clarinet tones were generated. Dissimilarity judgments were then collected and subjected to various statistical analyses to uncover the psychomechanical and psychoacoustical factors best predicting the perceptual representation of the various clarinet tones.

Timbre is the attribute of the auditory sensation that allows tones of equal pitch, loudness and duration to be distinguished [6]. From the cognitive point of view, timbre refers both to the identity of the sound sources (timbre is the attribute that allows two different musical instruments playing the same note, with identical sound lev-

Received 30 July 2009,
accepted 4 March 2010.

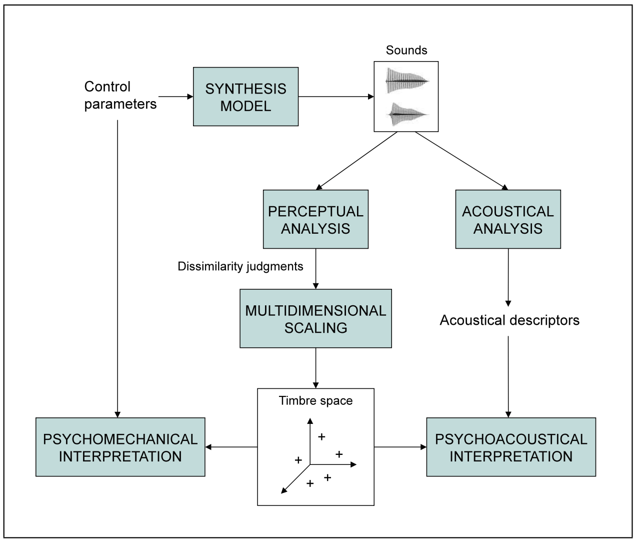


Figure 1. Multidimensional analysis of timbre.

els and durations, to be distinguished) and to the sound quality (e.g., two different guitar tones played with different plucking positions have different timbral qualities) [7]. The seminal works of Grey [8], Wessel [9], Krumhansl [10], Kendall and Carterette [11], and McAdams *et al.* [12] developed a methodology to represent and interpret the perceptual relationships between sounds differing in timbre. This methodology, which we rely upon in the present study, involves three main steps (see Figure 1): the collection of dissimilarity judgments among sounds varying in timbre, the construction of a geometrical representation of the dissimilarity judgments with a multidimensional scaling (MDS) procedure, and the interpretation of the representation with acoustical factors (timbre descriptors) and/or mechanical factors (control parameters). MDS techniques yield timbre spaces where the distances between the sounds represent the perceived dissimilarities statistically well [13]. Such techniques benefit from the fact that no assumptions are needed regarding the underlying acoustical descriptors that may predict the structure of the perceptual representation. MDS studies based on natural sounds from musical instruments, or synthetic tones generated to imitate the instruments of the orchestra, generally reported two- to four-dimensional timbre spaces. Caclin *et al.* [14] tested the perceptual relevance of some of the acoustical correlates of timbre space dimensions proposed in the psychoacoustic literature. This study confirmed that the Attack Time (linked to the rate of energy increase during the transient regime), the Spectral Centroid (frequency of the centroid of the spectrum), and the spectrum fine structure were major determinants of timbre. The results of MDS studies depend on the set of stimuli used in the experiment; the acoustical correlates found to characterize the broad differences of timbre between different musical instruments are not necessarily the same as the ones used to characterize the fine differences in timbre produced by the same instrument. One purpose of the present study is to find timbre descriptors well adapted to the range of timbres producible on the clarinet and which are potentially used by performers in a musical context (see [3]).

In section 2, the physics-based synthesis model used to generate clarinet tones will be presented. Some comparisons between natural and synthetic tones will be given in order to validate the use of the sounds produced by such a model in the study of the clarinet timbre. In section 3, the methods used in the perceptual experiment will be described. The statistical analyses conducted on the dissimilarity ratings will be presented and discussed in section 4. Finally, section 5 will present the conclusions and perspectives of the study.

2. Comparison of synthetic and natural clarinet tones

2.1. Description of the physics-based synthesis model

The physical model that was used to generate clarinet tones which formed the stimuli of the perceptual experiment is made of three coupled parts. The first part is linear and represents the bore of the instrument. The second part expresses the reed channel opening, linked to the reed displacement, by modeling the reed as a pressure driven mass-spring oscillator. The third part couples the previous ones in a nonlinear way. In what follows, we use dimensionless variables for the pressure, flow, and reed displacement, according to [15]. The internal variables of the model $p_e(t)$, $u_e(t)$ and $x(t)$ denote the dimensionless acoustic pressure, flow, and reed displacement, respectively. Capital letters denote their Fourier transforms. The dimensionless control variables $\gamma(t)$ and $\zeta(t)$ are related to the blowing pressure and the lip pressure on the reed, respectively, and will be detailed later.

2.1.1. Bore model

The bore is considered as a perfect cylinder of length L. Under classical hypothesis, its input impedance Z_e which links the acoustic pressure P_e and flow U_e in the mouthpiece, in the Fourier domain, is classically written as:

$$Z_e(\omega) = \frac{P_e(\omega)}{U_e(\omega)} = \mathrm{i} \tan\big(k(\omega)L\big). \qquad (1)$$

The wavenumber $k(\omega)$ corresponds to the classical approximation (the bore radius is large with respect to the thicknesses of the boundary layers). It is worth noting that for any flow signal, the acoustic pressure mostly contains odd harmonics since the input impedance corresponds to that of a quarter-wave resonator. At high frequencies, the increase of losses taken into account in the wavenumber induces non-zero values of the impedance for even harmonic frequencies. Hence, if the flow contains high frequency even harmonics, these will also appear in the pressure.

2.1.2. Reed model

The classical single mode reed model used here describes the dimensionless displacement $x(t)$ of the reed with respect to its equilibrium point when it is submitted to the dimensionless acoustic pressure $p_e(t)$:

$$\frac{1}{\omega_r^2}\frac{\mathrm{d}^2x(t)}{\mathrm{d}t^2}+\frac{q_r}{\omega_r}\frac{\mathrm{d}x(t)}{\mathrm{d}t}+x(t)=p_e(t), \qquad (2)$$

where $\omega_r = 2\pi f_r$ and $1/q_r$ are the angular frequency and the quality factor of the reed resonance, respectively. The reed displacement behaves like the pressure below the reed resonance frequency, as a pressure amplifier around the reed resonance frequency, and as a low-pass filter at higher frequencies.

2.1.3. Nonlinear characteristics

The classical nonlinear characteristics used here are based on the steady Bernoulli equation and link the acoustic flow (the product of the opening of the reed channel and the acoustic velocity) to the pressure difference between the bore and the mouth of the player. The effective reed channel opening $S(t)$ is expressed in terms of the reed displacement by

$$S(t)=\zeta(t)\Theta\big(1-\gamma(t)+x(t)\big)\big(1-\gamma(t)+x(t)\big), \qquad (3)$$

where Θ denotes the Heaviside function, the role of which is to keep the opening of the reed channel positive by cancelling it when $1-\gamma(t)+x(t)<0$. The parameter $\zeta(t)$ characterizes the whole embouchure and takes into account both the lip position and the section ratio between the mouthpiece opening and the resonator. It is proportional to the square root of the reed position at equilibrium and inversely proportional to the reed resonance frequency. The parameter $\gamma(t)$ is the ratio between the pressure $p_m(t)$ inside the player's mouth (assumed to be slowly varying on a sound period) and the static beating reed pressure P_M, i.e. the pressure needed to close the reed channel when there are no oscillations. In a lossless bore and a massless reed model, $\gamma(t)$ evolves from $\frac{1}{3}$ which is the oscillation threshold, to 1, which corresponds to the extinction threshold. The value $\frac{1}{2}$ corresponds to the so-called beating reed threshold, from which the reed touches (beats) the mouthpiece table during the course of each period of the oscillations.

Since the reed displacement corresponds to a linear filtering of the acoustic pressure, the reed opening mostly contains odd harmonics. Nevertheless, the function Θ introduces a singularity in $S(t)$ for playing conditions (given by $\zeta(t)$ and $\gamma(t)$) yielding a complete closing of the reed channel (dynamic beating reed case). This leads to a rise of even harmonics in $S(t)$ (saturating nonlinearity) and the generation of high frequencies.

The acoustic flow is finally given by

$$u_e(t)=S(t)\mathrm{sign}\big(\gamma(t)-p_e(t)\big)\sqrt{|\gamma(t)-p_e(t)|}. \qquad (4)$$

This nonlinear relation between pressure and opening of the reed channel explains why the flow spectrum contains all the harmonics.

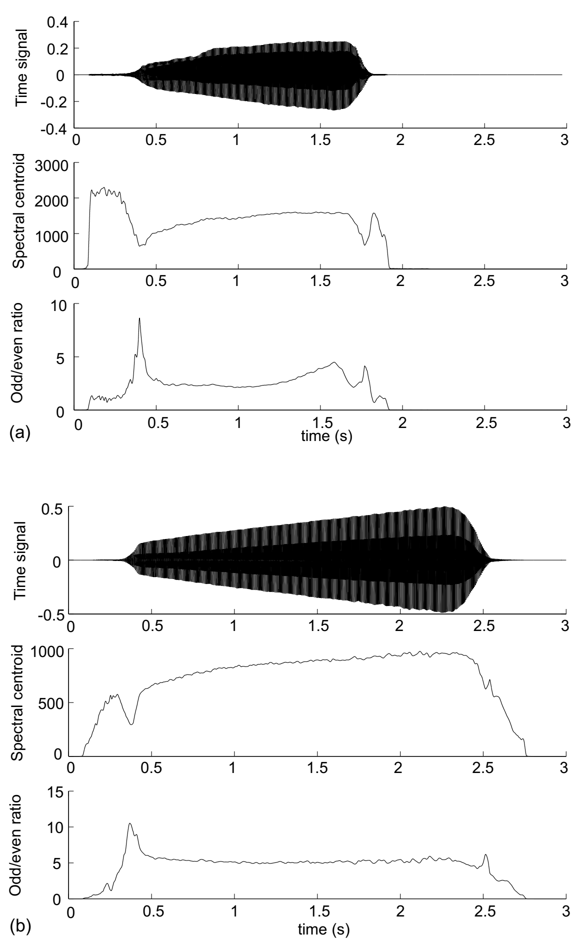


Figure 2. External pressure and timbre descriptors (Spectral Centroid and Odd/Even Ratio) in the cases of a natural (a) and a synthetic (b) clarinet tone.

2.1.4. Coupling of the reed and the resonator

Combining the impedance relation, the reed displacement and the nonlinear characteristics, the acoustic pressure, acoustic flow and reed displacement in the mouthpiece are solutions of the coupled equations (1), (2) and (3). The digital transcription of these equations and the computation scheme that explicitly solves this coupled system are achieved according to the method described in [5].

2.1.5. External pressure

From the pressure and the flow inside the resonator at the position of the mouthpiece, the external pressure is calculated by the relationship: $p_{ext}(t)=\frac{d}{dt}(p_e(t)+u_e(t))$, which corresponds to the simplest approximation of a monopolar radiation. This expression shows that in the clarinet spectrum, the odd harmonics are generated from both the flow and the pressure, while the even harmonics mostly come from the flow. Therefore, the ratio between odd and even harmonics can be considered as a signature of the "strength" of the nonlinearity in a non beating-reed situation.

2.2. Validation of the model

To investigate the relationships between the control of the instrument and the perception of the resulting sounds, we have chosen to use a synthesis model, since the control parameters $\gamma(t)$ and $\zeta(t)$ of such a model can be altered in a systematic and constrained manner. It is therefore necessary, as a first step, to check that the perceptual features considered here behave similarly for natural and synthetic sounds.

For that purpose, both an acoustic clarinet and a synthetic one were blown with the same time-varying pressure (parameter related to $\gamma(t)$). For the acoustic clarinet, an artificial mouth piloted in pressure by a PID (Proportional Integral Derivative) controller using a repetitive command was used. The role of the PID is to make an accurate pressure regulation to minimize the difference between the actual pressure in the mouth and the command pressure. Although the lip pressure on the reed can be measured on a human player, it varies significantly during transients. Furthermore, since the mechanical properties of the reed and the lip are unknown, the lip pressure cannot be linked to the reed channel opening at rest and therefore to the control parameter of the model $\zeta(t)$. Hence, the artificial mouth appeared to us as the only possibility to maintain this parameter as constant as possible in order to facilitate the comparison between the acoustic instrument and the model. The target pressure comprised a transient onset, followed by a linear pressure increase and then a decay transient. This specific pattern allows comparison of the behaviors of the timbre descriptors during transients and of the consistency of their variations with respect to a given steady-state pressure. The linear pressure increase has been chosen to test several constant blowing pressures within a single experiment. The actual mouth pressure (the result of the command) in the artificial mouth was recorded and used to feed the real-time synthesis model. In both experiments, the lip pressure on the reed was kept constant. The associated sound examples A1 and A2 are available at
`www.lma.cnrs-mrs.fr/~kronland/ClarinetTimbre/`.

Figure 2a shows respectively, from top to bottom, the variations in the external pressure, the Spectral Centroid (SC), and the Odd/Even Ratio (OER), over the duration of the note produced by the acoustic clarinet (the definitions of these timbre descriptors are given in section 3.5). At the oscillation threshold, around $t = 0.4$ s, the Spectral Centroid exhibits a minimum while the Odd/Even Ratio exhibits a maximum. This corresponds to the birth of the sound during which mostly the first harmonics are present. Then, the Spectral Centroid keeps on increasing steadily until $t = 1.8$ s, which corresponds to the release of the blowing pressure. A short increase followed by a short decrease of the centroid can be observed. After the attack, the Odd/Even Ratio decreases very quickly and then remains nearly constant with a slight decrease until $t = 1$ s. It then increases until $t = 1.6$ s. At the release of the blowing pressure, the OER exhibits an overshoot.

Non real-time simulations, for which all the physical variables of the problem are known, show that the amplitude of the flow increases steadily, and the balance between its even and odd harmonics remains constant, until the reed starts beating. This explains why the OER remains nearly constant with a slight decrease until $t = 1$ s. After that time, since the reed is beating, the amplitude of the flow oscillations are limited while those of the pressure are not. Moreover, the singularity in the flow due to its cessation over half a period introduces high frequency odd harmonics. This explains why the OER increases after $t = 1$ s.

The same features can be observed in Figure 2b for both the SC and the OER in the case of a synthetic tone. It can be observed that the duration of the sound is different to that produced by the acoustic clarinet. This difference can be attributed to the fact that the oscillation threshold is different. Indeed, it is important to point out that, although the lip pressure on the reed (related to $\zeta(t)$) is constant in both cases, they are different as the reed opening at rest cannot be measured on the artificial mouth. This probably also explains why the range of variations of the Spectral Centroid is similar in both cases while the average values are different, and why the average value of the Odd/Even Ratio is higher and its range of variations is smaller for the synthetic sound.

It is important to mention that the purpose of this example is not to make a model inversion using timbre descriptors, i.e to find the commands of the synthesis model from the analysis of an acoustic sound, but to demonstrate that the perceptual features under consideration evolve similarly with respect to a given transient or constant pressure command and that their variations can be related to the behavior of the physical variables of the problem.

3. Methods

3.1. Stimuli

A musician was asked to play a short sustained tone (E3, $f_0 \approx 164.81$ Hz) with the real-time synthesis model driven by a MIDI (Musical Instrument Digital Interface) controller adapted to wind instruments (Yamaha-WX5). The latter namely allows the control of the mouth pressure and the lip pressure on the reed. The blowing pressure was measured in the player's mouth cavity using a pressure probe (a very thin plastic tube, that does not disturb the player during tone production) linked to a pressure sensor (reference: Honeywell ASCX05DN, 0-5 psi differential piezzo-resistive pressure sensor). Figure 3 shows the variations in the measured blowing pressure over the duration of the note. The measured blowing pressure was normalized by dividing by the static beating reed pressure. This normalized blowing pressure, denoted $\gamma^{ref}(t)$, served as a reference when generating the stimuli by providing a maximum value, denoted γ_m.

A set of 150 tones ($f_0 \approx 164.81$ Hz, duration ≈ 2 s) was generated using the synthesis model by varying γ_m

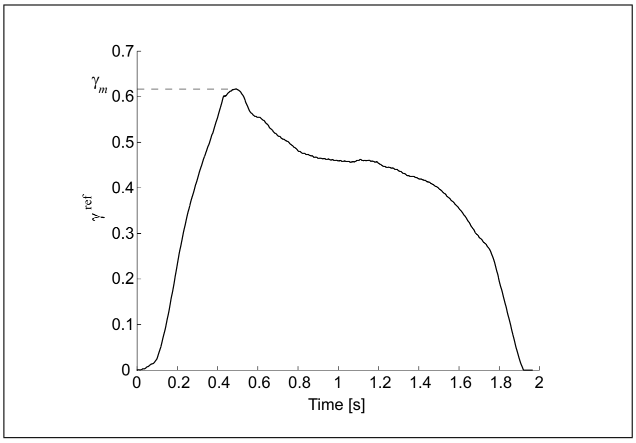


Figure 3. Blowing pressure of reference $\gamma^{ref}(t)$ (dimensionless). The maximal value of the pressure profile is denoted γ_m.

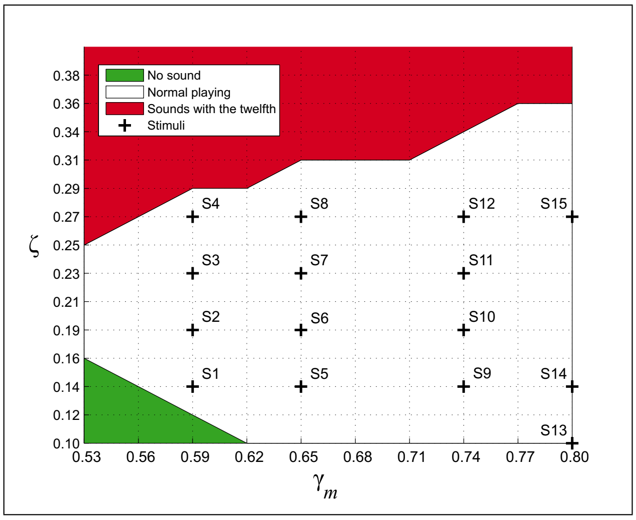


Figure 4. Control parameter space: the couples (γ_m, ζ) corresponding to the stimuli are indicated by crosses.

in the range $0.53 \leq \gamma_m \leq 0.80$ and ζ in the range $0.10 \leq \zeta \leq 0.40$. From this set, 15 tones were selected (see sound examples B1 to B15) which represented the palette of timbres producible with the model in traditional playing conditions (no squeaks, nor tones with the twelfth). Figure 4 represents the values of the control parameters associated with the stimuli. The sounds were equalized for loudness based on the informal loudness scaling of one listener. This was done by adjusting the loudness of the various sounds according to a reference signal (a 1kHz pure tone presented with an intensity of 70 dB SPL as measured at the position of the listener). Three other listeners validated that the amplitude-adjusted sounds were equal in loudness. This procedure was chosen since an automatic loudness equalization process based on the model defined by Zwicker and Fastl [16] for stationary sounds did not produce satisfying results. This may be due to the fact that clarinet tones are non stationary as they both contain a transient and a steady-state part. Furthermore, the synthesized tones present large differences in spectral richness, a factor which acts on the loudness [16] and which is poorly taken into account by the loudness model.

3.2. Participants

The experiment was conducted by a group of 16 participants (10 males, 6 females; age=23-47 years). 10 of them had received musical training and none reported any hearing loss.

3.3. Apparatus

The experiment took place in an audiometric cabin and the sounds were stored on the hard drive of an Apple iMac G4. The user interface was implemented in the Matlab environment[1]. The sounds were played to the participants using a STAX SRM-310 headphones system.

3.4. Procedure

The participants were first asked to listen to the 15 stimuli presented in random order to get used to the range of variation. After carrying out a training stage with randomly chosen practice trials, the participants had to rate the dissimilarity for all 105 possible pairs of non-identical stimuli. The within-pair order and the order of the pairs were randomized. The dissimilarity judgments were made by adjusting the position of a slider on a scale whose end points were labelled “Very similar" and “Very dissimilar" (in French) on the computer screen. The perceptual dissimilarities were digitized on a 0 to 1 scale with a discretization step of 0.01. The participants could listen to the stimuli as many times as they wished and were requested to keep the same rating strategy during the experiment. At the end of the test, they had to answer a questionnaire which was designed to establish the criteria they used to discriminate the various stimuli.

3.5. Presentation of the acoustical correlates of timbre

A large set of timbre descriptors was investigated in order to find some acoustical predictors of the perceptual structure underlying the discrimination of the various clarinet tones (see Table I). A precise description and formulation of these parameters can be found in [17]. The only definitions given here are those deemed most important in light of the results obtained in this study.

The **Attack Time (AT)** is given by

$$AT = t_{e_{AT}} - t_{s_{AT}}, \tag{5}$$

where $t_{s_{AT}}$ and $t_{e_{AT}}$ are the start and end of the attack times, respectively. They are defined as in [18], as the times at which the Root Mean Square (RMS) envelope attains 10% and 90% of its maximum value, respectively.

The short-term **Spectral Centroid (SC)** is defined as

$$SC(n) = \frac{\sum_{k=1}^{K} f(k) A_n(k)^2}{b_0 + \sum_{k=1}^{K} A_n(k)^2}, \tag{6}$$

[1] http://www.mathworks.com/products/matlab/

Table I. List of the timbre descriptors.

Timbre descriptor	Description
Attack Time (AT)	Linked to the rate of energy increase during the transient regime
Log. of the Attack Time (LAT)	Logarithm (decimal base) of AT
Release Time (RT)	Linked to the rate of energy decrease during the release part
Temporal Centroid (TC)	Time of the centroid of the acoustical energy
Spectral Centroid (SC)	Frequency of the centroid of the spectrum
Spectral Spread (SS)	Variance of the spectrum around its mean
Spectral Skewness (SSK)	Measure of the asymmetry of the spectrum around its mean
Spectral Kurtosis (SKU)	Measure of the flatness of the spectrum around its mean
Spectral Roll-off (SRO)	Cutoff frequency of the spectrum so that 95% of the energy is below
Spectral Flux (SF)	Measure of the fluctuation of the spectrum over time
Spectral Flux Attack (SFAT)	Spectral Flux calculated during the attack part
Harmonic Spectral Centroid (HSC)	Frequency of the centroid of the harmonic spectrum
Odd Spectral Centroid (OSC)	Frequency of the centroid of the spectrum of the odd harmonics
Even Spectral Centroid (ESC)	Frequency of the centroid of the spectrum of the even harmonics
Odd/Even Ratio (OER)	Ratio between the odd harmonics and the even harmonics energy
Spectral Irregularity (IRRKRI)	Measure of the irregularity of the spectral envelope (Krimphoff)
Spectral Irregularity (IRRKEN)	Measure of the irregularity of the spectral envelope (Kendall)
Spectral Irregularity (IRRJEN)	Measure of the irregularity of the spectral envelope (Jensen)
Tristimulus, 1^{st} coefficient (TR1)	Ratio between the fundamental component energy and the total energy
Tristimulus, 2^{nd} coefficient (TR2)	Ratio between the energy of harmonics 2, 3, and 4 and the total energy
Tristimulus, 3^{rd} coefficient (TR3)	Ratio between the energy of higher-order harmonics and the total energy

where $A_n(k)$ is the magnitude of the k^{th} coefficient of the Discrete Fourier Transform (DFT) associated with the frame centred at time n, $f(k)$ is the frequency associated with the k^{th} spectral component, and K denotes the last frequency bin to be considered. We used a threshold value b_0 as in the formulation given by Beauchamp [19], in order to force the descriptor to decrease at very low amplitudes when noise predominates.

The time-varying **Odd/Even Ratio (OER)** is obtained from

$$OER(t) = \frac{b_0 + \sum_{h=0}^{\frac{H}{2}-1} A_{2h+1}(t)^2}{b_0 + \sum_{h=1}^{\frac{H}{2}} A_{2h}(t)^2}, \tag{7}$$

where $A_h(t)$ is the instantaneous amplitude of the h^{th} harmonic component, H is the total number of considered harmonics (assumed here to be even so that an equal number of odd and even harmonics are compared). The b_0 threshold is used, as in equation 6, to prevent the descriptor from tending to infinity when noise predominates. OER is dimensionless; $OER < 1$, indicates that even harmonics are dominant, whereas $OER > 1$ indicates that odd harmonics are dominant.

The spectral and spectro-temporal descriptors were computed from the Short Term Discrete Fourier Transform (STDFT) by using a 1024-point Hann window (approximately 20 ms at the sampling frequency of 44.1 kHz) with a 50%-overlap. The DFT was calculated over 8192 points (the discretization step of the frequency scale is approximately 5 Hz). The computation of the spectral and spectro-temporal descriptors was made with two different amplitude scales (linear and power) in order to find the one which best fits the perceptual measures (see [20, 21] for a discussion on amplitude scales). b_0 was set at a value giving a spectral dynamic of 60 dB. The harmonic descriptors were calculated from the components' instantaneous amplitudes and frequencies. The latter were derived from short-band analytic signals associated with each component of the tone (see [22]). To obtain a global measure characterizing the whole signal, the instantaneous descriptors were averaged between the start of the attack part and the end of the release part.

4. Results and discussion

4.1. Multidimensional scaling (MDS)

Before analyzing the dissimilarity ratings using an MDS procedure, tests were first carried out to determine if the answers of the participants depended on whether they were musicians or non-musicians. To address this issue, correlations (Pearson) were computed among participants' dissimilarity ratings. A Hierarchical Cluster Analysis (complete linkage) was then performed on distances derived from the correlation measures (one minus Pearson's correlation coefficient) to detect whether certain participants systematically answered differently from the others. As the HCA did not reveal systematic differences between musicians and non-musicians, the dissimilarity ratings of all the participants were averaged for the MDS analysis.

A nonmetric MDS procedure (MDSCAL) was adopted (see e.g., [23]) as the dissimilarities possess ordinal scale properties. The procedure yields solutions such that the distances in the MDS space are in the same rank order as the original data. The goodness-of-fit criterion that it was necessary to minimize was Kruskal's stress. The number of dimensions of the MDS configuration was found by

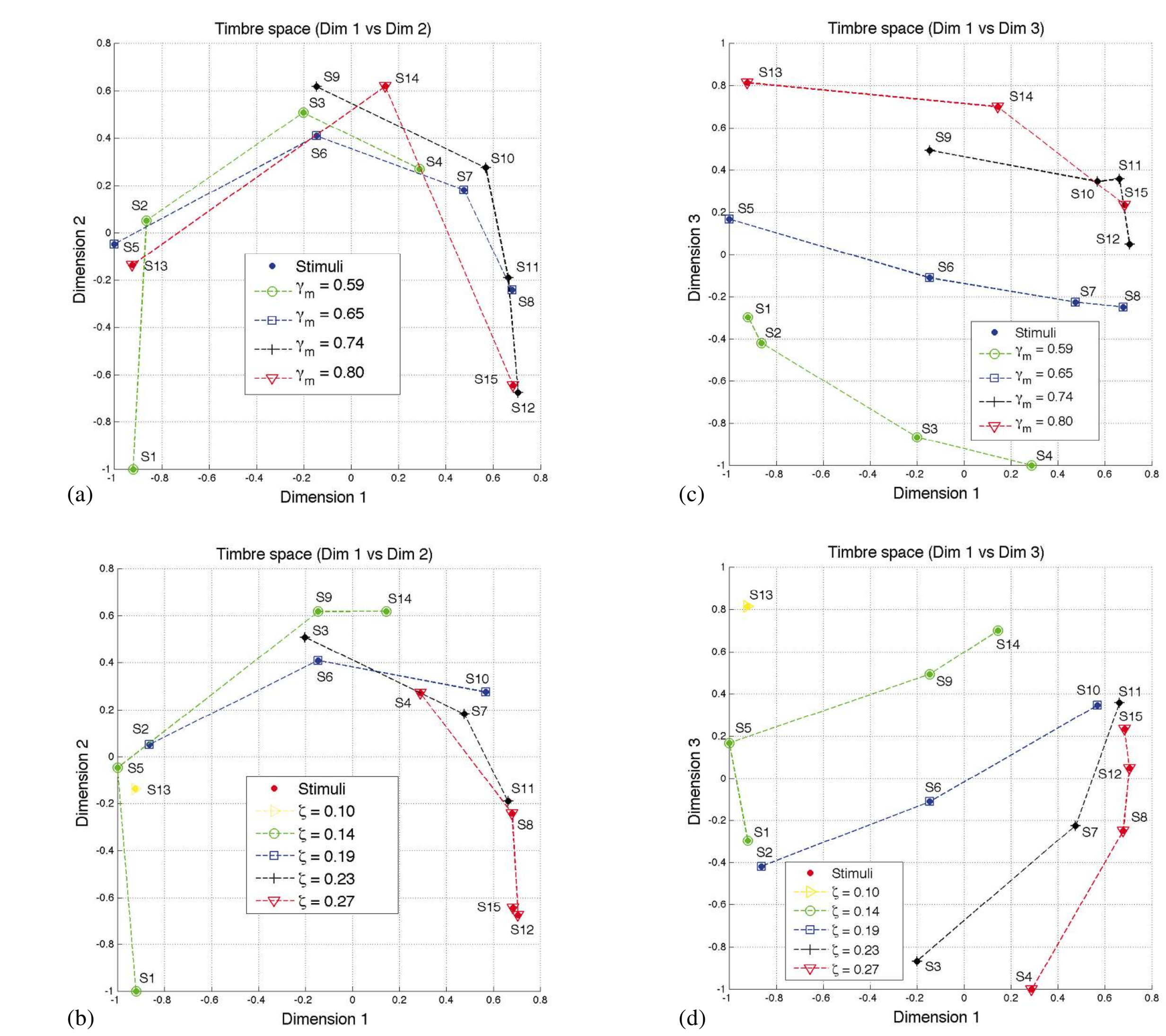


Figure 5. Two-dimensional projections of the MDS space. In Figures (a) and (c), the stimuli with same γ_m (related to the blowing pressure) are connected, in figures (b) and (d), the stimuli with same ζ (related to the lip pressure on the reed) are connected.

using several empirical criteria: the rate of decline of the stress as dimensionality increases (scree-elbow criterion) and the analysis of the Shepard diagram (representation of the distances in the MDS space as a function of the dissimilarities). The method indicated a 3-dimensional (3D) perceptual space (stress ≈ 0.05).

4.2. Interpretation of the clarinet timbre space

4.2.1. Mechanical correlates of the timbre space

Figure 5 shows the two-dimensional projections of the timbre space. To better visualize the influence of the control parameters on the positions of the stimuli in the perceptual space, the stimuli with identical γ_m values (Figures 5a and 5c) and ζ values (Figures 5b and 5d) have been linked together. The linear correlations (Pearson) between the positions of the stimuli in the space and the values of the control parameters are reported in Table II. It can be seen that the first dimension is correlated with ζ [r(13)=0.76, p<0.001] (the higher ζ, the higher the stimuli coordinate along the first dimension). Meanwhile, the third dimension is correlated with γ_m [r(13)=0.88, p<0.001] (the higher γ_m, the higher the stimuli coordinate along the third dimension). None of the control parameters were significantly correlated with the second dimension. Indeed, Figures 5a and 5b indicate a non-linear behavior of the stimuli coordinates along the second dimension as a

Table II. Pearson's correlation coefficient between the stimuli coordinates in the perceptual space and the control parameters γ_m and ζ (degrees of freedom = 13). Values that are not significant are not reported. The probabilities that the measures are independent are indicated as follows: *p<0.05, **p<0.01,***p<0.001.

	Dimension 1	Dimension 2	Dimension 3
γ_m	-	-	0.88^{***}
ζ	0.76^{***}	-	-0.54^{*}

Table III. Pearson's correlation coefficient between the stimuli coordinates in the perceptual space and the timbre descriptors (d.o.f. = 13). When both linear and power amplitude scales were used in the computation of the descriptors, the corresponding correlations are reported. Values that are not significant are not reported. The probabilities that the measures are independent are indicated as follows: *p<0.05, **p<0.01,***p<0.001. For each dimension, the two strongest correlations are indicated in bold.

Descriptors	Dimension 1		Dimension 2		Dimension 3	
Scale	*linear*	*power*	*linear*	*power*	*linear*	*power*
AT	-0.93^{***}		-		-	
LAT	$\mathbf{-0.95}^{***}$		-		-	
RT	-		-		-0.78^{***}	
TC	-		-		$\mathbf{0.80}^{***}$	
SC	0.90^{***}	$\mathbf{0.94}^{***}$	-	-	-	-
SS	0.63^{*}	0.88^{*}	$\mathbf{-0.63}^{*}$	-	-	-
SSK	-0.93^{***}	-0.72^{**}	-	-0.54^{*}	-	-
SKU	-0.88^{***}	-0.77^{***}	-	-	-	-
SRO	0.92^{***}		-	-	-	-
SF	-	-0.58^{*}	-	-	-	-
SFAT	$\mathbf{-0.95}^{***}$		-	-	-	-
HSC	0.91^{***}	$\mathbf{0.94}^{***}$	-	-	-	-
OSC	0.91^{***}	$\mathbf{0.94}^{***}$	-	-	-	-
ESC	0.90^{***}	0.81^{***}	-	-	-	-
OER	-	0.84^{***}	-	-	$\mathbf{0.87}^{***}$	-
IRRKRI	-0.79^{***}		-		-	
IRRKEN	-		0.53^{*}		0.69^{**}	
IRRJEN	0.79^{***}		-		-	
TR1	-0.92^{***}	-0.88^{***}	-	-	-	-
TR2	-0.71^{**}	0.62^{*}	$\mathbf{0.67}^{**}$	0.62^{*}	-	-
TR3	$\mathbf{0.95}^{***}$	$\mathbf{0.95}^{***}$	-	-	-	-

function of γ_m and ζ (see the bell-like shape when one of the parameters is fixed and the other varies). The second dimension might be correlated with a non-linear combination of the parameters γ_m and ζ, but we did not look for such a combination as it would not be easily interpretable from the physical point of view.

4.2.2. Acoustical correlates of the timbre space

Table III shows the correlations between the positions of the stimuli along the various dimensions of the perceptual space and the timbre descriptors previously presented in Table I.

It can be seen that the first dimension is correlated with the Logarithm of the Attack Time (LAT) [r(13)=-0.95, p<0.001], with various Spectral Centroid descriptors (SC, HSC, OSC [r(13)=0.94, p<0.001]), and with the third Tristimulus coefficient TR3 [r(13)=0.95, p<0.001]. Indeed, this dimension separates tones having long Attack Times and low Spectral Centroids, generated with small values of reed opening and blowing pressure (S1, S2, S5), or very small reed opening and high blowing pressure (S13), from tones having brief Attack Times and high Spectral Centroids, generated with high values of reed opening and mouth pressure (S8, S11, S12, S15). The fact that the Attack Time, the Spectral Centroid, and the third coefficient of Tristimulus are intercorrelated (see Table IV) is consistent with the physics of the instrument, as the greater the reed opening and mouth pressure, the faster the start of the self-sustained oscillations (small AT), and the greater the spectral richness (high SC and TR3) (see the Worman-Benade laws in [24]). Note that AT, SC and TR3 are not correlated with the other dimensions of the timbre space. It is also worth pointing out that the correlations with the Spectral Centroid increased when the computation of the descriptor was made using a power amplitude scale, which assigns a greater weight to the dominant harmonics.

Referring back to Table III, the second dimension is only weakly correlated with the second Tristimulus coefficient TR2 [r(13)=0.67, p<0.01], and the Spectral Spread [r(13)=-0.63, p<0.05] computed with a linear amplitude scale. Furthermore, these descriptors are more significantly correlated with the first dimension (see Table III), as they covary with AT, SC and TR3 (see Table IV). The descriptor TR2, based on two even harmonics (2 and 4) and one odd harmonic (5), is not adapted to characterize the difference in behavior between odd and even harmonics.

The third dimension is well correlated with the Odd/Even Ratio [r(13)=0.87, p<0.001] computed on a linear amplitude scale, and with the Temporal Centroid [r(13)=0.80, p<0.001]. Figure 6 displays the evolution of the Odd/Even Ratio as the mouth pressure and the reed aperture increases. The Odd/Even Ratio globally increases as the mouth pressure and reed opening increase (the odd harmonics grow faster than the even ones), attains a maximal value and then decreases past a certain threshold of ζ. We showed in a previous study that the decrease of OER is due to the beating reed situation, occurring when the reed

Table IV. Intercorrelations between the major timbre descriptors and the control parameters. The spectral descriptors have been computed using a power amplitude scale. ** and * indicate that the correlation is significant at the 0.01 and 0.05 levels, respectively.

	AT	SC	TR3	TR2	SS	OER	TC	γ_m	ζ
AT	1	-0.94^{**}	-0.95^{**}	-0.77^{**}	-0.83^{**}	-0.80^{**}	0.66^{**}	-0.20	-0.80^{**}
SC		1	0.99^{**}	0.69^{**}	0.96^{**}	0.89^{**}	-0.56^{*}	0.35	0.83^{**}
TR3			1	0.69^{**}	0.94^{**}	0.87^{**}	-0.58^{*}	0.31	0.84^{**}
TR2				1	0.52^{*}	0.74^{**}	-0.28	0.42	0.36
SS					1	0.89^{**}	-0.42	0.46	0.77^{**}
OER						1	-0.14	0.71^{**}	0.49
TC							1	0.56^{*}	-0.87^{**}
γ_m								1	-0.18
ζ									1

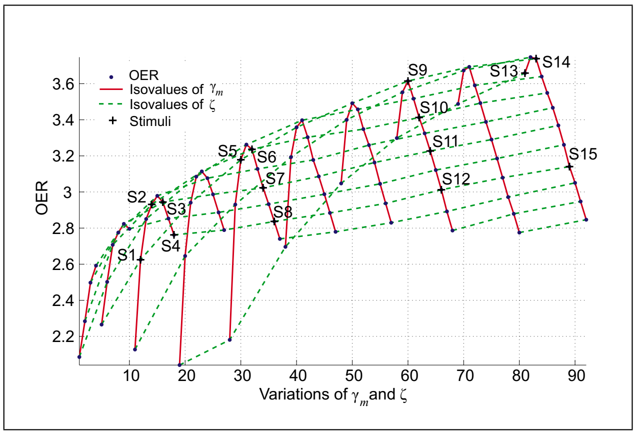


Figure 6. Evolution of the Odd/Even Ratio (OER) as a function of the control parameters γ_m (related to the blowing pressure) and ζ (related to the lip pressure on the reed). OER has been computed on a linear amplitude scale. The parameters used to generate the sounds are given in Figure 4 for normal playing conditions. The solid lines link the sounds with isovalues of γ_m. The dashed lines link the sounds with isovalues of ζ.

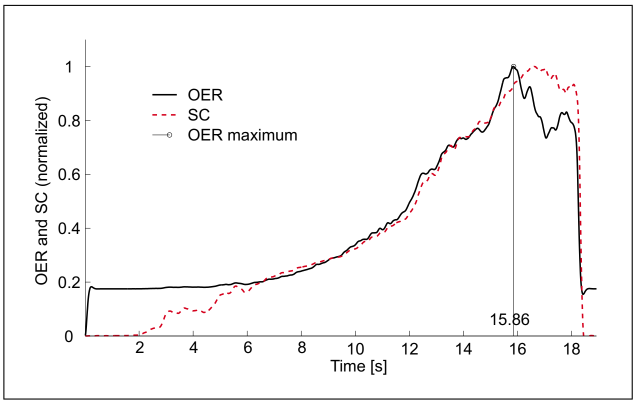


Figure 7. Evolution of the Odd/Even Ratio (OER) and Spectral Centroid (SC) for a natural clarinet crescendo.

beats the mouthpiece table and closes the reed channel aperture during a portion of its oscillations [25]. The closure of the reed channel induces the appearance of high-order even harmonics in the external pressure which grow faster than their relative odd harmonics. We verified that OER showed a similar behavior on a natural clarinet sound (see Figure 7 and sound example C). It is worth pointing out that a net timbre change is audible when OER attains its maximal value (t ≈ 15.9 s) and then decreases (due to the appearance of high-order even harmonics which give a metallic color to the sound). OER hence brings complementary information on timbre to that provided by the Spectral Centroid, as this timbre change at about 15.9 s is not revealed by SC, which keeps on increasing during the whole crescendo. These findings on OER are consistnt with the fact that the stimuli coordinates along the third dimension of the perceptual space globally increase as γ_m increases, and decrease as ζ increases (see Figure 5c).

4.3. Hierarchical Cluster Analysis (HCA)

The dissimilarity measures were also subjected to a Hierarchical Cluster Analysis (complete linkage) in order to test if certain groups of sounds presented systematic differences. Three main clusters of sounds were obtained. The first one associates tones generated using small (γ_m, ζ) values with the ones generated using high γ_m and very small ζ values (S1, S2, S5, S13). This cluster corresponds to tones which are not bright (small Spectral Centroid) and have a long attack part (high Attack Time). These results show that from the control point of view such timbral qualities (small brightness, long attack) can be obtained with two different techniques, either by applying a weak blowing pressure (small γ_m), while keeping a tight embouchure (small ζ), or by applying a strong blowing pressure (high γ_m), while keeping an even tighter embouchure (very small ζ). The second cluster gathers sounds generated using moderate values of γ_m and ζ (S3, S6, S9, S10, S14). The corresponding tones have moderate values of SC and AT. The third cluster associates tones generated using high (γ_m, ζ) values with ones generated using a high ζ value and a very small γ_m value (S4, S7, S8, S11, S12, S15). The tones gathered in this cluster are very bright (high SC) and present a brief attack part (small AT). Again, the results indicate that two combinations of mouth pressure and reed aperture can be used by the performer to obtain bright tones with a brief attack part.

Figure 8 represents the three-dimensional clarinet timbre space obtained with the MDS analysis and shows the

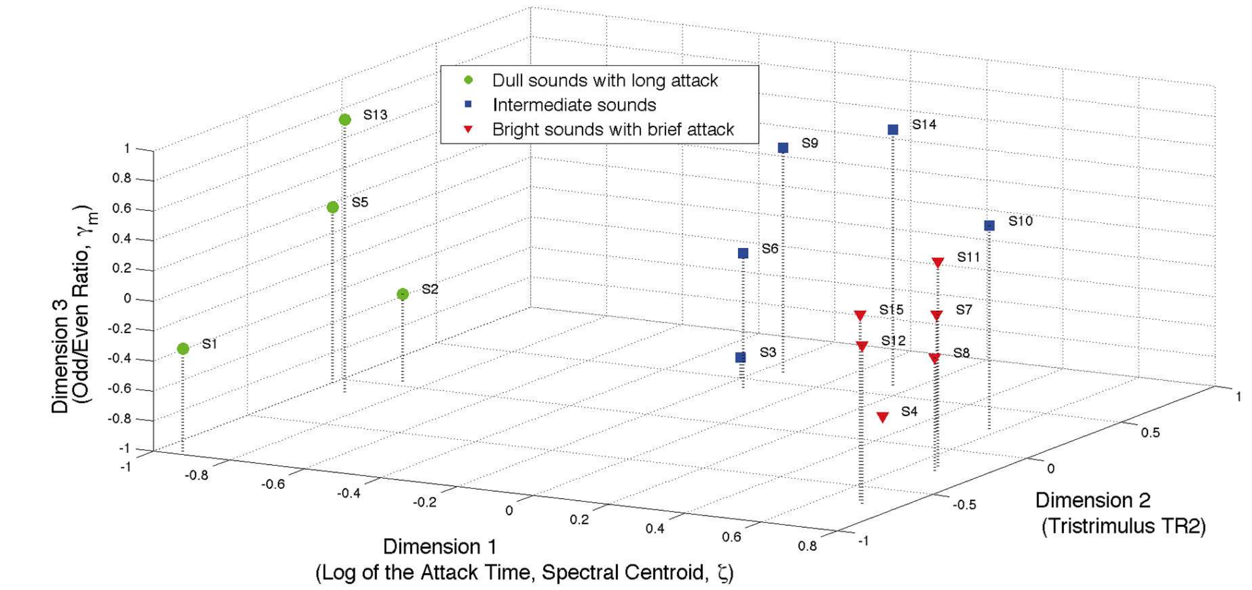


Figure 8. Three-dimensional clarinet timbre space and its mechanical and acoustical correlates. The three main clusters yielded by the HCA are represented with distinct markers.

Table V. Expressions used by the participants (P1 to P16) to describe their sound discrimination strategy (translated from French). They are gathered in three main categories: brightness, attack, and tension. Note that these categories are not necessarily independent. Certain expressions can belong to several categories at the same time.

	Brightness	Attack	Tension
P1	“nasal"		“soft"
P2	“rich and nasal"	“attack more or less soft"	
P3	“power of high-order harmonics"		
P4	“dull vs bright sounds"	“softness of the attack"	“softness of the attack"
P5		“attack"	
P6	“nasal"	“attack"	
P7	“richness of the sound in the trebles"	“dynamic of the transient" “duration of the transient"	
P8			“aggressive vs soft sounds"
P9	“sharp sound, bright sound"	“attack"	“sharp sound"
P10	“round notes" / “nasal notes"		
P11	“brightness of the sound after the attack"	“attack time", “attack intensity"	
P12	“sizzling"		
P13	“richness of the sounds (more high harmonics)"		
P14	“intensity of the sizzling"		
P15	“round sound, full versus narrow sound" “brightness"	“attack (soft or brief and intense)"	
P16	“brightness"	“attack (or more generally transient)"	

timbre descriptors and control parameters which best predict the distribution of the tones in the space. The three main clusters of tones yielded by the HCA are also reported in the figure (dull sounds with long attack, intermediate sounds, and bright sounds with brief attack).

4.4. Qualitative analysis of the verbal descriptions

The analysis of the questionnaire revealed three main factors of sound discrimination: the brightness of the sounds, the nature of their attack, and the sensation of tension (see Table V).

15 out of 16 participants employed expressions which refer to the percepts of *brightness* (also called *nasality*), either directly (e.g., “bright", “nasal"), or indirectly (“rich", “power of high-order harmonics"). 10 out of 16 participants also mentioned having relied on the nature of the attack (e.g., “attack time", “attack intensity") to discriminate the sounds. These qualitative observations are consistent with the preceding quantitative analyses of the timbre space. The first dimension indeed appeared to be well correlated with descriptors characterizing the nature of the attack (e.g., the Attack Time, characterizing its rapidity, the

Spectral Flux computed within the attack part and characterizing the non-synchronicity of the harmonic components), or with descriptors characterizing the spectral richness (e.g., the Spectral Centroid, the third Tristimulus coefficient TR3). It is worth pointing out that the Spectral Centroid strongly maps with the sensation of brightness [26], and has recurrently proved to be a good acoustical correlate of timbre space dimensions (see [8, 10, 12, 14]).

In [27], the Logarithm of the Attack Time and the Spectral Centroid are the best predictors of the first two dimensions of the 3D-timbre space associated with the discrimination of orchestral instruments' tones. As pointed out in section 4.2.2, the clarinet tones used in the perceptual experiment present covariant Attack Times and Spectral Centroids. It is then logical that AT and SC are both correlated with the first dimension. However, this does not prove that AT and SC would also be covariant with other configurations of the control parameters (e.g., different blowing pressure profiles). Hence, both descriptors seem to be needed to build a timbre model of clarinet tones.

Some expressions seem to refer to the timbral quality associated with the beating reed situation (e.g., "sizzling", "intensity of the attack", "richness of the sounds") which, as shown above, is highlighted by the shape of the Odd/Even Ratio. The timbre changes associated with Spectral Irregularity variations are well discriminated by listeners [26]. These results support the fact that the Odd/Even Ratio showed itself to be a good acoustical correlate of the third dimension of the timbre space.

Certain participants have discriminated the sounds according to the sensation of tension that they procured (e.g., "soft", "aggressive"). The soft tones may correspond to those which are dull and have a long attack, whereas the aggressive ones may be associated with those which are very bright and have a fast attack. Indeed, Zwicker and Fastl showed that the sensation of annoyance procured by sounds equal in loudness increased as the acuity (Spectral Centroid formulation based on an auditory model) increased [16].

5. Conclusions

This study shows that the perceptual relationships among different clarinet timbres generated with a synthesis model can be modelled in a 3-dimensional space. The underlying attributes used by participants when rating the dissimilarities between pairs of stimuli were investigated both from quantitative (mechanical and acoustical correlates of the timbre space dimensions) and qualitative (verbal descriptions) analyses. These analyses revealed that the participants were sensitive to changes of timbre induced by variations of the control parameters of the model (correlated to the blowing pressure and the lip pressure on the reed). The perceptual representation of clarinet timbres was best explained by timbre descriptors characterizing the attack of the tones (Attack Time), the spectral richness (Spectral Centroid, third Tristimulus coefficient), and the irregularity of the spectrum (Odd/Even Ratio). It is worth pointing out that the results from a study on the perceptual categorization of natural clarinet tones also showed that the Spectral Centroid is a major determinant of clarinet timbre [28].

Similar behaviors of the perceptual features (Spectral Centroid, Odd/Even Ratio) were found between natural and synthetic clarinet tones which validate the use of the physics-based synthesis model to explore the relationships between the control gestures of the instrument, the generated timbres and their perception. In particular, these analyses indicated that the Odd/Even Ratio is a good predictor of the change of timbral quality occurring during the beating reed situation for which high-order even harmonics are amplified.

The verbal descriptions given by the participants highlighted the influence of certain timbre changes on the sensation of tension. These results are interesting in the musical context, as music usually involves a succession of tensions and releases. According to Vines *et al.* [29], the sensation of tension is indeed correlated to the emotional responses of listeners. In other work, we have shown that musicians use timbre variations to vary their expression, and that such timbre variations induce changes in the emotional responses of listeners [3, 4].

6. Acknowledgments

This project was partly supported by the French National Research Agency (ANR JC05-41996, "senSons" http://www.sensons.cnrs-mrs.fr/, and ANR "Consonnes" http://www.consonnes.cnrs-mrs.fr/).
The authors would like to thank the reviewers for their helpful comments and corrections.

References

[1] J. Brymer: Clarinette. Hatier, Paris, 1979, (Collection Yehudi Menuhin).

[2] J. Kergomard: Le timbre des instruments à anche. – In: Le timbre, métaphore pour la composition. C. Bourgeois (ed.). I.R.C.A.M., 1991, 224–235.

[3] M. Barthet, P. Depalle, R. Kronland-Martinet, S. Ystad: Acoustical correlates of timbre and expressiveness in clarinet performance. Music Perception (in press) (2010).

[4] M. Barthet, P. Depalle, R. Kronland-Martinet, S. Ystad: Analysis-by-synthesis of timbre, timing, and dynamics in expressive clarinet performance. Music Perception (in press) (2010).

[5] P. Guillemain, J. Kergomard, T. Voinier: Real-time synthesis of clarinet-like instruments using digital impedance models. J. Acoust. Soc. Am. **118** (2005) 483–494.

[6] ANSI: USA Standard Acoustical Terminology. American National Standards Institute, New York, 1960.

[7] J. M. Hajda, R. A. Kendall, E. C. Carterette, M. L. Harshberger: Methodological issues in timbre research. 2nd ed. Psychology Press, New York, 1997, 253–306.

[8] J. M. Grey: Multidimensional perceptual scaling of musical timbres. J. Acoust. Soc. Am. **61** (1977) 1270–1277.

[9] D. L. Wessel: Timbre space as a musical control structure. Computer Music Journal **3** (1979) 45–52.

[10] C. L. Krumhansl: Why is musical timbre so hard to understand ? Proc. of the Marcus Wallenberg Symposium held in Lund, Sweden, Amsterdam, 1989, S. Nielzén, O. Olsson (eds.), Excerpta Medica, 43–53.

[11] R. A. Kendall, E. C. Carterette: Perceptual scaling of simultaneous wind instrument timbres. Music Perception **8** (1991) 369–404.

[12] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, J. Krimphoff: Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychological Research **58** (1995) 177–192.

[13] R. Plomp: Timbre as a multidimensional attribute of complex tones. – In: Frequency Analysis and Periodicity Detection in Hearing. R. Plomp, G. F. Smoorenburg (eds.). A. W. Sijthoff, Leiden, 1970.

[14] A. Caclin, S. McAdams, B. K. Smith, S. Winsberg: Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. J. Acoust. Soc. Am. **118** (2005) 471–482.

[15] J. Kergomard: Elementary considerations on reed-instrument oscillations. – In: Mechanics of Musical Instruments. A. Hirschberg et al. (eds.). Springer-Verlag, New York, 1995.

[16] E. Zwicker, H. Fastl: Psychoacoustics: Facts and models. Springer-Verlag, New York, 1990.

[17] M. Barthet: De l'interprète à l'auditeur: une analyse acoustique et perceptive du timbre musical. Dissertation. Université Aix-Marseille II, 2008.

[18] G. Peeters: A large set of audio features for sound description (similarity and description) in the cuidado project. Tech. Rept. version 1.0, I.R.C.A.M., Paris, 2004.

[19] J. W. Beauchamp: Synthesis by spectral amplitude and brightness matching of analyzed musical instrument tones. J. Audio Eng. Soc. **30** (1982) 396–406.

[20] J. M. Grey, J. W. Gordon: Perception of spectral modifications on orchestral instrument tones. Computer Music Journal **11** (1978) 24–31.

[21] J. Marozeau, A. de Cheveigné, S. McAdams, S. Winsberg: The dependency of timbre on fundamental frequency. J. Acoust. Soc. Am. **114** (November 2003) 2946–2957.

[22] B. Picinbono: On instantaneous amplitude and phase of signals. IEEE Transactions on Signal Processing **45** (1997) 552–560.

[23] W. R. Dillon, M. Goldstein: Multivariate analysis. John Wiley & Sons, New York, 1984, (Wiley series in probability and mathematical statistics).

[24] A. H. Benade, S. N. Kouzoupis: The clarinet spectrum: Theory and experiment. J. Acoust. Soc. Am. **83** (January 1988) 292–304.

[25] M. Barthet, P. Guillemain, R. Kronland-Martinet, S. Ystad: On the relative influence of even and odd harmonics in clarinet timbre. Proc. Int. Comp. Music Conf. (ICMC'05), Barcelona, Spain, 2005, 351–354.

[26] R. A. Kendall, E. C. Carterette: Verbal attributes of simultaneous wind instrument timbres: I. von bismarck's adjectives. II. adjective induced from piston's orchestration. Music Perception **10** (1993) 445–468; 469–502.

[27] S. McAdams, J. W. Beauchamp, S. Meneguzzi: Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. J. Acoust. Soc. Am. **105** (1999) 882–897.

[28] M. A. Loureiro, H. B. de Paula, H. C. Yehia: Timbre classification of a single instrument. ISMIR 2004 5th International Conference on Music Information Retrieval, Barcelona, Spain, 2004, Audiovisual Institute, Universitat Pompeu Fabra.

[29] B. W. Vines, C. L. Krumhansl, M. M. Wanderley, D. J. Levitin: Cross-modal interactions in the perception of musical performance. Cognition **101** (2006) 80–113.

Sound Categorization and Conceptual Priming for Nonlinguistic and Linguistic Sounds

Mitsuko Aramaki[1,2], Céline Marie[1,2], Richard Kronland-Martinet[3], Sølvi Ystad[3], and Mireille Besson[1,2]

Abstract

■ The aim of these experiments was to compare conceptual priming for linguistic and for a homogeneous class of nonlinguistic sounds, impact sounds, by using both behavioral (percentage errors and RTs) and electrophysiological measures (ERPs). Experiment 1 aimed at studying the neural basis of impact sound categorization by creating typical and ambiguous sounds from different material categories (wood, metal, and glass). Ambiguous sounds were associated with slower RTs and larger N280, smaller P350/P550 components, and larger negative slow wave than typical impact sounds. Thus, ambiguous sounds were more difficult to categorize than typical sounds. A category membership task was used in Experiment 2. Typical sounds were followed by sounds from the same or from a different category or by ambiguous sounds. Words were followed by words, pseudowords, or nonwords. Error rate was highest for ambiguous sounds and for pseudowords and both elicited larger N400-like components than same typical sounds and words. Moreover, both different typical sounds and nonwords elicited P300 components. These results are discussed in terms of similar conceptual priming effects for nonlinguistic and linguistic stimuli. ■

INTRODUCTION

Word processing is specific in that the sequences of phonemes that form spoken words (or the sequence of graphemes that form written words) are not meaningful by themselves but acquire meaning through the process of double articulation. Thus, there is generally no direct relationship between the sound (or form) of a word and its meaning (de Saussure, 1916). By contrast, a causal relationship exists between the perceptual characteristics of environmental sounds (e.g., broken glass) and the meaning that is derived from them (e.g., a glass was broken; Ballas, 1993; Ballas & Howard, 1987). On the basis of such differences, one may expect words and environmental sounds to be processed differently. However, results of several experiments that have used the ERP method to address this question argue in favor of the similarities rather than the differences in the processing of words and environmental sounds (Orgs, Lange, Dombrowski, & Heil, 2006, 2007, 2008; Cummings et al., 2006; Plante, Van Petten, & Senkfor, 2000; Van Petten & Rheinfelder, 1995).

One of the first studies of conceptual priming was conducted by Van Petten and Rheinfelder (1995). In their first experiment, environmental sounds were used as primes and related words, unrelated words, and pseudowords were used as targets. Participants were asked to decide whether the target stimulus was a word or not (lexical decision task). Results showed higher error rate for pseudowords than for words and faster RTs for related than for unrelated words. Thus, these results demonstrated conceptual priming between environmental sounds and words. To examine the neural basis of this effect, they took advantage of the N400 component (Kutas & Hillyard, 1980) to compare priming when the prime is an environmental sound and the target a related or an unrelated printed word and vice versa. Results revealed that conceptual priming, as reflected by the N400 effect (i.e., the difference between unrelated and related targets), was very similar for environmental sounds and for words. This result argues in favor of the similarity of the neural processes involved in computing the meaning of words and environmental sounds. However, although the typical, slightly larger over the right than left hemisphere distribution of the N400 effect was found for target words (Kutas, Van Petten, & Besson, 1988; Kutas & Hillyard, 1982), the N400 effect to environmental sounds was larger over the left hemisphere. Similar interhemispheric differences in the N400 effect were reported in a subsequent study (Plante et al., 2000) in which priming for pairs of line drawing and environmental sounds, on one side, and for pairs of printed and spoken words, on the other side, was compared using the same task as in Van Petten and Rheinfelder (1995).

More recently, Orgs et al. (2006) used printed words as primes followed by environmental sounds as targets (or vice versa). Primes and targets were semantically related or

[1]CNRS—Institut de Neurosciences Cognitives de la Méditerranée, Marseille Cedex, France, [2]Université Aix-Marseille, Marseille Cedex, France, [3]CNRS—Laboratoire de Mécanique et d'Acoustique, Marseille Cedex, France

unrelated, and participants were asked to decide whether the words and environmental sounds fitted together or not. Results showed slower RTs and larger N400 amplitude for unrelated than for related targets. Moreover, the N400 effect showed an earlier onset for environmental sounds than for words over parieto-occipital sites. In subsequent studies with similar design and stimuli, participants were asked to perform either a physical task or a semantic task (Orgs et al., 2007, 2008). In both tasks, the authors found an N400 effect that was taken to reflect the automatic processes that mediate analysis of sound meaning. Finally, Cummings et al. (2006) also used a cross-modal priming design with pictures presented as primes and related or unrelated spoken words, environmental sounds, or nonmeaningful sounds (defined by the authors as "computer-generated sounds that were not easily associated with any concrete semantic concept"; Cummings et al., 2006, p. 104) presented as targets. Participants were asked to decide whether the two stimuli matched or mismatched. They reported an N400 effect for both words and environmental sounds but not for nonmeaningful sounds. Again, the N400 effect had an earlier onset for environmental sounds than for words, but in contrast with Orgs et al. (2006), the N400 effect was larger for environmental sounds than for words over frontal sites (F3/F4). Moreover, in contrast to Plante et al. (2000) and Van Petten and Rheinfelder (1995), they reported no interhemisphere differences. Findings from an experiment using event-related desynchronization to compare words and environmental sounds suggest the involvement of left-lateralized phonological and semantic processes for words and of distributed processes in both hemispheres for environmental sounds (Lebrun et al., 2001). On the basis of these results, Lebrun et al. (2001) suggested a common semantic system for both words and environmental sounds but with more specific perceptual processing for the later.

Other studies have also examined conceptual priming between music and language. Results have shown the occurrence of an N400 component to unrelated visual word targets when primes were long musical excerpts (several seconds; Koelsch et al., 2004). Recently, Daltrozzo and Schön (2009) used either words or short musical excerpts (1 sec) as primes and targets and found an N400 effect in both cases. However, the scalp distribution of the N400 effect was temporal for musical excerpts and parietal for visual word targets. To our knowledge, only one study has examined priming effects for pairs of musical excerpts presented both as prime and as target stimuli (Frey et al., 2009). Although the musical excerpts used as primes conveyed specific concepts, the musical excerpts used as targets either conveyed the same concept as the prime (congruous) or started with the same concept but shifted midstream into another concept (incongruous). Results showed an N400-like component, which was largest over right frontal regions, to incongruous musical excerpts in nonmusicians.

Taken together, results of these studies, which used different experimental designs, stimuli, and tasks (i.e., lexical decision task, matching tasks, physical or semantic priming), nevertheless concur in showing that environmental sounds or musical excerpt targets that are unrelated to word or picture primes elicit increased negative components in the N400 latency band compared with related targets. However, the scalp distribution of these effects varies either between hemispheres (Lebrun et al., 2001; Plante et al., 2000; Van Petten & Rheinfelder, 1995) or along the anterior-posterior dimension (Daltrozzo & Schön, 2009; Cummings et al., 2006; Orgs et al., 2006). Thus, whether conceptual priming for environmental sounds, music, and words rely on similar or different processes still remains an open issue.

The Present Studies

In the different studies summarized above (except for Frey et al., 2009), some form of cross-modal priming was always used between pictures, printed or spoken words, musical excerpts, on one side, and nonlinguistic sounds (environmental sounds, musical excerpts), on the other side. As a consequence, the presence of words or pictures in the experimental design may have encouraged the use of linguistic encoding strategies. Thus, the possibility remains that, within such experimental contexts, participants associated a verbal label to each sound thereby explaining the similarity of the N400 for words and environmental sounds. In a recent study, Schön, Ystad, Kronland-Martinet, and Besson (2009) also used a cross-modal priming design but they tried to minimize linguistic mediation by presenting sounds with no easily identifiable sound sources (i.e., a verbal label could not easily be associated to the sounds). Again, larger N400 components were found for targets (sounds or words) that were unrelated to the primes (words or sounds). However, this cross-modal design still involved words as prime or target stimuli. Therefore, our first aim was to reduce the influence of linguistic mediation by using only nonlinguistic sounds as prime and target. Moreover, in previous studies, the set of nonlinguistic sounds that were used were often very diverse and nonhomogeneous (e.g., animal or human nonspeech sounds, instrumental sounds, and everyday life sounds). As these different types of sounds may engage different processes, our second aim was to use only one homogeneous class of environmental sounds: impact sounds from wood, metal, or glass. Finally, to our knowledge, priming effects including only linguistic primes and linguistic targets on one side and nonlinguistic primes and nonlinguistic targets on the other side have never been directly compared within subjects. Thus, our third aim was to directly compare within-subjects conceptual priming for impact sounds and for linguistic sounds by using the same task with both types of stimuli. However, to use a priming design including only nonlinguistic sounds as stimuli, we first needed to create typical sounds from different impact sound categories (i.e.,wood, metal, glass) and ambiguous sounds. The aims of Experiment 1 were to create such sounds and to study the neural basis of impact sound categorization.

EXPERIMENT 1: SOUND CATEGORIZATION

To create typical and ambiguous impact sounds, we used a morphing technique. First, sounds from three material categories (wood, metal, and glass) were recorded, analyzed, and resynthesized using an analysis–synthesis method (Aramaki & Kronland-Martinet, 2006) to generate realistic synthetic sounds. Second, sound continua were created that simulate progressive transitions between sounds from different materials (i.e., wood–metal, wood–glass, and metal–glass continua). Although sounds at extreme positions on the continua were synthesized to be as similar as possible to natural sounds, sounds at intermediate positions were synthesized by interpolating the acoustic parameters characterizing sounds at extreme positions. They were consequently ambiguous (e.g., neither wood nor metal). Sounds from the different continua were randomly presented, and participants were asked to categorize each sound as wood, metal, or glass. If sounds at extreme positions of the continua are indeed perceived as typical exemplars of their respective categories, they should be categorized faster and with lower error rate than sounds at intermediate positions on the continua.

Little is known about the neural basis of impact sound perception. To investigate this issue, we also recorded ERPs while participants performed the categorization task. Results of studies on the categorization of nonspeech stimuli have shown that the amplitude of the N200–P300 complex, which typically follows the N100–P200 exogenous complex, is influenced by the difficulty of the categorization task: The N200 component is larger and the P300 component is smaller to stimuli that are more difficult to categorize (Donchin & Coles, 1988; Ritter, Simson, & Vaughan, 1983; Donchin, 1981; Ritter, Simson, Vaughan, & Friedman, 1979; Simson, Vaughan, & Ritter, 1977). Thus, if ambiguous sounds are more difficult to categorize than typical sounds because they are composed of hybrid acoustic features, they should elicit larger N200 and smaller P300 components than typical sounds.

Methods

Subjects

A total of 25 participants were tested in this experiment that lasted for about 1 hour. Three participants were excluded from final data analysis because of the high number of trials contaminated by ocular and muscular artifacts. The remaining 22 participants (11 women and 11 men; 19–35 years old) were all right-handed, were nonmusicians (no formal musical training), had normal audition, and had no known neurological disorders. They all gave written consent to participate to the experiment and were paid for their participation.

Stimuli

We first recorded sounds by impacting everyday life objects made of different materials (i.e., wooden beams, metallic plates, and various glass bowls) to insure the generation of realistic familiar sounds. Then, we used a simplified version of the model described in Aramaki and Kronland-Martinet (2006) on the basis of an additive synthesis technique to resynthesize these recorded sounds (44.1-kHz sampling frequency). From a physical point of view, because the vibrations of an impacted object (under free oscillations) can generally be written as a sum of exponentially damped sinusoids, the recorded sounds are considered to be well described by

$$s(t) = \theta(t) \sum_{m=1}^{M} A_m \sin(2\pi f_m t) e^{-\alpha_m t} \qquad (1)$$

where $\theta(t)$ is the Heaviside unit step function, M is the number of sinusoidal components, and the parameters A_m, f_m, and α_m are the amplitude, frequency, and damping coefficient of the mth component, respectively. The synthesis parameters of this model (i.e., M, A_m, f_m, and α_m) were estimated from the analysis of the recorded sounds (examples of analysis–synthesis processes can be found in Kronland-Martinet, Guillemain, & Ystad, 1997). In practice, many sounds from each material category were recorded and resynthesized and the five most representative sounds per category (as judged by seven listeners) were selected for the current study.

All sounds were tuned to the same chroma (note C that was closest to the original pitch) and were equalized in loudness by gain adjustments. Averaged sound duration was 744 msec for wood, 1667 msec for metal, and 901 msec for glass category. Because the damping is frequency dependent (see Equation 1), the damping coefficient of each tuned component was modified according to a damping law estimated from the original sound (Aramaki, Baillères, Brancheriau, Kronland-Martinet, & Ystad, 2007).

A total of 15 sound continua were created as progressive transitions between two material categories (i.e., 5 different continua for each transition: wood–metal, wood–glass, and glass–metal). Each continuum comprised 20 sounds that were generated using additive synthesis (see Equation 1). Each sound of the continuum was obtained by combining the spectral components of the two extreme sounds and by varying the amplitude and damping coefficients. Amplitude variations were obtained by applying a cross-fade technique between the two extreme sounds. Damping coefficients were estimated by a hybrid damping law resulting from the interpolation between the damping laws of the two extreme sounds. Note that this manipulation allowed creating hybrid sounds that differed from a simple mix between the two extreme sounds because at each step, the spectral components are damped following a same hybrid damping law. All sounds are available at http://www.lma.cnrs-mrs.fr/~kronland/Categorization/sounds.html.

Procedure

A total of 300 sounds were presented in a random order within five blocks of 60 sounds through one loudspeaker (Tannoy S800) located 1 m in front of the participants. They were asked to listen to each sound and to categorize it as wood, metal, or glass as quickly as possible by pressing one button on a three-button response box. The association between response buttons and sounds was balanced across participants. The experiment was conducted in a faradized room. A row of XXXX was presented on the screen 2500 msec after sound onset for 1500 msec to give participants time to blink, and the next sound was then presented after a 500-msec delay.

Recording ERPs

The EEG was recorded continuously from 32 Biosemi Pin-type active electrodes (Amsterdam University) mounted on an elastic headcap and located at standard left and right hemisphere positions over frontal, central, parietal, occipital, and temporal areas (international extended 10/20 system; Jasper, 1958): Fz, Cz, Pz, Oz, Fp1, Fp2, AF3, AF4, F7, F8, F3, F4, Fc5, Fc6, Fc1, Fc2, T7, T8, C3, C4, Cp1, Cp2, Cp5, Cp6, P3, P4, PO3, PO4, P7, P8, O1, and O2. Moreover, to detect horizontal eye movements and blinks, the EOG was recorded from flat-type active electrodes placed 1 cm to the left and right of the external canthi and from an electrode beneath the right eye. Two additional electrodes were placed on the left and right mastoids. EEG was recorded at 512-Hz sampling frequency using Biosemi amplifier. The EEG was rereferenced off-line to the algebraic average of the left and right mastoids and filtered with a band-pass of 0–40 Hz.

Data were analyzed using Brain Vision Analyzer software (Brain Products, Munich), segmented in single trials of 2500 msec starting 200 msec before the onset of the sound and averaged as a function of the type of sound (i.e., typical vs. ambiguous).

Results

Behavioral Data

Participants' responses and RTs were collected for each sound and were averaged across participants. Sounds categorized within a category (wood, metal, or glass) by more than 70% of the participants were considered as typical sounds; sounds categorized within a category by less than 70% of the participants were considered as ambiguous sounds. As can be seen in Figure 1 (top), participants' responses are consistent with the position of the sound on the continua so that typical sounds are located at extreme positions and ambiguous sounds at intermediate positions on the continua.

RTs to typical and ambiguous sounds were submitted to repeated measures ANOVAs (for this and following statistical analyses, effects were considered significant if the p value was equal to or less than .05) that included Type of Sounds (typical vs. ambiguous) and Continua (wood–metal, wood–glass, and glass–metal) as within-subject factors. For typical sounds, only RTs associated to correct responses were taken into account. As shown in Figure 1 (bottom), RTs to typical sounds (984 msec) were shorter than RTs to ambiguous sounds (1165 msec), $F(1, 21) = 74.00, p < .001$. Moreover, the Type of Sounds × Continua interaction was significant, $F(2, 42) = 22.24, p < .001$. Results of post hoc comparisons (Tukey tests) showed that although RTs were shorter for typical than for ambiguous sounds for each continuum ($p < .01$), this difference was larger for the wood–glass continuum (281 msec) than for the wood–metal (184 msec) and glass–metal (79 msec) continua.

Electrophysiological Data

Separate ANOVAs were conducted for midline and lateral electrodes. Type of sounds[1] (typical vs. ambiguous) and Electrodes (Fz, Cz, Pz) were used as factors for midline analyses. Type of Sounds, Hemispheres (left vs. right), ROIs (fronto-central R1, centro-temporal R2, and centro-parietal R3), and Electrodes (three for each ROI: [AF3, F3, FC5]/[AF4, F4, FC6]; [T7, C3, CP5]/[T8, C4, CP6]; and [P7, P3, CP1]/[P8, P4, CP2]) were included for lateral analyses. On the basis of visual inspection of the ERP traces (Figure 2) and results of successive analyses in 50-msec latency windows, the following time windows were chosen for statistical analysis:[2] 0–250 msec (P100–N100–P200),

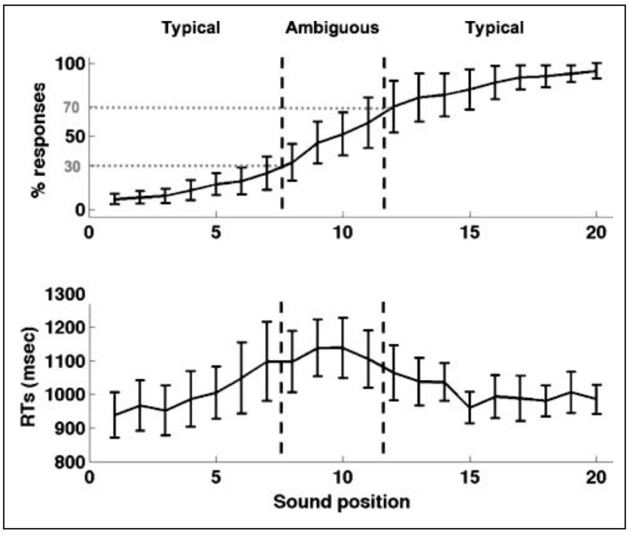


Figure 1. Categorization function (top) and mean RTs in millisecond (bottom) averaged across the 15 continua as a function of sound position (from 1 to 20). Standard deviations are indicated at each sound position. The categorization function represents the percentage of responses in the material category corresponding to the right extreme of the continuum. The vertical dotted gray lines delimit the zones of typical (extremes positions) and of ambiguous (intermediate positions) sounds.

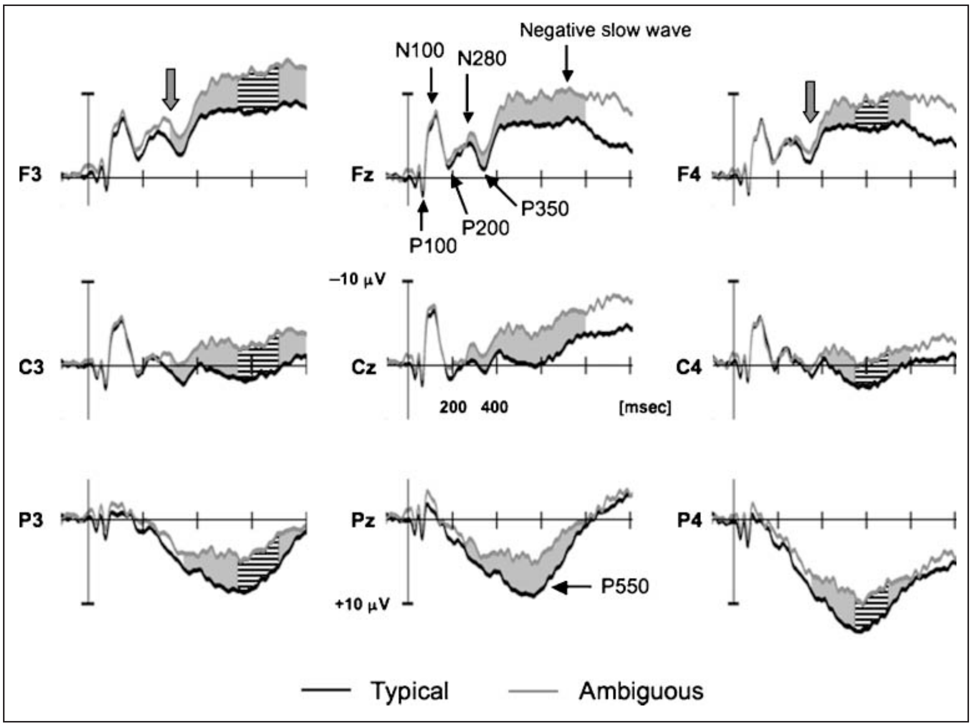


Figure 2. ERPs to typical (black line) and ambiguous sounds (gray line) at midline and at selected lateral electrodes (the most representative electrodes for each ROI). For this and following figures, the amplitude of the effects is represented on the ordinate (in µV; negativity is up). The time from sound onset is on the abscissa (in msec). The gray zones indicate the latency ranges in which differences between typical and ambiguous sounds were significant. These differences started earlier (vertical gray arrows) and were larger (striped zones) over the left than the right hemisphere.

250–400 msec (N280-P350), and 400–800 msec (negative slow wave [NSW] and P550).

Figure 2 shows the ERPs for midline and selected lateral electrodes (the most representative electrode for each ROI). Both typical and ambiguous sounds elicited similar P100, N100, and P200 components at midline and lateral electrodes (no significant effect in the 0- to 250-msec latency band). The ERPs in the two conditions then start to diverge with larger N280 and smaller P350 components in the 250- to 400-msec latency band for ambiguous (−0.89 µV) than for typical sounds (0.54 µV) at midline electrodes, $F(1, 21) = 4.19, p < .05$. Results of fine-grained analyses in successive 50-msec latency bands revealed that these differences started earlier over fronto-central regions of the left—from 300 msec at F3 ($p < .001$) and C3 ($p < .01$) electrodes; Type of Sounds × Hemispheres × ROI × Electrodes interaction: $F(4, 84) = 2.68, p < .05$—than of the right hemisphere—from 350 msec at F4 electrode ($p < .05$); Type of Sounds × Hemispheres × ROI × Electrodes interaction: $F(4, 84) = 2.32, p = .06$ (see Figure 2).

In the 400- to 800-msec latency range, the main effect of Type of Sounds was still significant: midline, $F(1, 21) = 23.60, p < .001$; lateral, $F(1, 21) = 17.91, p < .001$. Although typical sounds were associated with larger positivity (P550) than ambiguous sounds over parietal regions, ambiguous sounds elicited larger NSW than typical sounds over fronto-central regions: Type of Sounds × ROI interaction, $F(2, 42) = 4.11, p < .05$. These differences were larger over the left (3.4 µV) than the right (2.78 µV) hemisphere in the 550- to 700-msec latency band[2]: Type of sounds × Hemispheres interaction, $F(1, 21) = 4.60, p < .05$ (see striped zones in Figure 2).

Discussion

Analysis of behavioral data showed that sounds categorized within a material category by less than 70% of the participants (ambiguous sounds) were associated with slower RTs than sounds that were categorized within a category by more than 70% of the participants (typical sounds). This was found for each continuum. Thus, as hypothesized, ambiguous sounds were more difficult to categorize than typical sounds. This result is in line with previous findings in the literature showing slower RTs for nonmeaningful than for meaningful sounds (e.g., Cummings et al., 2006). The differences between typical and ambiguous sounds were smaller in the wood–metal and glass–metal continua than in the wood–glass continuum. This is interesting from an acoustic perspective because metal sounds typically present higher spectral complexity (related to the density and repartition of spectral components) than both wood and glass sounds that show closer sound properties. Thus, ambiguous sounds in wood–metal and glass–metal continua were easier to categorize than those in the wood–glass continuum and the ambiguity effect was smaller.

Electrophysiological data showed that ambiguous sounds elicited more negative ERPs (a negative component, N280, followed by an NSW) than typical sounds over fronto-central regions. By contrast, typical sounds elicited more positive ERPs (P350 and P550 components) than ambiguous sounds over frontal and parietal regions. These findings were expected on the basis of previous results in categorization tasks showing that the amplitude of the N200 component is larger and the amplitude of the P300 component is smaller to stimuli that are more difficult to categorize (Donchin & Coles, 1988; Ritter et al., 1983; Duncan-Johnson & Donchin, 1982; Donchin, 1981; Kutas, McCarthy, & Donchin, 1977). Moreover, in line with the long duration of the RTs (around 1 sec), the long latency of the positive component (P550) is taken to reflect the difficulty of the categorization task (participants were required to categorize sounds in one of three possible categories) and the relatively long duration of the sounds (860 msec, on average, over the three categories). Thus, both behavioral and ERP data showed that we were able to create ambiguous sounds that were more difficult to categorize than typical sounds.

The differences between ambiguous and typical sounds started earlier over the left (300 msec) than the right (350 msec) hemisphere and were also larger over the left than right hemisphere in the 550- to 700-msec latency band (see striped zones in Figure 2). This scalp distribution is similar to the left-hemisphere distribution reported for sounds by Van Petten and Rheinfelder (1995). Moreover, as found by these authors, a long-lasting NSW developed over frontal sites that lasted for the entire recording period. Late NSW are typically interpreted as reflecting processes linked with the maintenance of stimuli in working memory, expectancy (Walter, Cooper, Aldridge, McCallum, & Winter, 1964), and attention (King & Kutas, 1995). In the present experiment, the NSW may indeed reflect expectancy processes because a row of XXXX followed sound offset, but it may also reflect sound duration processing (as the "sustained potential" reported by Alain, Schuler, & McDonald, 2002) and categorization difficulty because this fronto-central negativity was larger for ambiguous than for typical sounds. In particular, as it has been proposed for linguistic stimuli (see Kutas & Federmeier, 2000), this larger negativity may reflect the difficulty of accessing information from long-term memory.

Finally, it should be noted that no significant differences were found on the P100 and N100 components. These components are known to be sensitive to sound onset (e.g., attack time) and temporal envelope (for a review, see Kuriki, Kanda, & Hirata, 2006; Shahin, Roberts, Pantev, Trainor, & Ross, 2005; Shahin, Bosnyak, Trainor, & Roberts, 2003; Hyde, 1997). However, because differences in attack time between typical and ambiguous sounds were 0.1 msec, on average, they were consequently not perceptible as this value is below the temporal resolution of the human hearing system (Gordon, 1987), thereby explaining the lack of differences in the ERPs.

EXPERIMENT 2: CONCEPTUAL AND SEMANTIC PRIMING

Results of Experiment 1 showed that we were able to create typical and ambiguous sounds. The goal of Experiment 2 was to use these sounds in a priming design to address the three aims described in the introduction: (1) test for conceptual priming between pairs of nonlinguistic sounds, (2) use only one homogeneous class of sounds (impact sounds), and (3) directly compare conceptual priming for nonlinguistic stimuli on one side and for linguistic stimuli on the other side. To achieve these aims, it was important to use the same task with both types of sounds. Thus, participants were asked to decide whether the target belonged to the same or to a different category than the prime. For the linguistic sounds and based on the design used by Holcomb and Neville (1990), primes were always words, and targets were words, pseudowords, or nonwords (i.e., words played backward). To use similar design and experimental conditions with nonlinguistic sounds, primes were always impact sounds, and targets were typical impact sounds from the same category as the prime, ambiguous sounds and typical impact sounds from a different category than the prime.[3]

On the basis of previous results in the literature with linguistic stimuli (Holcomb & Neville, 1990; Bentin, McCarthy, & Wood, 1985) and results of Experiment 1 with nonlinguistic stimuli, we hypothesized that pseudowords and ambiguous sounds should be more difficult to categorize (i.e., higher error rates and slower RTs) than stimuli from the other two categories. Moreover, as reported by Holcomb and Neville (1990) and in previous studies (for a review, see Kutas, Van Petten, & Kluender, 2006), pseudowords should also elicit larger N400 than words. Holcomb and Neville (1990) argued that "Perhaps this was because their word-like characteristics also produce lexical activation, but because no complete match was achieved, the amount of activation produced was greater and more prolonged" (p. 306). More generally, this result has been taken to reflect the (unsuccessful) search for meaning of orthographically and phonologically legal constructions that nevertheless have no meaning (see Kutas & Federmeier, 2000). However, the N400 to pseudowords may also reflect their lower familiarity than words and their ambiguous nature: They are word-like at the orthographic and phonological levels but are not real words at the semantic level. In such case, ambiguous sounds that share acoustic properties with typical sounds of a material category but nevertheless are not typical exemplars of any categories may also elicit N400-like components. It was therefore of interest to determine whether ambiguous sounds would be processed as pseudowords and elicit N400-like components or would rather elicit increased N280 and NSW as found in Experiment 1. Finally, nonwords (i.e., words played backward) should elicit larger P300 components than words, as reported by Holcomb and Neville (1990). Indeed, although words played backward keep the main attributes of vocal sounds (i.e., the formantic structure of the spectrum due to the resonance of the vocal tract), they should readily be perceived as belonging to a different category than the word prime. Similarly, if typical sounds from a different category than the prime are easily categorized as such, they should also elicit larger P300 components than typical sounds from the same category than the prime.

Methods

Subjects

A total of 19 students (8 women and 11 men; 24 years old, on average) participated in this experiment that lasted for about 1 hour 30 min. None had participated in Experiment 1. They were right-handed, nonmusicians (no formal musical training), French native speakers with no known neurological disorders. They all gave written consent to participate in the experiment and were paid for their participation.

Stimuli

A total of 120 pairs of nonlinguistic stimuli were presented. Primes were always typical sounds from a given material category (i.e., wood, metal, or glass), and targets were either sounds from the same category as the prime (Same condition, 30 pairs), ambiguous sounds (Ambiguous condition, 60 pairs), or sounds from a different category than the prime (Different condition, 30 pairs). The number of ambiguous pairs was twice the number of pairs in the Same and Different conditions to balance the number of Yes and No responses. (On the basis of the results of Experiment 1, we expected participants to give as many Yes as No responses to ambiguous targets.) The averaged duration of nonlinguistic stimuli was 788 msec.

A total of 180 pairs of linguistic sounds were presented. Primes were always French spoken words and targets were spoken words (Same condition, 90 pairs), pseudowords (Ambiguous condition, 45 pairs), or nonwords (Different condition, 45 pairs). Word targets were bisyllabic nouns. Pseudowords were constructed by modifying one vowel from word targets (e.g., *b$\underline{o}$teau* from *b$\underline{a}$teau*). Nonwords were words played backward. The averaged duration of linguistic stimuli was 550 msec.

Procedure

Participants were asked to listen to each pair of stimuli and to determine whether the prime and the target belonged to the same category by pressing one of two response buttons. Nonlinguistic and linguistic stimuli were presented in two separate sessions 10 min apart with the linguistic session always presented after the nonlinguistic session. In each session, pairs of stimuli belonging to the three experimental conditions were randomly presented within three blocks of 40 trials for nonlinguistic pairs and three blocks of 60 trials for linguistic pairs (less nonlinguistic stimuli were presented within a block because sounds were longer in duration than linguistic stimuli).

To balance the number of Yes and No responses, each block of nonlinguistic stimuli comprised 10 same (yes), 20 ambiguous (yes/no), and 10 different pairs (no). Each block of linguistic stimuli comprised 30 Same (yes), 15 Ambiguous (no), and 15 Different (no) pairs. The order of block presentations within the nonlinguistic and linguistic sessions and the association between responses (Yes/No) and buttons (left/right) were balanced across participants.

For both nonlinguistic and linguistic pairs, targets followed prime offset with a 20-msec interstimulus interval. A row of XXXX was presented on the screen 2000 msec after target onset for 2000 msec to give participants time to blink. The prime of the next pair was then presented after a 1000-msec delay.

Recording ERPs

EEG was continuously recorded using the same procedure as in Experiment 1 and later segmented in single trials of 2200 msec starting 200 msec before target onset. Data were analyzed using the Brain Vision Analyzer software (Brain Products, Munich).

Nonlinguistic Sounds

Results

Behavioral data. For ambiguous sounds, there are no correct or incorrect responses because they can be associated to yes or to no responses. Thus, on the basis of the participants' responses, ANOVAs included Category as a factor with four conditions: Same, Ambiguous/Yes, Ambiguous/No, and Different targets. Results revealed a main effect of Category, $F(3, 54) = 111.71$, $p < .001$: Same and Different targets were associated with low error rates (6% and 4%, respectively) and did not differ from each other ($p = .92$). They differed from Ambiguous/Yes (46%; $p < .001$) and Ambiguous/No targets (53%; $p < .001$) that did not differ from each other ($p = .15$). Mean RTs were not significantly different ($p = .09$: 917 msec for Same, 900 msec for Ambiguous/Yes, 863 msec for Ambiguous/No, and 892 msec for Different targets).

Electrophysiological data. Two separate ANOVAs were conducted for midline and lateral electrodes. Category (Same, Ambiguous,[4] Different) and Electrodes (Fz, Cz, Pz) were included as factors for midline analyses. Category, Hemispheres (left vs. right), ROIs (fronto-central R1, centro-temporal R2, and centro-parietal R3), and Electrodes (3 for each ROI: [F7, F3, FC1]/[F8, F4, FC2]; [FC5, T7, C3]/[FC6, T8, C4]; and [CP1, CP5, P3]/[CP6, CP2, P4]) were included for lateral analyses. On the basis of visual inspection and results of successive analyses in 50-msec latency windows, time windows chosen for the statistical analysis were 0–150, 150–350, 350–450, 450–550, and 550–700 msec. For Same and Different targets, only correct responses were taken into account. Results are reported in Table 1.

Figure 3 (top) illustrates ERPs to nonlinguistic targets. In all conditions, sounds elicited P100, N100, P200, and

Table 1. Nonlinguistic Targets

(I)	*Factors*	*df*	*0–150 msec*	*150–350 msec*	*350–450 msec*	*450–550 msec*	*550–700 msec*
Midline	C	2,36	–	–	14.92***	17.13***	12.26***
	C × E	4,72	–	2.60*	2.50*	–	–
Lateral	C	2,36	–	–	15.70***	19.04***	12.10***
	C × ROI	4,72	–	–	2.44*	2.66*	4.89**

(II)	*150–350 msec*			*350–450 msec*		
C × E	*Fz*	*Cz*	*Pz*	*Fz*	*Cz*	*Pz*
A − S	–	–	–	−1.74**	−2.24	−2.52
A − D	−1.42*	–	–	−4.23	−3.75	−3.05
S − D	−1.9	–	–	−2.49	−1.51*	–

(III)	*350–450 msec*			*450–550 msec*			*550–700 msec*		
C × ROI	*R1*	*R2*	*R3*	*R1*	*R2*	*R3*	*R1*	*R2*	*R3*
A − S	−1.57	−1.8	−2.55	−1.54	−1.86	−2.37	−1.12**	−1.43	−1.44
A − D	−2.82	−2.51	−3.04	−3.17	−2.97	−4.09	−2.07	−2.04	−3.4
S − D	−1.25	–	–	−1.63	−1.11**	−1.72	−0.95*	–	−1.96

(I) *F* statistics for the main effect of Category (C), for the Category by Electrodes (C × E) interaction, and for the Category by ROI (C × ROI) interaction in the latency ranges chosen for analyses. (II and III) Mean amplitude differences (in μV) between Same (S), Ambiguous (A), and Different (D) conditions when the C × E and C × ROI interactions were significant. The reported difference values were always significant at $p < .001$ (results of post hoc tests) except when indicated by $p < .05$ or $p < .01$.

*$p < .05$.

**$p < .01$.

***$p < .001$.

N280 components followed by large negative components over fronto-central regions and P550 components over parietal regions. No significant effects were found in the latency range 0–150 msec (see Table 1-I). In the 150- to 350-msec latency range, the main effect of Category was not significant but the Category × Electrodes interaction was significant: Ambiguous and Same targets elicited larger N280 than Different targets at Fz (see Table 1-II). In the 350- to 450-msec latency range, the main effect of Category was significant with larger negativity to Ambiguous than to both Same and Different targets that did not differ from each other except over fronto-central region (see Table 1-II and 1-III). In the 450- to 550-msec latency range, the three conditions significantly differed from each other with the negativity being largest for Ambiguous, intermediate for Same, and the positivity largest for Different targets, with largest differences over centro-parietal regions (i.e., over R3 in Table 1-III). Finally, in the 550- to 700-msec latency range, at midline electrodes, Different targets elicited larger positivity (−0.01 μV) than Same (−2.05 μV) and Ambiguous targets (−3.28 μV) that did not differ from each other. By contrast, at lateral electrodes, the three conditions still differed significantly from each other with largest differences over the centro-parietal region.

Discussion

As hypothesized, the error rate for Same and Different targets was very low (6% and 4%, respectively), which shows that typical sounds were easily categorized as belonging to the same or to a different impact sound category than the prime. By contrast and as expected, on the basis of the results of Experiment 1, ambiguous targets were more difficult to categorize and were categorized as often as belonging to the same (46%) or to a different category (54%) from the prime. This clearly confirms the ambiguous nature of these sounds. The lack of effects on RTs may result from the relatively long duration of the prime sounds (788 msec, on average). Priming effects generally are short lived (Meyer & Schvaneveldt, 1971) and may consequently have vanished by the time the target sound was presented.

Regarding the ERPs, all targets elicited a P100–N100–P200 complex that, as expected (see Discussion of Experiment 1), did not differ between Same, Ambiguous,

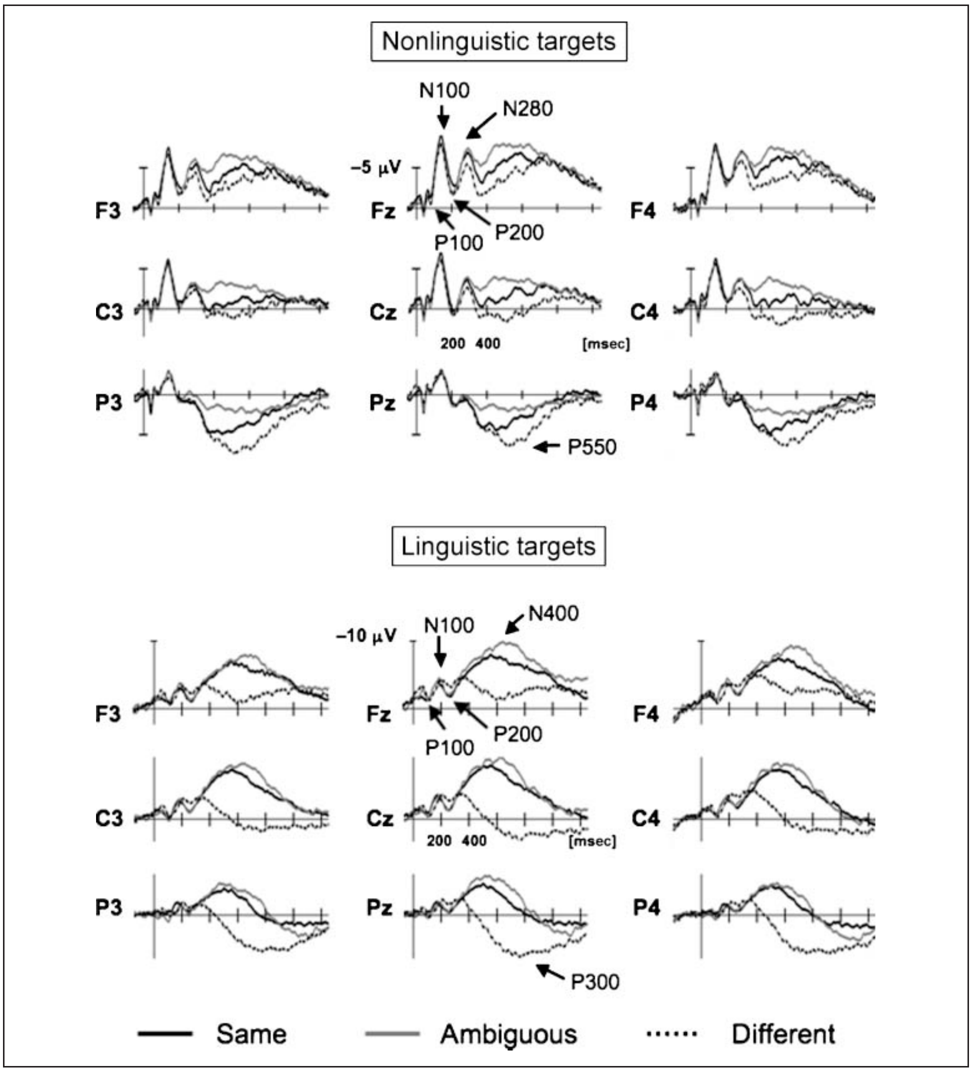


Figure 3. ERPs to Same (solid line), Ambiguous (gray line), and Different (dashed line) targets at midline and at selected lateral electrodes (the most representative electrodes for each ROI) for nonlinguistic (top) and linguistic (bottom) stimuli.

and Different target sounds. An N280 component was also elicited as in Experiment 1. Its amplitude was larger for Same and Ambiguous sounds than for Different sounds over frontal regions. However, the ERPs were morphologically different in Experiments 1 and 2. Although an NSW followed the N280 in Experiment 1 (and lasted until the end of the recording period), a temporally (between 350 and 700 msec) and spatially (fronto-central) localized negativity followed the N280 in Experiment 2. Fine-grained analyses allowed to specify the spatiotemporal dynamics of the effects. First, between 350 and 450 msec, the amplitude of this negative component was largest over fronto-central sites for Ambiguous targets, intermediate for Same targets, and smallest for Different targets. Then, between 450 and 550 msec, typical sounds from a different category than the prime elicited large P300 components over parietal sites thereby reflecting the fact that they were easily categorized as different (4% errors; Holcomb & Neville, 1990; Kutas et al., 1977). Because the same stimuli were used in Experiments 1 and 2, these differences are clearly linked with the task at hand (i.e., in Experiment 1, isolated impact sounds were to be categorized in one of three categories, whereas in Experiment 2, target sounds were compared with a prime). Thus, and as typically shown by fMRI data, these results demonstrate the strong influence of task demands on stimulus processing (e.g., Thierry, Giraud, & Price, 2003).

In Experiment 2, we used a priming design to be able to compare results with previous ones in the literature, and we presented two sounds and no words to reduce the use of linguistic strategies that, as described in the introduction, may have influenced previous results (Orgs et al., 2006, 2007, 2008; Cummings et al., 2006; Plante et al., 2000; Van Petten & Rheinfelder, 1995; and, to a lesser extent, Schön et al., 2009). The finding that a negative component developed in the 350- to 700-msec latency band with largest amplitude to Ambiguous sounds is in line with these previous studies and shows that conceptual priming can occur within sound–sound pairs. Moreover, this result was found when using the homogeneous class of impact sounds. However, before considering the implications of these results for conceptual priming, it is important to examine results obtained for linguistic targets preceded by linguistic primes.

Linguistic Sounds

Results

Behavioral data. The main effect of Category (word [W], pseudoword [PW], nonword [NW]; within-subject factor) was significant, $F(2, 36) = 48.15$, $p < .001$: The error rate was higher for PW (13%) than for W (4.9%; $p < .001$) and NW (2.3%; $p < .001$). RTs were not significantly different ($p = .61$) for PW (1067 msec), W (1057 msec), and NW (1054 msec).

Electrophysiological data. Similar ANOVAs were conducted as for nonlinguistic sounds. Statistical analysis was conducted in the 0–150, 150–350, 350–600, 600–750, and 750–1100 msec latency ranges. Only correct responses were taken into account. Results of statistical analyses are reported in Table 2.

Figure 3 (bottom) illustrates ERPs to linguistic targets. No significant differences were found in the latency ranges 0–150 and 150–350 msec either at midline or at lateral

Table 2. Linguistic Targets

(I)	*Factors*	*df*	*0–150 msec*	*150–350 msec*	*350–600 msec*	*600–750 msec*	*750–1100 msec*
Midline	C	2,36	–	–	41.32***	90.52***	40.81***
	C × E	4,72	–	–	5.00**	8.97***	11.24***
Lateral	C	2,36	–	–	36.71***	85.53***	54.50***
	C × ROI	4,72	–	–	4.80**	7.41***	7.51***
	C × ROI × H	4,72	–	–	4.93**	2.87*	3.16*

(II)	*350–600 msec*			*600–750 msec*			*750–1100 msec*		
C × E	*Fz*	*Cz*	*Pz*	*Fz*	*Cz*	*Pz*	*Fz*	*Cz*	*Pz*
P − W	–	–	–	−2.57	−2.1	−2.22	−1.22*	–	–
P − N	−4.48	−6.33	−5.68	−7.79	−10.47	−9.58	−3.34	−5.61	−4.61
W − N	−3.53	−5.67	−4.62	−5.22	−8.37	−7.36	−2.12	−5.4	−4.6

(III)	*350–600 msec*			*600–750 msec*			*750–1100 msec*		
C × ROI	*R1*	*R2*	*R3*	*R1*	*R2*	*R3*	*R1*	*R2*	*R3*
P − W	–	–	–	−2.03	−1.57	−1.45	−0.95*	–	–
P − N	−3.03	−4.07	−4.37	−6.08	−7.11	−7.39	−2.58	−3.68	−3.47
W − N	−2.49	−3.65	−3.91	−4.05	−5.54	−5.94	−1.63	−3.25	−3.71

(IV)	*350–600 msec*		*600–750 msec*		*750–1100 msec*	
C × ROI × H	*L*	*R*	*L*	*R*	*L*	*R*
R1	–	−0.68*	−1.72	−2.33	–	−1.43
R2	–	–	−1.52	−1.61	–	–
R3	–	–	−1.61	−1.29	–	–

(I) *F* statistics for the main effect of Category (C), for the Category by Electrodes (C × E) interaction, for the Category by Regions of Interest (C × ROI) interaction, and for the Category by Regions of Interest by Hemispheres (C × ROI × H) interactions in the latency ranges of interest. (II and III) Mean amplitude differences (in μV) between Words (W), Pseudowords (P), and Nonwords (N) conditions for C × E and C × ROI interactions when effects were significant. The reported difference values were always significant at $p < .001$ (results of post hoc tests) except when indicated by $p < .05$. (IV) Mean amplitude differences P − W (in μV) for C × ROI × H interaction. The reported difference values were always significant at $p < .001$ (results of post hoc tests) except when indicated by $p < .05$.

*$p < .05$.

**$p < .01$.

***$p < .001$.

electrodes, but a main effect of Category was found in the 350–600, 600–750, and 750–1100 msec latency ranges at both midline and lateral electrodes (see Table 2-I). In these three latency ranges, NW always elicited larger positivity than both PW and W with largest differences at Cz and over centro-parietal regions (Table 2-II and 2-III). In addition, between 600 and 1100 msec, PW elicited larger negativity than W over right fronto-central regions (Table 2-IV).

Discussion

Behavioral data, showing higher error rate for PW than for both W and NW, are in line with previous results (e.g., Holcomb & Neville, 1990; Bentin et al., 1985). However, no effect was found on RTs, which again may reflect the relatively long duration of stimuli and of RTs (over 1 sec, on average) together with short-lived priming effects (Meyer & Schvaneveldt, 1971). As expected, on the basis of Holcomb and Neville's (1990) results, PW produced larger N400 components than W over anterior sites. Moreover, this N400 effect was larger over the right than the left hemisphere. This "paradoxical lateralization" (Plante et al., 2000, p. 1680) is consistent with previous results showing right-greater-than-left asymmetry of the N400 effect (Kutas et al., 1988; Kutas & Hillyard, 1982). Finally, the rather long latency of this N400 effect is also consistent with the results of Holcomb and Neville (1990), showing that the N400 effect starts earlier and lasts longer in the auditory than in the visual modality. It may also reflect the difficulty to categorize PW that were very similar to words (they were constructed by replacing only one vowel from an existing word). By contrast, NW (i.e., words played backward) was easy to categorize as different from the prime words and elicited a large P300 component with a posterior scalp distribution (Holcomb & Neville, 1990).

Nonlinguistic versus Linguistic Sounds

Because the same design was used for both nonlinguistic and linguistic sounds within the same group of participants, conceptual and semantic priming effects were directly compared by including Stimulus (nonlinguistic vs. linguistic) as a factor.

ANOVAs were conducted in the 350- to 800-msec time window, where significant differences were found for both nonlinguistic and linguistic sounds. Results of statistical analyses are reported in Table 3. The main effect of Stimulus was significant: ERPs to linguistic stimuli were overall more negative than to nonlinguistic stimuli (Table 3-II). Moreover, the main effect of Category was significant with largest N400 to Ambiguous (ambiguous impact sounds and PW), intermediate to Same (same impact sounds and W), and largest positivity to Different targets (different impact sounds and NW) (Table 3-III). Finally, the Stimulus × Category interaction was significant. Although the difference between Ambiguous and Same targets was similar for both linguistic and nonlinguistic stimuli, the difference between Different and Same targets was significantly larger for linguistic than for nonlinguistic stimuli (Table 3-IV and Figure 4).

Table 3. Nonlinguistic versus Linguistic Targets

(I)	*Factors*	*df*	*F*
Midline	Stimulus	1,18	17.48***
	C	2,36	68.57***
	Stimulus × C	2,36	14.43***
Lateral	Stimulus	1,18	24.18***
	C	2,36	65.95***
	Stimulus × C	2,36	12.69***

(II)	*Midline*	*Lateral*
Nonlinguistic	−1.52	−0.42
Linguistic	−4.00	−2.98

(III)	*Midline*	*Lateral*
S	−3.51	−2.19
A	−4.95	−3.35
D	0.17	0.43

(IV)		*Midline*	*Lateral*
A − S	Nonlinguistic	1.43	1.31
	Linguistic	1.46	1.01
D − S	Nonlinguistic	1.80	1.17
	Linguistic	5.57	4.06

(I) *F* statistics for the main effect of Stimulus and Category (C) and for the Stimulus by Category (Stimulus × C) interaction in the 350- to 800-msec latency range. (II) Mean amplitude (in μV) of the main effect of Stimulus. (III) Mean amplitude (in μV) of the main effect of Category: Same (S), Ambiguous (A), and Different (D) conditions. (IV) Mean amplitude differences A − S and D − S (in μV) for Nonlinguistic and Linguistic stimuli.

$^{***}p < .001$.

GENERAL DISCUSSION

Results of the general ANOVA highlighted clear similarities between conceptual priming for nonlinguistic and linguistic sounds. In both cases, behavioral data showed higher error rates in the Ambiguous than in the Same and Different conditions with no effects on RTs. This ambiguity effect most likely reflects the difficulty to correctly categorize Ambiguous targets as different because they are similar to the prime (e.g., orthographic and phonologic similarity for PW and acoustic proximity for impact sounds). Several studies using priming designs showed higher error rates

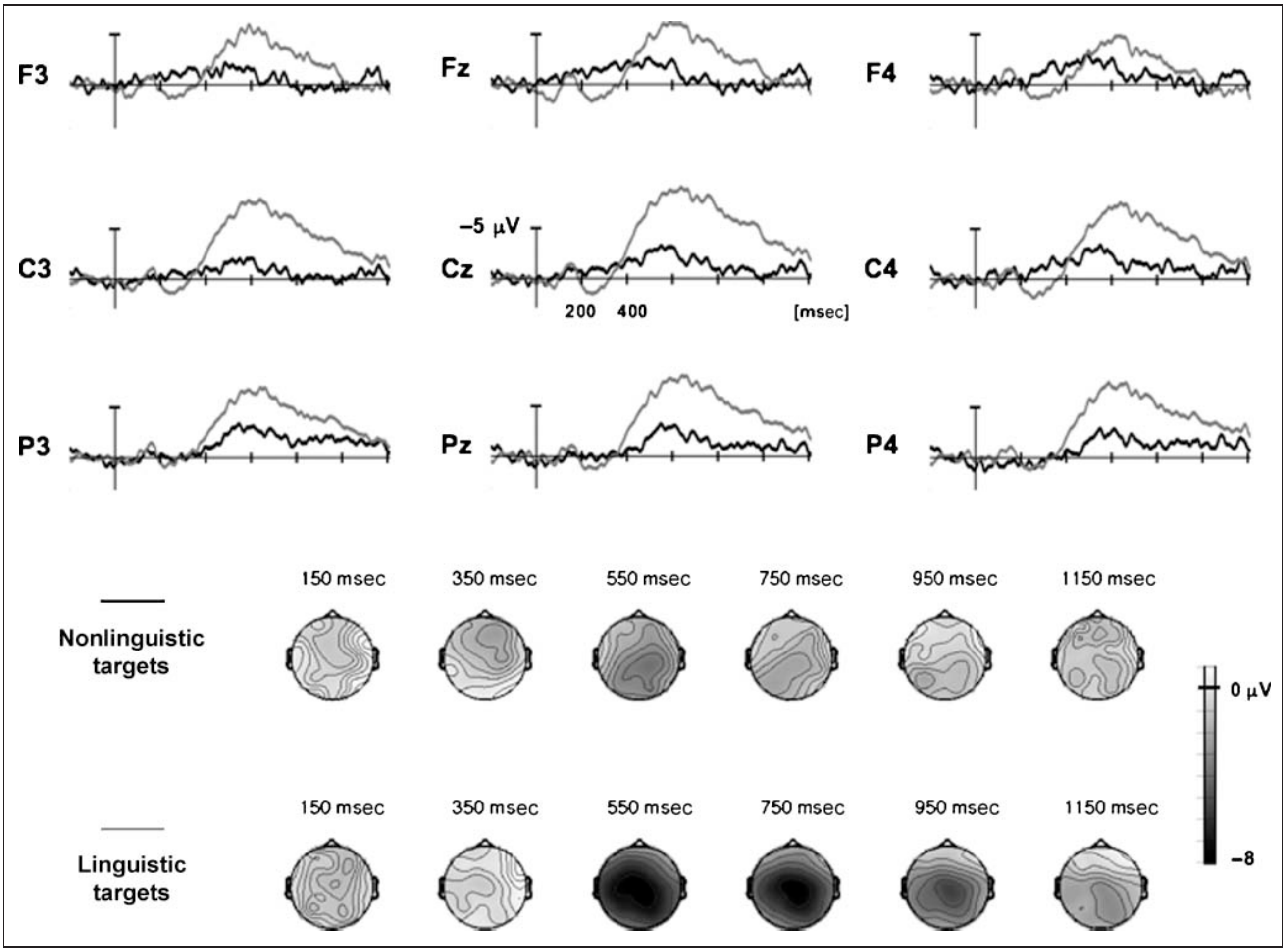


Figure 4. Same-minus-Different Difference Waves. ERPs to nonlinguistic (black line) and linguistic (gray line) targets at midline and at selected lateral electrodes (the most representative electrodes for each ROI). Temporal dynamics of the scalp distribution of the effects from 150 to 1150 msec for nonlinguistic and linguistic targets.

for PW than for W (e.g., Holcomb & Neville, 1990) as well as for unrelated than for related words (e.g., Bentin et al., 1985; Boddy, 1981). By contrast, results differ in some studies using nonlinguistic sounds. For instance, Orgs et al. (2006, 2008) found higher error rates in related compared with unrelated pairs. They explained this result by the greater ambiguity of environmental sounds due to causal uncertainties that influence their labeling.

Most interestingly, analyses of the ERPs revealed similar modulation of the late components elicited by nonlinguistic and linguistic sounds: largest negativity for Ambiguous, intermediate for Same, and largest positivity for Different targets. These differences emerged with similar onset latencies in both cases (i.e., at 350 msec after target onset). Importantly, the Stimulus × Category interaction was significant: differences between Same and Different targets were larger for linguistic than for nonlinguistic sounds (see Figure 4). Because linguistic Different targets were words played backward, they were unfamiliar stimuli. Therefore, they were probably more surprising than nonlinguistic Different targets that were typical impact sounds and consequently more familiar but still different from the prime. By contrast, the priming effect for Ambiguous stimuli was similar in the linguistic and in the nonlinguistic conditions (i.e., the difference between Ambiguous and Same categories was not significantly different, either in amplitude or in scalp distribution for nonlinguistic and for linguistic stimuli).

However, results of separate ANOVAs nevertheless revealed that the spatiotemporal dynamics of the ambiguity effect was somewhat different for nonlinguistic and linguistic sounds, with an earlier onset for ambiguous impact sounds than for PW and a slight predominance over right frontal sites for PW. As noted in the introduction, although priming studies using environmental sounds have reported ERP effects that closely resemble the verbal N400 effect, they also showed differences in scalp distribution. As found here, priming effects were larger over the right hemisphere for words and over the left hemisphere for environmental sounds (Plante et al., 2000; Van Petten & Rheinfelder, 1995). By contrast, Orgs et al. (2006, 2008) found no interhemispheric differences but larger priming effects for sounds over posterior than anterior sites, and Cummings et al. (2006) found larger differences over anterior than posterior regions (as found here). Thus, the scalp topography seems somewhat variable between experiments, which most likely reflects differences in the acoustic properties of the stimuli and in task demands.

This conclusion is in line with results in the fMRI literature on verbal and environmental sounds showing mixed evidence in favor of the similarity of conceptual priming with nonlinguistic and linguistic sounds. For instance, although both spoken words and environmental sounds activate bilateral temporal regions (Giraud & Price, 2001; Humphries, Willard, Buchsbaum, & Hickok, 2001), Thierry et al. (2003) have demonstrated larger activation of the left anterior and posterior temporal areas for spoken words and larger activation of the right posterior superior temporal areas for environmental sounds. These between-experiments differences were taken to reflect differences in the task semantic requirements. Recently, Steinbeis and Koelsch (2008) provided evidence for both similar and different neural activations related to the processing of meaning in music and speech.

Taken together, our results are in line with previous literature (Schön et al., 2009; Daltrozzo & Schön, 2009; Orgs et al., 2006, 2007, 2008; Cummings et al., 2006; Plante et al., 2000; Van Petten & Rheinfelder, 1995) and argue in favor of the similarity of conceptual priming for nonlinguistic and linguistic sounds. Interestingly, the present results extend previous ones in several aspects. Most importantly, previous results were problematic in that words were always included in the design. As a consequence, the reported conceptual priming effects were possibly due to a linguistic strategy of generating words when listening to sounds. Although we also used linguistic stimuli to be able to compare priming effects within subjects, they were always presented in a separate session. The finding of N400-like components in a sound–sound design, as used in Experiment 2, shows that linguistic mediation is not necessary for an N400-like component to be elicited. Thus, this component may reflect a search for meaning that is not restricted to linguistic meaning. This interpretation is in agreement with the idea that variations in N400 amplitude are related to the "ease or difficulty of retrieving stored conceptual knowledge associated with a word or other meaningful stimuli" (Kutas et al., 2006, p. 10). Moreover, although two conditions (related vs. unrelated) were used in most previous studies, we used three conditions (Same vs. Ambiguous vs. Different) to more closely examine conceptual priming effects. In line with early studies of category membership effects (Ritter et al., 1983; Vaughan, Sherif, O'Sullivan, Herrmann, & Weldon, 1982; Boddy, 1981; Boddy & Weinberg, 1981), stimuli that clearly did not belong to the prime category elicited late positivity (P300 components), whereas stimuli that were ambiguous elicited late negativity (N400-like components) compared with stimuli that belonged to the prime category. Most importantly for our purposes, we were able to demonstrate similar relationships between categories for both nonlinguistic and linguistic target sounds.

Conclusion

These results add interesting information to the vast and still largely unexplored domain of the semiotics of sounds. Other experiments using different tasks and stimuli are needed to further explore the similarities and differences in conceptual priming for nonlinguistic and linguistic sounds. However, by using a homogeneous class of environmental sounds (impact sounds), by varying the relationship between prime and target sounds, and by comparing conceptual priming for nonlinguistic and for linguistic sounds within the same participants, we were able to make one step further and to show that conceptual priming develops in a sound–sound design without words and, consequently, that conceptual priming can develop without (or with reduced) linguistic mediation.

Acknowledgments

This research was supported by a grant from the Human Frontier Science Program (HFSP #RGP0053) to Mireille Besson (PI) and from the French National Research Agency (ANR, JC05-41996, "senSons") to Sølvi Ystad (PI).

Reprint requests should be sent to Mitsuko Aramaki, CNRS—Institut de Neurosciences Cognitives de la Méditerranée, 31, Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France, or via e-mail: Mitsuko.Aramaki@incm.cnrs-mrs.fr.

Notes

1. The Continua factor was not taken into account to keep enough trials in each condition.

2. Fine-grained analyses were computed as separated ANOVAs (that included the same factors as described in the Results section) in successive 50-msec latency windows from 0 to 800 msec after sound onset. Then, the 50-msec latency windows within which statistically similar effects were found were grouped together between 0–250, 250–400, and 400–800 msec or more specifically between 550 and 700 msec, and an ANOVA was conducted in each latency band.

3. To increase the similarities between the Different conditions for nonlinguistic and linguistic sounds, we also considered the possibility of playing impact sounds backward as was done for the words. However, although such sounds conserve the spectral characteristics of the original sound (i.e., acoustic cues characterizing the material category), they do no longer sound as impact sounds (i.e., the perception of impact disappears). They are therefore ambiguous and difficult to categorize. Because it was important to equate task difficulty for nonlinguistic and linguistic sounds (the words played backward are easy to categorize as different from the prime) and because words played backward keep the main attributes of vocal sounds, we decided to use typical sounds from another material category that are not ambiguous and easy to categorize as different.

4. On the basis of the results of behavioral data showing no differences between ambiguous targets associated with yes and no responses, we averaged ERPs in these two categories together to increase the signal to noise ratio. Moreover, no differences were found when ERPs to Ambiguous/Yes and Ambiguous/No were averaged separately.

REFERENCES

Alain, C., Schuler, B. M., & McDonald, K. L. (2002). Neural activity associated with distinguishing concurrent auditory objects. *Journal of the Acoustical Society of America, 111,* 990–995.

Aramaki, M., Baillères, H., Brancheriau, L., Kronland-Martinet, R., & Ystad, S. (2007). Sound quality assessment of wood for xylophone bars. *Journal of the Acoustical Society of America, 121,* 2407–2420.

Aramaki, M., & Kronland-Martinet, R. (2006). Analysis–synthesis of impact sounds by real-time dynamic filtering. *IEEE Transactions on Audio, Speech, and Language Processing, 14,* 695–705.

Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 250–267.

Ballas, J. A., & Howard, J. H., Jr. (1987). Interpreting the language of environmental sounds. *Environment and Behavior, 19,* 91–114.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision, and semantic priming. *Electroencephalography and Clinical Neurophysiology, 60,* 343–355.

Boddy, J. (1981). Evoked potentials and the dynamics of language processing. *Biological Psychology, 13,* 125–140.

Boddy, J., & Weinberg, H. (1981). Brain potentials, perceptual mechanisms and semantic categorization. *Biological Psychology, 12,* 43–61.

Cummings, A., Ceponiene, R., Koyama, A., Saygin, A. P., Townsend, J., & Dick, F. (2006). Auditory semantic networks for words and natural sounds. *Brain Research, 1115,* 92–107.

Daltrozzo, J., & Schön, D. (2009). Conceptual processing in music as revealed by N400 effects on words and musical targets. *Journal of Cognitive Neuroscience, 21,* 1882–1892.

de Saussure, F. (1916). *Cours de linguistique générale*. Paris: Payot.

Donchin, E. (1981). Surprise!…surprise? *Psychophysiology, 18,* 493–513.

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences, 11,* 357–374.

Duncan-Johnson, C. C., & Donchin, E. (1982). The P300 component of the event-related brain potential as an index of information processing. *Biological Psychology, 14,* 1–52.

Frey, A., Marie, C., Prod'Homme, L., Timsit-Berthier, M., Schön, D., & Besson, M. (2009). Temporal semiotic units as minimal meaningful units in music? An electrophysiological approach. *Music Perception, 26,* 247–256.

Giraud, A. L., & Price, C. J. (2001). The constraints functional neuroimaging places on classical models of auditory word processing. *Journal of Cognitive Neuroscience, 13,* 754–765.

Gordon, J. W. (1987). The perceptual attack time of musical tones. *Journal of the Acoustical Society of America, 82,* 88–105.

Holcomb, P. J., & Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language and Cognitive Processes, 5,* 281–312.

Humphries, C., Willard, K., Buchsbaum, B., & Hickok, G. (2001). Role of anterior temporal cortex in auditory sentence comprehension: An fMRI study. *NeuroReport, 12,* 1749–1752.

Hyde, M. (1997). The N1 response and its applications. *Audiology & Neuro-otology, 2,* 281–307.

Jasper, H. H. (1958). The ten–twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology, 10,* 371–375.

King, J., & Kutas, M. (1995). Who did what and when? Using word- and clause-level ERPs to monitor working memory usage in reading. *Journal of Cognitive Neuroscience, 7,* 376–395.

Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., & Friederici, A. (2004). Music, language and meaning: Brain signatures of semantic processing. *Nature Neuroscience, 7,* 302–307.

Kronland-Martinet, R., Guillemain, P., & Ystad, S. (1997). Modelling of natural sounds by time-frequency and wavelet representations. *Organised Sound, 2,* 179–191.

Kuriki, S., Kanda, S., & Hirata, Y. (2006). Effects of musical experience on different components of MEG responses elicited by sequential piano-tones and chords. *Journal of Neuroscience, 26,* 4046–4053.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension, Language comprehension and the N400. *Trends in Cognitive Sciences, 4,* 463–470.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207,* 203–204.

Kutas, M., & Hillyard, S. A. (1982). The lateral distribution of event-related potentials during sentence processing. *Neuropsychologia, 20,* 579–590.

Kutas, M., McCarthy, G., & Donchin, E. (1977). Augmenting mental chronometry: The P300 as a measure of stimulus evaluation time. *Science, 197,* 792–795.

Kutas, M., Van Petten, C., & Besson, M. (1988). Event-related potential asymmetries during the reading of sentences. *Electroencephalography and Clinical Neurophysiology, 69,* 218–233.

Kutas, M., Van Petten, C., & Kluender, R. (2006). Handbook of psycholinguistics. In M. A. Gernsbacher & M. Traxler (Eds.), *Psycholinguistics electrified II (1994–2005)* (2nd ed., pp. 659–724). New York: Elsevier Press.

Lebrun, N., Clochon, P., Etévenon, P., Lambert, J., Baron, J. C., & Eustache, F. (2001). An ERD mapping study of the neurocognitive processes involved in the perceptual and semantic analysis of environmental sounds and words. *Cognitive Brain Research, 11,* 235–248.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90,* 227–234.

Orgs, G., Lange, K., Dombrowski, J., & Heil, M. (2006). Conceptual priming for environmental sounds and words: An ERP study. *Brain and Cognition, 62,* 267–272.

Orgs, G., Lange, K., Dombrowski, J., & Heil, M. (2007). Is conceptual priming for environmental sounds obligatory? *International Journal of Psychophysiology, 65,* 162–166.

Orgs, G., Lange, K., Dombrowski, J. H., & Heil, M. (2008). N400-effects to task-irrelevant environmental sounds: Further evidence for obligatory conceptual processing. *Neuroscience Letters, 436,* 133–137.

Plante, E., Van Petten, C., & Senkfor, A. J. (2000). Electrophysiological dissociation between verbal and nonverbal semantic processing in learning disabled adults. *Neuropsychologia, 38,* 1669–1684.

Ritter, W., Simson, R., & Vaughan, H. G. (1983). Event-related potential correlates of two stages of information processing in physical and semantic discrimination tasks. *Psychophysiology, 20,* 168–179.

Ritter, W., Simson, R., Vaughan, H. G., & Friedman, D. (1979). A brain event related to the making of sensory discrimination. *Science, 203,* 1358–1361.

Schön, D., Ystad, S., Kronland-Martinet, R., & Besson, M. (2009). The evocative power of sounds: Conceptual priming between words and nonverbal sounds. *Journal of Cognitive Neuroscience, 22,* 1026–1035.

Shahin, A., Bosnyak, D. J., Trainor, L. J., & Roberts, L. E. (2003). Enhancement of neuroplastic P2 and N1c auditory evoked potentials in musicians. *Journal of Neuroscience, 23,* 5545–5552.

Shahin, A., Roberts, L. E., Pantev, C., Trainor, L. J., & Ross, B. (2005). Modulation of P2 auditory-evoked responses by the spectral complexity of musical sounds. *NeuroReport, 16,* 1781–1785.

Simson, R., Vaughan, H. G., & Ritter, W. (1977). The scalp topography of potentials in auditory and visual discrimination tasks. *Electroencephalography and Clinical Neurophysiology, 42,* 528–535.

Steinbeis, N., & Koelsch, S. (2008). Comparing the processing of music and language meaning using EEG and fMRI provides evidence for similar and distinct neural representations. *PLoS ONE, 3,* e2226. doi:10.1371/journal.pone.0002226.

Thierry, G., Giraud, A.-L., & Price, C. (2003). Hemispheric dissociation in access to the human semantic system. *Neuron, 38,* 499–506.

Van Petten, C., & Rheinfelder, H. (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia, 33,* 485–508.

Vaughan, J., Sherif, K., O'Sullivan, R. L., Herrmann, D. J., & Weldon, D. A. (1982). Cortical evoked responses to synonyms and antonyms. *Memory and Cognition, 10,* 225–231.

Walter, W. G., Cooper, R., Aldridge, V. J., McCallum, W. C., & Winter, A. L. (1964). Contingent negative variation: An electrical sign of sensorimotor association and expectancy in the human brain. *Nature, 230,* 380–384.

The Evocative Power of Sounds: Conceptual Priming between Words and Nonverbal Sounds

Daniele Schön[1], Sølvi Ystad[2], Richard Kronland-Martinet[2], and Mireille Besson[1]

Abstract

■ Two experiments were conducted to examine the conceptual relation between words and nonmeaningful sounds. In order to reduce the role of linguistic mediation, sounds were recorded in such a way that it was highly unlikely to identify the source that produced them. Related and unrelated sound–word pairs were presented in Experiment 1 and the order of presentation was reversed in Experiment 2 (word–sound). Results showed that, in both experiments, participants were sensitive to the conceptual relation between the two items. They were able to correctly categorize items as related or unrelated with good accuracy. Moreover, a relatedness effect developed in the event-related brain potentials between 250 and 600 msec, although with a slightly different scalp topography for word and sound targets. Results are discussed in terms of similar conceptual processing networks and we propose a tentative model of the semiotics of sounds. ■

INTRODUCTION

Most research on the question of how we are able to derive meaning from the external world has been investigated by studies on language. Although this line of research turned out to be very fruitful, leading to models of semantic processing (see McNamara, 2005), it remains a highly debated question whether mechanisms for meaning allocation rely on cerebral resources that are specific to language or that are common to other domains. Understanding the meaning of language may require specific functional and anatomical pathways. Alternatively, similar neural networks may be involved for linguistic information and for other types of meaningful information such as objects, pictures, nonlinguistic sounds, or music. In this article, we will prefer the term concept to the term meaning, because the former is a general term, whereas the latter is often associated to semantics and linguistics. One way of studying conceptual processing is to look at context effects on the processing of a target stimulus. In a seminal study, Kutas and Hillyard (1980) showed that the amplitude of a negative component of the event-related potentials (ERPs) peaking around 400 msec postword onset, the N400 component, is larger for final words unrelated to the preceding sentence context than for related words (*The pizza was too hot to cry/eat*). Thereafter, the N400 has been widely used to investigate semantic processing in language, using the classical semantic priming paradigm, wherein one single word is used to create a context that influences the processing of a following target word (Bentin, McCarthy, & Wood, 1985). More recently, several researchers have become interested in studying whether an N400 can be elicited and modulated by the conceptual relation in a nonlinguistic context. Indeed, several studies have been published on conceptual processing with pictures (Holcomb & McPherson, 1994), odors (Castle, Van Toller, & Milligan, 2000; Sarfarazi, Cave, Richardson, Behan, & Sedgwick, 1999), and music (Daltrozzo & Schön, 2009; Frey et al., 2009; Koelsch et al., 2004).

Within the auditory domain, one way of comparing linguistic and nonlinguistic conceptual processing has been to use spoken words and environmental sounds. Environmental sounds are interesting in that they bear a direct relation with the source of the sound. They establish a reference to an object (bottle, cork, corkscrew) or an action (turn, pull, open). A number of studies have used the ERP method and the classical priming paradigm to study the conceptual processing of environmental sounds. To our knowledge, the first study was conducted by Van Petten and Rheinfelder (1995). They presented spoken words followed by environmental sounds and vice-versa. Words preceded by unrelated sounds evoked a larger N400 than those preceded by related sounds. This N400 effect (i.e., the difference between unrelated and related targets) was slightly lateralized to the right hemisphere. Sounds preceded by unrelated words also evoked a larger N400 than those preceded by related words but this effect was larger over the left hemisphere. Orgs, Lange, Dombrowski, and Heil (2006, 2007) used a similar design but with shorter stimuli (300 msec instead of 2500 msec)

[1]CNRS & Université de la Méditerranée, Marseille, France, [2]CNRS, Marseille, France

and also found similar effects on the N200 and N400 components. Finally, Cummings et al. (2006, p. 104) compared behavioral and electrophysiological responses to words, environmental sounds, and nonmeaningful sounds ("not easily associated with any concrete semantic concept") in semantically matching or mismatching visual contexts (photos). They found that words and environmental sounds mismatching the visual context evoked a larger N400 than words and environmental sounds matching the visual context. By contrast, no differences were found for the so-called nonmeaningful sounds. These sounds were selected so that they always fit either a smooth or a jagged category, and should, as such, have evoked concepts related to smoothness or roughness. However, the repetitive character (always smooth or jagged) might have greatly reduced the influence of the visual context on these "nonmeaningful" sounds.

Although the result of these experiments are most often interpreted as reflecting some form of conceptual priming between words or pictures and environmental sounds, they may also reflect linguistic mediated effects. For instance, looking at a picture of a cat and listening to the meowing of a cat may automatically activate the verbal label {cat}. This conceptual effect cannot be considered as purely nonlinguistic because there could be a semantic mediation between the (linguistic) label assigned to the drawing and the label assigned to the sound.

The purpose of the present study was to try to reduce, as much as possible, the chance that such labeling takes place. To this end, we generated, recorded, and, in some cases, also resynthesized sounds so that it was highly unlikely to identify a source (Ystad, Kronland-Martinet, Schön, & Besson, 2008). Thus, while people, when hearing a sound, may try to identify the source that produced it, our sounds should greatly reduce the likelihood of labeling compared to previous studies using environmental sounds.

We conducted two experiments. In Experiment 1, sounds were used as a context and were followed by visual words. In Experiment 2, visual words were used as a context and were followed by sounds. In Experiment 1, we predicted a larger N400 to words preceded by conceptually unrelated sounds compared to words preceded by related sounds. In Experiment 2, we predicted a larger N400 to sounds preceded by conceptually unrelated words compared to sounds preceded by related words.

EXPERIMENT 1

Methods

Participants

Sixteen nonmusician volunteers were tested in this experiment. All were right-handed, neurologically normal, had normal or corrected-to-normal vision, normal audition, and were native French speakers (age: $M = 27.5$ years, 7 women). All participants were paid for their participation to the experiment. Due to large drifts in EEG data, two participants were discarded from analyses.

Stimuli

Stimuli were built to favor what Pierre Schaeffer (1966) called "acousmatic listening" in his book *Traité des Objets Musicaux.* The term acousmatic relates to the ability of listening to a sound without considering the object(s) that created it, hence, reflecting the perceptual reality of a sound independently of the way it is produced or transmitted. By extension, sounds with no recognizable sources are "acousmatic sounds." These sounds are typically used as compositional resources in contemporary music such as "musique concrète" or electroacoustic music.

Stimuli included sounds originally intended for musical composition in electroacoustic music, as well as sounds specifically recorded for the experiment. Recordings aimed at decontextualizing the sounds to force listeners to pay attention to the sound itself. Some sounds were also obtained from traditional instruments, but their familiarity was altered by untraditional playing techniques or modified by signal processing techniques. To obtain a sound corpus representative of the main sound morphologies found in nature, we used the classification system proposed by Schaeffer (1966) and called "typology of sound objects," where sounds are mainly sorted as a function of their mass and shape. Schaeffer's typology of sound objects contains 35 classes of sounds, but only the nine main classes called "balanced sounds" (*sons équilibrés*) were used in this study. Two main aspects determine balanced sounds: maintenance (the way the energy is spread over time) and mass (linked to the spectral content of sounds and to the potential existence of pitch). Maintenance is used to distinguish sustained, iterative, and impulsive sounds. Mass distinguishes sounds with constant, varying, or indefinable pitch. The nine sound categories used here resulted from the combination of the three types of maintenances with the three types of masses (sustained with constant pitch, sustained with varying pitch, sustained with indefinable pitch; iterative with constant pitch, iterative with varying pitch, iterative with indefinable pitch; impulse with constant pitch, impulse with varying pitch, impulse with indefinable pitch).

We first selected 70 sounds representative of the nine categories of balanced sounds. Seven participants were then asked to listen to the sounds and to write down the first few words that came to mind. Although we did not measure the time participants needed to write the words evoked by each sound, this procedure lasted for almost 2 hr (i.e., more than one minute/sound in average). Participants were specifically asked to focus on the associations evoked by the sounds without trying to identify the physical sources that produced them. For instance, a particular sound evoked the following words: dry, wildness, peak, winter, icy, polar, cold. Sounds that evoked identical or semantically close words for at least three of the seven participants were selected for the experiment resulting in a final set of 45 sound–word pairs. Each sound was paired with the proposed word of highest lexical frequency among the three words (e.g., "cold" was chosen between icy, polar, and cold). Finally, 45 unrelated pairs were built from this material by recombining words and sounds in a different manner. Average sound duration was 820 msec, standard deviation was 280 msec.

Procedure

Participants were comfortably seated in a Faraday box. Presentation of the sound was followed by the visual presentation of a word for 200 msec, with a stimulus onset asynchrony (SOA) of 800 msec (i.e., close to the average sound duration). Words were displayed in white lowercase on a dark background in the center of a 13-inches 88-Hz computer screen, set at about 70 cm from the participant's eyes. Participants were instructed to decide whether or not the sound and the target word fitted together by pressing one of two buttons. They were also told that the criterion for their relatedness judgment was of the domain of evocation rather than some direct relation such as the barking of a dog and the word "dog." It was also made clear that there were no correct or incorrect responses and participants were asked to respond as quickly as possible without much explicit thinking. A training session comprising 10 trials (with sounds and words different from those used in the experiment) was used to familiarize participants with the task.

Two seconds after word presentation, a series of "X" appeared on the screen signaling that participants could blink their eyes. A total of 45 related and 45 unrelated pairs were presented in pseudorandom order (no more than 5 successive repetitions of pairs belonging to the same experimental condition). Response side association (yes or no/left or right) was balanced across participants. A debriefing followed the experiment, questioning on possible strategies used by each participant (e.g., Do you have the feeling that you used a specific strategy? Do you have the feeling that the relation popped out from the stimuli or did you have to look for a relation? Did it happen that you gave a verbal label to sounds? Could you tell how sounds were generated? Did you try to find out?).

Data Acquisition and Analysis

Electroencephalogram (EEG) was recorded continuously at 512 Hz from 32 scalp electrodes (International 10–20 System sites) using a BioSemi Active Two system. Data were re-referenced off-line to the algebraic average of left and right mastoids. Trials containing ocular artifacts, movement artifacts, or amplifier saturation were excluded from the averaged ERP waveforms. Data were detrended and low-pass filtered at 40 Hz (12 dB/octave).

ERP data were analyzed by computing the mean amplitude, starting 100 msec before the onset of word presentation and ending 1000 msec after. Because there were no a priori correct responses, averages for related and unrelated pairs were based on the participants' responses. Repeated measures analyses of variance (ANOVAs) were used for statistical assessment of the independent variable (relatedness) To test the distribution of the effects, six regions of interest (ROIs) were selected as levels of two topographic within-subject factors (i.e., anteroposterior and hemisphere): left (AF3, F3, F7) and right (AF4, F4, F8) frontal; left (FC1, C3, CP1) and right (FC2, C4, CP2) central; and left (P3, PO3, P7) and right (P4, PO4, P8) parietal. Data were analyzed using latency windows of 50 msec in the 0 to 1000 msec range. Only results that were statistically significant in at least two successive 50-msec windows are reported. All p values were adjusted with the Greenhouse–Geisser correction for nonsphericity, when appropriate. Dunn–Sidak test was used in correcting post hoc multiple comparisons. The statistical analyses were conducted with Cleave (www.ebire.org/hcnlab) and Matlab.

Results

Behavioral Data

Even if the experimental conditions were defined by the participants' responses, we considered behavioral accuracies as the matching degree between the participants' responses and the related and unrelated pairs based upon the material selection procedure.

This analysis indicated an average accuracy of 77% that is significantly above the 50% chance level ($\chi^2 = 516$, $df = 25$, $p < .001$). No significant differences were found on RTs (mean ± SD = 1018 ± 170 and 1029 ± 220 msec, respectively, Wilcoxon matched pair test, $p = .97$).

Event-related Brain Potentials Data

As can be seen in Figure 1, visual word presentation elicited typical ERP components. An N1 component, peaking around 100 msec after stimulus onset, is followed by a P2 peaking around 150 msec. No differences are visible over these components for the related and unrelated targets. However, around 300 msec after stimulus onset, ERPs to related and unrelated words start to diverge with unrelated words associated to a larger negativity in the 300–700 msec latency window. This effect is maximal over central electrodes.

To analyze in detail how ERP components were modulated by the independent variables manipulated in this experiment, we first computed repeated measure ANOVAs with Relatedness (related/unrelated) × Anteroposterior (frontal, central, and parietal ROIs) × Hemisphere (left/right) as within factors. An interaction between relatedness, anteroposterior, and hemisphere factors was significant in

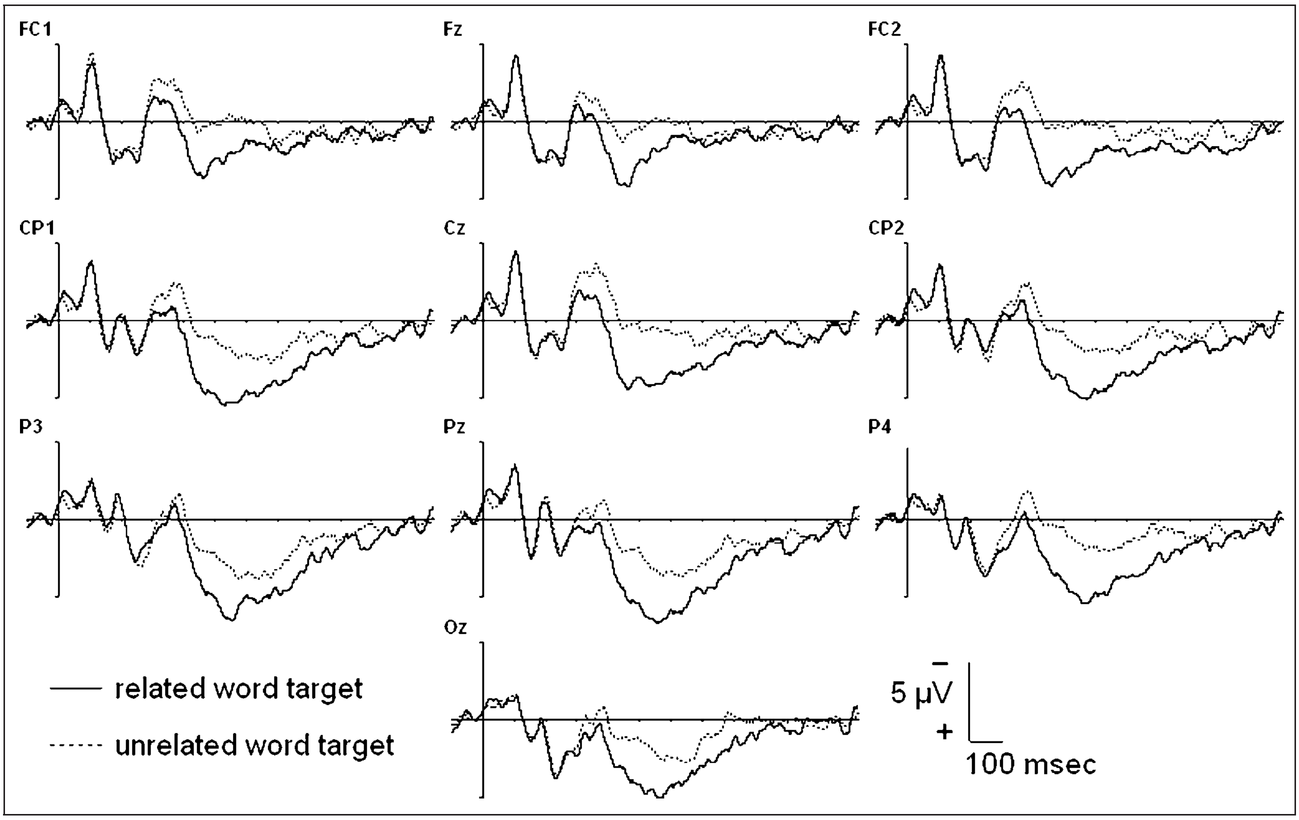


Figure 1. Grand-averaged ERPs to related and unrelated target words according the participants' responses (14 participants). Stimulus onset is the vertical calibration bar.

the 250–350 msec latency range [$F(2, 24) = 5.3, p < .05$]. Post hoc comparisons revealed a larger negativity to unrelated compared to related words over the left frontal region ($p < .0001$, effect size = 1.3 μV). In the 350–450 msec latency range, although the triple interaction was no longer significant, there was a significant Relatedness × Anteroposterior interaction [$F(2, 24) = 7.1, p < .01$] due to a larger effect over frontal and central regions compared to parietal regions, for both hemispheres (frontal: $p < .001$, effect size = 1.9 μV; central: $p = .001$, effect size = 2.0 μV; parietal: $p > .05$). Finally, in the 450–600 msec latency range, the relatedness effect was equally distributed over all ROIs and hemispheres [main effect of relatedness: $F(1, 12) = 6.83, p < .05$, effect size = 1.9 μV].

Discussion

Although, as stated above, there are no correct and incorrect responses in this experiment, the 77% participants' accuracy can be interpreted as a sign of a rather low intersubject variability between the experimental group and the group used in the pilot study. Moreover, this also shows that participants well understood the task, namely, to use the evocative features of a sound to judge its relation to a word. The fact that we did not find any significant difference on RTs between related and nonrelated responses is not so surprising insofar as the design is not a typical priming paradigm in that the task is more a value judgment than a categorization. This makes the task rather difficult, and RTs rather long. Indeed, priming experiments using lexical decision or categorical decision tasks report RTs between 500 and 800 msec (as compared to more than 1 sec here). Therefore, it might be the case that task difficulty, linked to the stimuli used in the present study, did override conceptual priming effects.

Electrophysiological results strongly resemble previous studies using target words preceded by semantically related or unrelated words (Chwilla, Brown, & Hagoort, 1995; Bentin et al., 1985), related or unrelated environmental sounds (Orgs et al., 2006, 2007; Van Petten & Rheinfelder, 1995), or musical excerpts (Daltrozzo & Schön, 2009; Koelsch et al., 2004). The similarity is mainly seen at the morphological level, with a negative component peaking between 300 and 400 msec. As in previous experiments, the amplitude of this N400-like component is modulated by the relatedness of the preceding sound. These results, together with the frontocentral distribution of the relatedness effect, will be further discussed in light of the results of Experiment 2.

EXPERIMENT 2

The same experimental design and stimuli as in Experiment 1 were used in Experiment 2, except that the order of stimulus presentation was reversed with words presented as primes and sounds as targets.

Methods

Participants

In order to reduce stimuli repetition effects, we tested a new group of 18 nonmusician volunteers. All were right-handed, neurologically normal, had normal or corrected-to-normal vision, normal audition, and were native French speakers (age: $M = 26$ years, 8 women). All participants were paid for their participation in the experiment. Due to large drifts in the EEG data, five participants were discarded from analyses.

Stimuli

Same materials were used as in Experiment 1.

The procedure was identical to Experiment 1, except that words were used as primes and sounds as targets. This has a direct effect on the SOA. In Experiment 1, wherein sounds were used as primes, we used an 800-msec SOA because the average sound duration was 820 msec and the end of the sound is generally not very informative due to natural damping. In Experiment 2, visual target words are used as primes. The use of 800 msec SOA would be too long as words do not require 800 msec to be read and also by comparison to typical durations used in priming experiments with visual words. Thus, a 500-msec SOA was used in Experiment 2.

Procedure

Data acquisition and analysis. Same as in Experiment 1.

Results

Behavioral Data

Participants showed an average accuracy of 78% that is significantly above the 50% chance level ($\chi^2 = 681$, $df = 35$, $p < .001$). These results were not significantly different from those of Experiment 1 for both unrelated and related trials (Mann–Whitney U test: $p = .70$ and $p = .67$, respectively). No significant relatedness effect was found on RTs (1320 ± 230 msec and 1295 ± 217 msec, respectively, Wilcoxon matched pair test, $p = .3$). However, RTs were slower than in Experiment 1, for both related and unrelated responses (Mann–Whitney U test: $p = .001$).

Event-related Brain Potentials Data

As can be seen in Figure 2, sound presentation elicited an N1 component, peaking around 130 msec after stimulus onset, followed by a P2 peaking around 220 msec. No

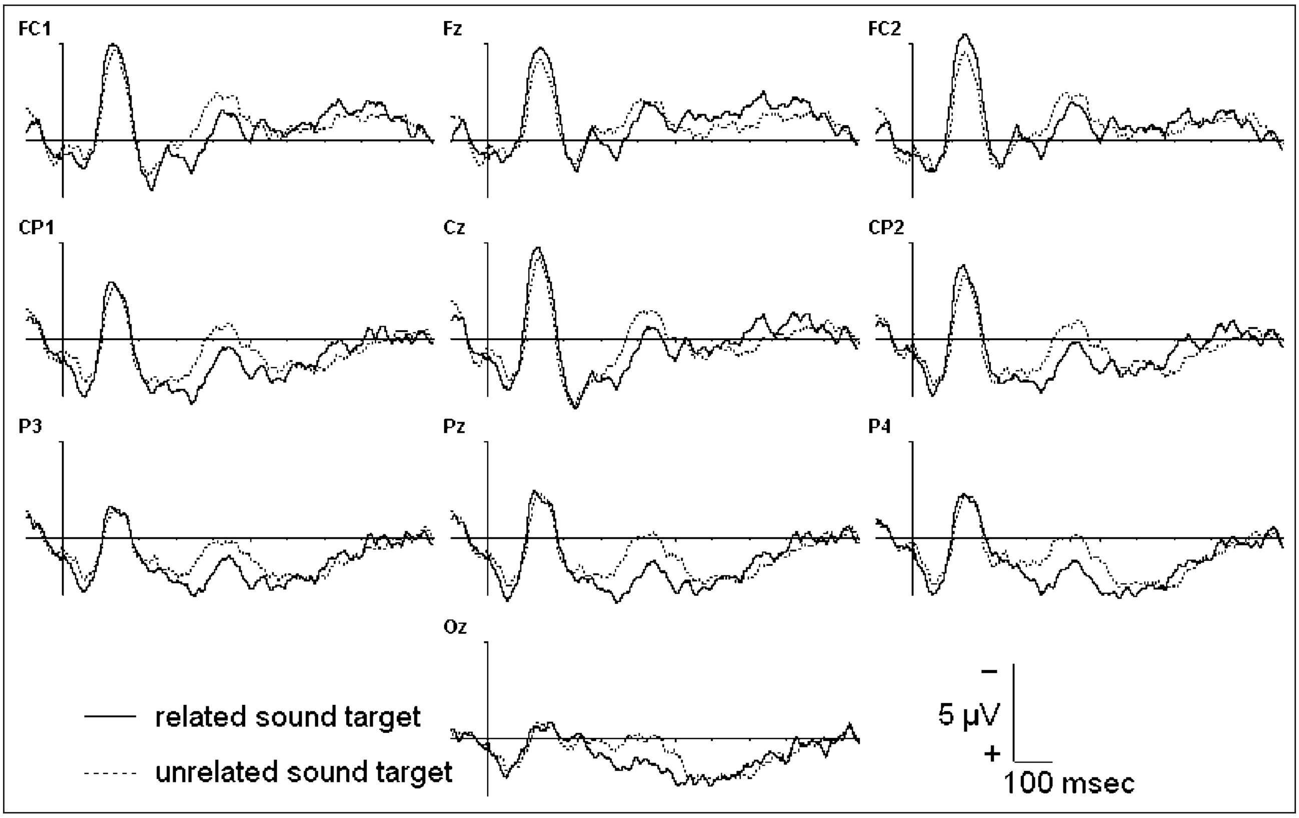


Figure 2. Grand-averaged ERPs to related and unrelated target sound according the participants' responses (13 participants). Stimulus onset is the vertical calibration bar.

differences are visible over these components for related and unrelated targets. However, around 300 msec after stimulus onset, ERPs to related and unrelated sounds start to diverge: Compared to related sounds, ERPs to unrelated sounds elicit a larger negativity in the 300–500 msec latency window. This effect is maximal over parietal electrodes.

To analyze in detail how these components were modulated by the independent variables manipulated in this experiment, we computed repeated ANOVAs with Relatedness (related/unrelated) × Anteroposterior (frontal, central, and parietal ROIs) × Hemisphere (left/right) as within factors using 50-msec latency windows.

The main effect of relatedness was significant in the 300–400 msec latency range [$F(1, 13) = 6.95$, $p < .05$; effect size = 1.2 μV]. In the 400–600 msec latency range, there was a significant Relatedness × Anteroposterior interaction [$F(2, 26) = 6.03$, $p < .05$] due to a larger relatedness effect over central and parietal regions compared to frontal regions (400–500 msec: frontal, $p > .7$; central, $p < .01$, effect size = 1 μV; parietal, $p < .001$, effect size = 1.1 μV; 500–600 msec: frontal and central, $p > .1$; parietal, $p < .05$, effect size = 1.0 μV).

In order to compare the relatedness effect found in the two experiments, we computed a four-way ANOVA including the same within-subject factors and target modality (word/sound) as between-subjects factor. Results showed a significant triple interaction of target modality, relatedness, and anteroposterior factors between 250 and 600 msec [$F(2, 50) = 5.2$, $p < .05$]. Post hoc analyses showed that this interaction was due to a lack of relatedness effect over frontal regions when sounds were used as targets (see Figure 3).

Discussion

Participants' accuracy was similar to what found in Experiment 1, which again shows that participants were sensitive to the conceptual word–sound relation. Moreover, the similarity of results in both experiments also points to an equivalent level of task difficulty. However, RTs were longer in Experiment 2 than in Experiment 1. We interpret this difference in terms of stimulus modality and familiarity. Although in Experiment 1 the target word was presented in the visual modality and lasted for 200 msec, in Experiment 2 the target sound average duration was around 800 msec. Therefore, although for words all information necessary to make a relatedness decision was available within 200 msec for sounds, information only gradually becomes available over time and the relatedness decision cannot be made as fast as for words. Moreover, whereas target words (Experiment 1) were highly familiar, target sounds (Experiment 2) were unfamiliar, as they were specifically created for the purpose of the experiments in order to minimize linguistic mediation. Thus, both modality and familiarity may account for longer RTs to sound than to word targets.

Electrophysiological data strongly resemble previous studies using target sounds preceded by semantically related or unrelated words (Orgs et al., 2006; Van Petten & Rheinfelder, 1995). The amplitude of a negative component, peaking between 300 and 400 msec, was modulated

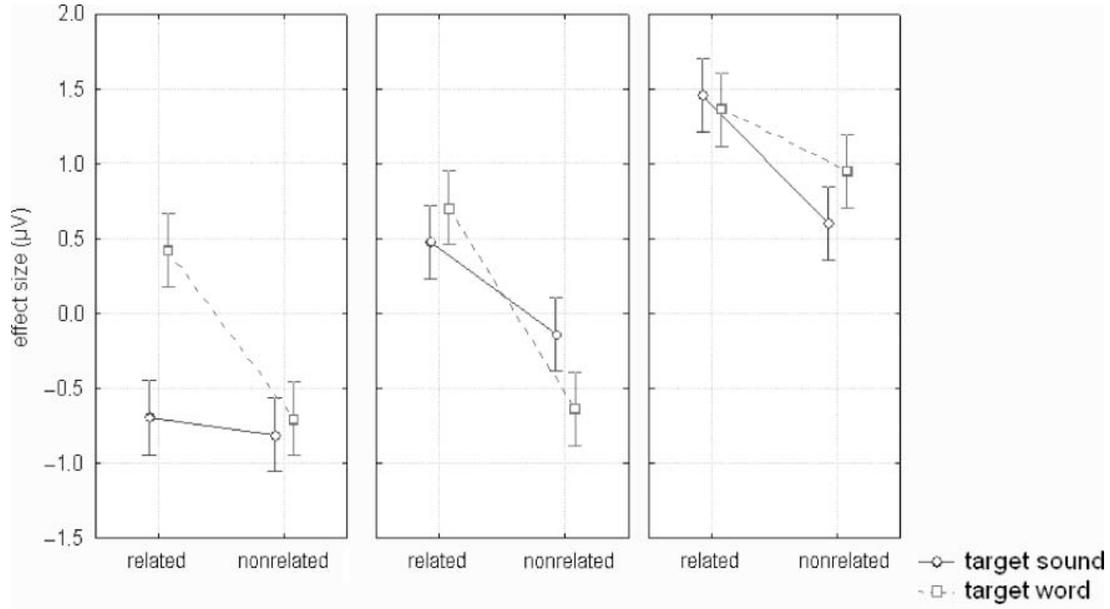


Figure 3. Interaction of Target modality × Relatedness × ROI (Anteroposterior factor). The relatedness effect is almost absent at frontal sites for sound targets. Vertical bars denote 95% confidence interval.

by its relatedness to the preceding sound. However, the centro-parietal distribution of this effect differed from the fronto-central distribution of the effect observed in Experiment 1. These differences are discussed in the General Discussion in view of the literature.

Finally, although N400-like components have been reported in most studies using environmental sounds, Cummings et al. (2006) failed to report such an effect in what they called the "nonmeaningful sound condition." In their study, the authors compared ERPs evoked during the processing of words, environmental sounds, and nonmeaningful sounds in semantically matching or mismatching visual contexts. However, there are two major differences between our acousmatic sound stimuli and the one used in the nonmeaningful sound condition by Cummings et al. that may explain why we found an effect while they did not. First, their criterion in choosing sounds was very different from ours. They chose sounds from an internet database (www.sounddog.com) in order "to portray either a smooth sound (e.g., a harmonic tone), or a jagged sound (e.g., a cracking sound)." Therefore, they used a binary categorical criterion, which most probably generated a rather repetitive sound corpus. By contrast, we chose sounds that could be used in electroacoustic music composition and we tried to maximize sound variability. Second, the context was not linguistic as in our study but comprised abstract visual patterns: "colorful, nonobject-looking patterns chosen to represent one of two categories," smooth or jagged. Therefore, the conceptual relation between context and sound target was very different in the two studies: each sound matching one out of two categories (smooth/jagged) versus our choice of looking for an "optimal" verbal descriptor for each given sound.

These differences in sound selection, context, and context–sound relations may possibly explain the different results. Thus, the lack of relatedness effect for the nonmeaningful sound condition in Cummings et al.'s study could be due to a weaker conceptual processing of their sound database compared to our more "musical" sounds, a weaker conceptual processing of the abstract images compared to words used in our experiment and a weaker relation between context and sounds. Unfortunately, not enough details concerning the nonmeaningful sound condition are given in the manuscript in order to clearly understand the reasons of the different results.

GENERAL DISCUSSION

The general aim of the present experiments was to compare conceptual priming effects when either words or sounds were used as targets. The originality of our approach was to create sounds whose sources were, in most cases, impossible to identify (i.e., "acousmatic sounds") in order to reduce the influence of linguistic mediation. In both experiments, behavioral data showed that participants were able to evaluate the sound–word or word–sound relations with relative low intersubject variability and good consistency. No relatedness effect was found on RTs. However, electrophysiological data revealed an enhanced negativity in the 250–600 msec latency range to unrelated compared to related targets in both experiments, although with a more fronto-central distribution to word targets in Experiment 1 and more centro-parietal distribution to sound targets in Experiment 2. These findings are discussed relative to the linguistic versus amodal theories of concepts.

To Label or Not to Label

The reason for using nonverbal stimuli in previous studies was to determine whether behavioral effects such as priming effects and electrophysiological effects such as the N400 effect are specific to language or not. This is not a trivial question to answer because a recurrent question in the behavioral and electrophysiological literature on conceptual priming with nonverbal stimuli is whether nonverbal items are given a verbal label or not (Koelsch et al., 2004). Indeed, if labeling takes place, then behavioral or electrophysiological differences between related and unrelated items may simply reflect a linguistic relatedness effect. Such results would therefore support a linguistic theory of concepts. By contrast, if the effects are found independently of language mediation, such results would support a more general and amodal theory of concepts.

In this respect, the relevant aspect of our study is the low probability that labeling takes place. Indeed, although labeling the picture or line drawing of a cat (Holcomb & McPherson, 1994), the odor of a lemon (Sarfarazi et al., 1999), the barking of a dog (Van Petten & Rheinfelder, 1995), or the ringing of a telephone (Orgs et al., 2006, 2007) is easy and rather automatic (i.e., we cannot avoid labeling), the stimuli used in the present experiment are rather difficult to label because they are uncommon sounds and it is difficult to identify the source that produced them (sound examples are available at www.sensons.cnrs-mrs.fr/Schaeffer/Schaeffer.html). Not surprisingly, in the pilot study, when participants were asked to find a related verbal label for each sound (see Stimuli section), they asked to listen to the sound several times and they needed 1 min rather than 1 sec to find an appropriate label. Of course, this does not completely prevent the possibility that participants of the two experiments still imagine and label a possible source of sounds (although incorrect), or attach verbal associations evoked by the sounds, for instance, by using adjectives to describe or characterize them. The argument is not that verbal labeling is completely prevented (which is probably impossible to do), but that it was strongly reduced in our experiments compared to previous experiments using, for instance, environmental sounds.

Another related issue is that the strength of the conceptual relation between sounds and words is probably weaker in our experiments compared to studies using environmental sounds. Indeed, although a barking sound immediately and strongly calls for a precise label {dog}, the sounds we used rather evoked a larger set of concepts and feelings, but in a weaker manner (see Figure 4). This might explain why the relatedness effect sizes in our experiment are smaller than those found in some studies using environmental sounds (Orgs et al., 2006, 2007; Cummings et al., 2006), although they do not seem to greatly differ from other studies (Plante, Petten, & Senkfor, 2000; Van Petten & Rheinfelder, 1995).

Differences in scalp topography between Experiments 1 and 2 and, more generally, between the several experiments that used environmental sounds can be taken to argue that the N400 effect encompasses different processes. Indeed, it may be influenced by both the high-level cognitive processing of the conceptual relation between two stimuli (that may be similar in all experiments) and the lower-level perceptual processes linked with the specific acoustic features of the sounds (that would be different across experiments). In this respect, it is interesting to note that the scalp distribution of the relatedness effect associated to unrelated and related targets changes dynamically over time, which reflects the spatio-temporal dynamics and potential interactions in the underlying processes (Experiment 1: left frontal in the early latency band, fronto-central in the 350–450 msec range, and spread across scalp sites in the 450–600 msec range; Experiment 2: widely distributed across scalp sites in the 300–400 msec range and spatially more localized over centro-parietal sites in the 400–600 msec range).

Finally, it is also important to note that all negativities in the 300–600 msec latency band are not necessarily N400 components. For instance, what Orgs et al. (2006, 2007) consider as an early onset of the N400 for sounds (between 200 and 300 msec) may also be an N200 component reflecting a physical mismatch between the sound expected on the basis of the context and the sound actually presented or an offset potential trigged by a constant sound offsets (all sounds had 300 msec duration). In conclusion, one way to reconcile these different results is to consider that the cognitive processing of the conceptual relation, as reflected by the amplitude modulation of the N400, is present in all the abovementioned experiments, but that other perceptive processes, which differ depending upon the specific characteristics of the stimuli, are also involved and influence the scalp distribution of the N400 effect.

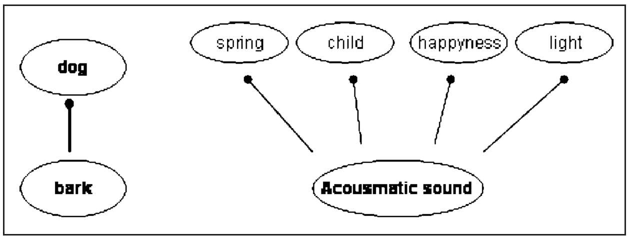


Figure 4. Different strengths between an environmental sound and a concept, and an acousmatic sound and the related concepts.

Modeling the Semiotics of Sounds

It might be the case that, after signal analysis, taking place in the brainstem and in the primary auditory regions, sound representations automatically spread to a whole range of concepts. With this respect, an attractive view is that of distributed network models (Anderson, 1993; McClelland, Rumelhart, & the PDP Research Group, 1986). According to these models, concepts are represented as patterns of activation of an interconnected set of units. Similar concepts share similar patterns of activation. What is interesting in these models is the fact that the units can be thought of as representing aspects of a given object (e.g., sound, word). Most importantly, these aspects "need not be nameable or correspond in any obvious way to the features people might list in a description of the entity" (McNamara, 2005, p. 29). Within such a theoretical framework, sound features, such as attack time, spectral centroid, spectral variation, energy modulation, inharmonicity, and others, might become input units shared by several patterns of activation for a set of concepts. This means that a combination of sound features, so-called invariants, might be used by the listeners to determine specific aspects of sounds. In Figure 5, we propose a model that may explain how conceptual priming takes place in our studies. Depending on whether we read it from left to right or from right to left, the model explains the effects of Experiment 1 or Experiment 2, respectively. We will quickly go through it beginning from the left, that is, for Experiment 1, wherein the probe is a sound and the target is a word. Once a sound is presented, acoustic features are extracted and represented in a sparse manner at the acoustic feature level. In the case of an environmental sound, these feature representations may feed forward to activate a precise item in the sound lexicon and possibly find a good match. If no good match is found in the lexicon, as in the case of the present experiment using acousmatic sounds, competition might be rather high, slowing down sound processing. In such cases, a direct path to an amodal concept representation level may take over, possibly influenced by the emotional connotation of the sound, carried by the signal features (Juslin & Västfjäll, 2008; Koelsch & Siebel, 2005; Juslin & Laukka, 2003). This amodal representation of concepts would be the link between concepts evoked by sounds and concepts evoked by words. Indeed, activation of the concepts in the amodal concept lexicon would spread to the semantic level (and possibly to the lexical and letter level), therefore priming the semantic processing of a following presented word.

Of course, this is just a tentative model and several issues need clarification. First, the existence of a direct pathway from the feature extraction level to the amodal concept lexicon needs to be proved. Indeed, it might be the case that processing always transits via a sound lexicon. Second, although we believe that sound features can be automatically directed to an "emotional parser" without transiting via the sound lexicon, this needs to be

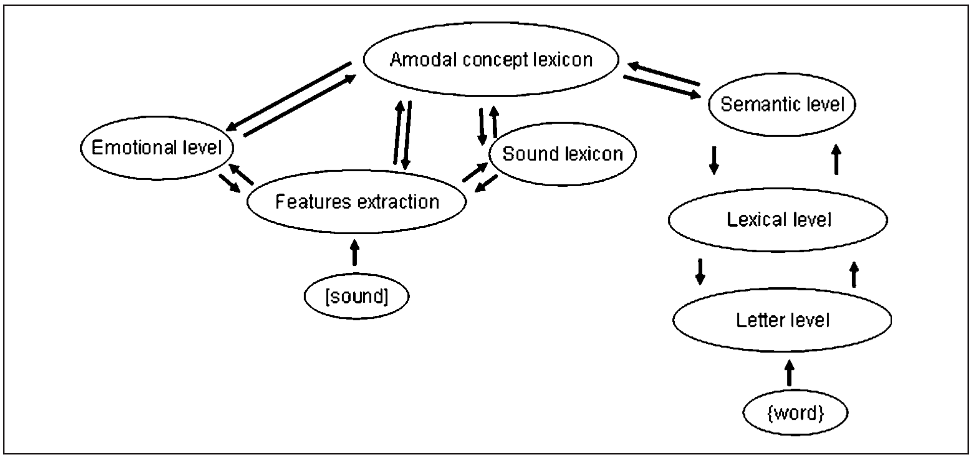


Figure 5. Tentative model describing how sounds can evoke concepts.

demonstrated. These two issues could possibly be addressed by studying patients with nonverbal auditory agnosia (Vignolo, 1982). These patients are particularly interesting because they do not have linguistic deficits, but they cannot anymore recognize environmental sounds. It is interesting to note that these patients can have difficulties in discriminating acoustically related sounds or semantically related sounds, often depending upon the lesion site (right or left, respectively; Schnider, Benson, Alexander, & Schnider-Klaus, 1994; Faglioni, Spinnler, & Vignolo, 1969). Unfortunately, testing procedures used in previous studies do not allow for qualitative error analysis, which would be most informative in order to understand whether these patients have a deficit at the sound lexicon level, at the amodal concept lexicon or both. Moreover, little is also known concerning whether these patients, experiencing difficulties in discriminating acoustically or semantically related sounds, can still attribute the correct emotion to these sounds. This could be an elegant way of showing that sound features can be processed by an "emotional parser" without transiting via the sound lexicon.

From Sound to Music

We previously said that the sound stimuli used in the present study are acousmatic sounds intended for "musique concrète." Of course, by no means would we claim that we studied music processing in this study, insofar as music goes well beyond a single sound. However, for a theoretical purpose, it is interesting to think about the relation between a single "acousmatic sound" and music.

Indeed, the fact that conceptual processing can take place for a single sound, independently of its source, is also of interest for the understanding of the meaning of music. Surprisingly, although timbre variations are consciously used by composers and by musicians during performance (Barthet, Kronland-Martinet, & Ystad, 2008), the sound structure or "sound matter" (having a quasi-physical connotation) is marginal or not considered at all in the taxonomy of musical meanings (see Patel, 2008). The musically meaningful elementary unit is, most of the time, considered to be a set of sounds composing a motif, a sentence, a theme, and so on. Of course, the way sounds combine in music is of utmost importance and, indeed, most theories on the meaning of music focus on the relation between musical events (e.g., Jackendoff, 1991; Meyer, 1956, see also Frey et al., 2009, for experimental evidence). However, if a single sound, out of a musical context, can generate meaning, we should question the possibility that, in music, elementary units, much shorter than motifs or themes, may also convey part of the musical meaning, via the property of the "sound matter" they carry at each single lapse of time. With respect to this hypothesis, and extending the work of Koelsch et al. (2004), we recently used a similar design to show that 1 sec of music can communicate concepts and influence the processing of a following target word (Daltrozzo & Schön, 2009). Most importantly, we also showed that when music is preceded by a verbal context, the amplitude of an N400 component to music is modulated by the degree of conceptual relation between the context and the musical excerpt, as soon as 300 msec after music onset. The fact that concepts carried by words can influence the processing of a following musical excerpt can be interpreted as a strong sign that the time window of elementary meaningful units in music might be very small, well below the time window of a motif or a theme. Therefore, the model we propose here for conceptual processing of sounds might also be at work in music listening. The meaning of music will, therefore, be the result of a rather complex process, taking into account the structural properties of music, the personal and cultural background of the listener, the aesthetic and emotional experience, and also the structure or matter of the sounds whereof a given excerpt is composed.

Acknowledgments

This project has been supported by the French National Research Agency (ANR 05-BLAN-0214 "Music and Memory" to D. Schön; ANR, JC05-41996, "senSons," www.sensons.cnrs-mrs.fr/ to S. Ystad).

Reprint requests should be sent to Daniele Schön, INCM-CNRS, 31 Ch J Aiguier, Marseille 13402, France, or via e-mail: schon@incm.cnrs-mrs.fr.

REFERENCES

Anderson, J. A. (1993). The BSB Model: A simple nonlinear autoassociative neural network. In M. Hassoun (Ed.), *Associative neural memories* (pp. 77–103). New York, NY: Oxford University Press.

Barthet, M., Kronland-Martinet, R., & Ystad, S. (2008). Improving musical expressiveness by time-varying brightness shaping. In *Lecture notes in computer science* (Vol. 4969, pp. 313–337). Berlin: Springer-Verlag.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology, 60,* 343–355.

Castle, P. C., Van Toller, S., & Milligan, G. J. (2000). The effect of odour priming on cortical EEG and visual ERP responses. *International Journal of Psychophysiology, 36,* 123–131.

Chwilla, D. J., Brown, P. M., & Hagoort, P. (1995). The N400 as a function of the level of processing. *Psychophysiology, 32,* 274–285.

Cummings, A., Ceponiene, R., Koyama, A., Saygin, A. P., Townsend, J., & Dick, F. (2006). Auditory semantic networks for words and natural sounds. *Brain Research, 1115,* 92–107.

Daltrozzo, J., & Schön, D. (2009). Conceptual processing in music as revealed by N400 effects on words and musical targets. *Journal of Cognitive Neuroscience, 21,* 1882–1892.

Faglioni, P., Spinnler, H., & Vignolo, L. A. (1969). Contrasting behavior of right and left hemisphere-damaged patients on a discriminative and a semantic task of auditory recognition. *Cortex, 5,* 366–389.

Frey, A., Marie, C., Prod'Homme, L., Timsit-Berthier, M., Schön, D., & Besson, M. (2009). Temporal semiotic units as minimal meaningful units in music? An electrophysiological approach. *Music Perception, 26,* 247–256.

Holcomb, P. J., & McPherson, W. B. (1994). Event-related brain potentials reflect semantic priming in an object decision task. *Brain and Cognition, 24,* 259–276.

Jackendoff, R. (1991). Musical parsing and musical affect. *Music Perception, 9,* 199–230.

Juslin, P. N., & Laukka, P. (2003). Emotional expression in speech and music: Evidence of cross-modal similarities. *Annals of the New York Academy of Sciences, 1000,* 279–282.

Juslin, P. N., & Västfjäll, D. (2008). All musical emotions are not created equal: The cost of neglecting underlying mechanisms. *Behavioral and Brain Sciences, 31,* 559–575.

Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., & Friederici, A. D. (2004). Music, language and meaning: Brain signatures of semantic processing. *Nature Neuroscience, 7,* 302–307.

Koelsch, S., & Siebel, W. (2005). Towards a neural basis of music perception. *Trends in Cognitive Sciences, 9,* 578–584.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 204,* 203–205.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. II). Cambridge, MA: MIT Press.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. New York: Psychology Press.

Meyer, L. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.

Orgs, G., Lange, K., Dombrowski, J., & Heil, M. (2006). Conceptual priming for environmental sounds and words: An ERP study. *Brain and Cognition, 62,* 267–272.

Orgs, G., Lange, K., Dombrowski, J. H., & Heil, M. (2007). Is conceptual priming for environmental sounds obligatory? *International Journal of Psychophysiology, 65,* 162–166.

Patel, A. (2008). *Music, language, and the brain*. New York: Oxford University Press.

Plante, E., Petten, C. V., & Senkfor, A. J. (2000). Electrophysiological dissociation between verbal and nonverbal semantic processing in learning disabled adults. *Neuropsychologia, 38,* 1669–1684.

Sarfarazi, M., Cave, B., Richardson, A., Behan, J., & Sedgwick, E. M. (1999). Visual event related potentials modulated by contextually relevant and irrelevant olfactory primes. *Chemical Senses, 24,* 145–154.

Schaeffer, P. (1966). *Traité des Objets Musicaux*. Paris: Editions du Seuil.

Schnider, A., Benson, D. F., Alexander, D. N., & Schnider-Klaus, A. (1994). Non-verbal environmental sound recognition after unilateral hemispheric stroke. *Brain, 117,* 281–287.

Van Petten, C., & Rheinfelder, H. (1995). Conceptual relations between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia, 33,* 485–508.

Vignolo, L. A. (1982). Auditory agnosia. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences, 298,* 49–57.

Ystad, S., Kronland-Martinet, R., Schön, D., & Besson, M. (2008). Vers une approche acoustique et cognitive de la sémiotique des objets sonores. In E. Rix & M. Formosa (Eds.), *Vers une sémiotique générale du temps dans les arts* (pp. 73–83). Paris: Ircam-Centre Pompidou/Delatour.

Sound quality assessment of wood for xylophone bars

Mitsuko Aramaki[a)]
CNRS Laboratoire de Mécanique et d'Acoustique 31, chemin Joseph Aiguier 13402 Marseille Cedex 20, France

Henri Baillères and Loïc Brancheriau
CIRAD-Forêt, TA 10/16, avenue Agropolis, 34398 Montpellier Cedex 5, France

Richard Kronland-Martinet and Sølvi Ystad
CNRS, Laboratoire de Mécanique et d'Acoustique 31, chemin Joseph Aiguier 13402 Marseille Cedex 20, France

(Received 15 March 2006; revised 22 January 2007; accepted 22 January 2007)

Xylophone sounds produced by striking wooden bars with a mallet are strongly influenced by the mechanical properties of the wood species chosen by the xylophone maker. In this paper, we address the relationship between the sound quality based on the timbre attribute of impacted wooden bars and the physical parameters characterizing wood species. For this, a methodology is proposed that associates an analysis-synthesis process and a perceptual classification test. Sounds generated by impacting 59 wooden bars of different species but with the same geometry were recorded and classified by a renowned instrument maker. The sounds were further digitally processed and adjusted to the same pitch before being once again classified. The processing is based on a physical model ensuring the main characteristics of the wood are preserved during the sound transformation. Statistical analysis of both classifications showed the influence of the pitch in the xylophone maker judgement and pointed out the importance of two timbre descriptors: the frequency-dependent damping and the spectral bandwidth. These descriptors are linked with physical and anatomical characteristics of wood species, providing new clues in the choice of attractive wood species from a musical point of view. © *2007 Acoustical Society of America.* [DOI: 10.1121/1.2697154]

PACS number(s): 43.75.Kk, 43.66.Jh, 43.60.Uv [NFH] Pages: 2407–2420

I. INTRODUCTION

The mechanical and anatomical properties of woods are of importance for the sound quality of musical instruments. Yet, depending on the role of the wooden elements, these properties may differ. Xylophone sounds are produced by striking wooden bars with a mallet, and thus the mechanical properties of the wood are important. This study is the first step towards understanding what makes the sound of an impacted wooden bar attractive for xylophone makers from a musical point of view. For this purpose, we recorded sounds from a wide variety of wood species to compare their sound quality and relate it to the wood properties. An original methodology is proposed that associates analysis-synthesis processes and perceptual classification analysis. Perceptual classification was performed by a renowned instrument maker.

The xylophone maker community agrees on the choice of wood species. This choice is driven by the sound quality, but other nonacoustically relevant properties are considered as well (e.g., robustness; esthetic aspects). The wood species most used in xylophone manufacturing is *Dalbergia* sp. Several authors have sought to determine which physical characteristics are of importance for the generated sound. In particular, Holz (1996) concluded that an "ideal" xylophone wood bar is characterized by a specific value range of density, Young modulus, and damping factors. Ono and Norimoto (1983) demonstrated that samples of spruce wood (*Picea excelsa, P. glehnii, P. sitchensis*)—considered a suitable material for soundboards—all had a high sound velocity and low longitudinal damping coefficient as compared to other softwoods. The cell-wall structure may account for this phenomenon. Internal friction and the longitudinal modulus of elasticity are markedly affected by the microfibril angle in the S2 tracheid cell layer, but this general trend does not apply to all species. For instance, pernambuco (*Guilandina echinata* Spreng.), traditionally used for making violin bows, has an exceptionally low damping coefficient relative to other hardwoods and softwoods with the same specific modulus (Bucur, 1995; Matsunaga *et al.*, 1996; Sugiyama *et al.*, 1994). This feature has been explained by the abundance of extractives in this species (Matsunaga and Minato, 1998). Obataya *et al.*(1999) confirmed the importance of extractives for the rigidity and damping qualities of reed materials. Matsunaga *et al.* (1999) reduced the damping coefficient of spruce wood by impregnating samples with extractives of pernambuco (*Guilandina echinata* Spreng.). The high sound quality conditions are met by the wood species commonly used by xylophone makers (like *Dalbergia* sp.), but other tropical woods may serve. We propose to focus on the perceptual properties of impacted wood bars as the basis for pointing out woods suitable for xylophone manufacturing. Several studies using natural or synthetic sounds have been conducted to point out auditory clues associated with geom-

[a)] Author to whom correspondence should be addressed. Electronic mail: aramaki@lma.cnrs-mrs.fr etry and material properties of vibrating objects (Avanzini and Rocchesso, 2001; Giordano and McAdams, 2006; Lutfi and Oh, 1997; Klatzky *et al.*, 2000; McAdams *et al.*, 2004). These studies revealed the existence of perceptual clues allowing the source of the impact sound to be identified merely by listening. In particular, the perception of material correlated mainly with the internal friction (related to the damping factors of the spectral components) as theoretically shown by Wildes and Richards (1988). Nevertheless, it has not been determined whether the perceptual clues highlighted in the distinction of different materials are those used to establish the subjective classification of different species of wood.

The perceptual differences reported in the literature are linked with subtle changes in timbre, defined as "the perceptual attribute that distinguishes two tones of equal, pitch, loudness, and duration" (ANSI, 1973). This definition points out the importance of comparing sounds with similar loudness, duration, and pitch. Concerning loudness and duration, the sounds of interest can easily be adjusted in intensity by listening, and they have about the same duration since they correspond to the very narrow category of impacted wooden bars. Concerning pitch, the bars do not have the same values because the pitch depends on the physical characteristics of the wood, i.e., essentially of the Young modulus and the mass density. To tune the sounds to the same pitch, we propose to digitally process the sounds recorded on bars of equal length. Synthesis models can be used for this purpose, allowing virtual tuning by altering the synthesis parameters. Such an approach combining sound synthesis and perceptual analysis has already been proposed. Most of the proposed models are based on the physics of vibrating structures, leading to a modal approach of the synthesis process (Adrien, 1991; Avanzini and Rocchesso, 2001) or to a numerical method of computation (Bork, 1995; Chaigne and Doutaut, 1997; Doutaut *et al.*, 1998). Yet, although these models lead to realistic sounds, they do not easily allow for an analysis-synthesis process implicating the generation of a synthetic sound perceptually similar to an original one. To overcome this drawback, we propose an additive synthesis model based on the physics of vibrating bars, the parameters of which can be estimated from the analysis of natural sounds.

The paper is organized as follows: in Sec. II, we discuss the design of an experimental sound data bank obtained by striking 59 wooden bars made of different woods carefully selected and stabilized in a climatic chamber. In Sec. III, we then address the issue of digitally tuning the sounds without changing the intrinsic characteristics of the wood species. This sound manipulation provided a tuned sound data bank in which each sound was associated with a set of descriptors estimated from both physical experiments and signal analysis. The experimental protocol is described in Sec. IV. It consists of the classification carried by a professional instrument maker. The classification was performed with both the original and the tuned data banks to better understand the influence of pitch on the classification. These results are discussed in Sec. VII, leading to preliminary conclusions that agree with most of the knowledge and usage in both wood mechanics, xylophone manufacturing, and sound perception.

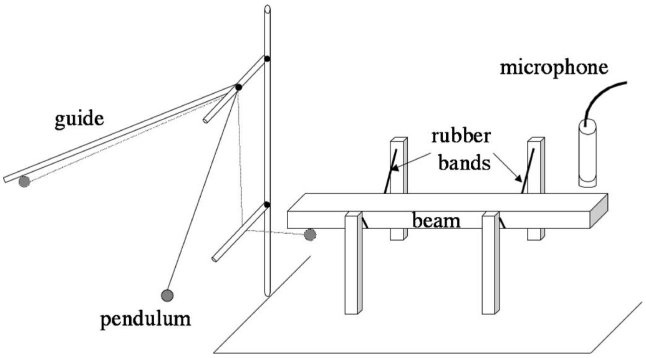


FIG. 1. Experimental setup used to strike the wood samples and record the impact sounds. The setup was placed in an anechoic room.

II. DESIGN OF AN EXPERIMENTAL SOUND DATA BANK

a. Choice of wood species. Most percussive instruments based on wooden bars are made of specific species (for example, *Dalbergia* sp. or *Pterocarpus* sp.). In this experiment, we used tropical and subtropical species, most of which were unknown to instrument makers. A set of 59 species presenting a large variety of densities (from 206 to 1277 $\mathrm{kg/m^3}$) were chosen from the huge collection (about 8000) of the CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement, Montpellier, France). Their anatomical and physical characteristics have been intensely studied and are well described. The name and density of each species are in Table III.

b. Manufacturing wooden bars. Both geometry and boundary conditions govern the vibration of bars. By considering bars with the same geometry and boundary conditions, sounds can be compared to determine the intrinsic quality of the species. Hence, a set of bars was made according to the instrument maker recommendations. The bars were manufactured to be as prismatic as possible, with dimensions $L = 350\ \mathrm{mm} \times W = 45\ \mathrm{mm} \times T = 20\ \mathrm{mm}$, without singularities and cut in the grain direction. We assume that the growth rings are parallel to the tangential wood direction and that their curvature is negligible. The longitudinal direction is collinear to the longitudinal axis of the bars. The bars were stabilized in controlled conditions.

c. Recording of impact sounds under anechoic conditions. An experimental setup was designed that combines an easy way to generate sounds with a relative precision ensuring the repeatability of the measurements, as shown in Fig. 1. In this way, impact excitation was similar for all the impacted bars. Moreover, to minimize the sound perturbations due to the environment, the measurements took place in an anechoic room.

The bar was placed on two rubber bands, ensuring free-free-type boundary conditions. The rubbers minimized perturbations due to suspension (see, for example, Blay *et al.*, 1971 for more details). Bars were struck with a small steel pendulum. The ball on the string was released from a constrained initial position (guide), and after the string wrapped around a fixed rod, the ball struck the bar from underneath. The robustness of this simple procedure showed the radiated sounds were reproducible: the determination error was less than 0.1% for the fundamental frequency and 4.3% for the damping coefficient of the first mode (Brancheriau *et al.*, 2006a). To ensure broad spectral excitation, the ball was chosen to generate a sufficiently short pendulum/bar contact (to be as close as possible to an ideal Dirac source). The excitation spectrum is given by the Fourier transform of the impact force, so that the shorter the impact, the broader the spectrum excitation. For that, a steel ball was used since the modulus of elasticity of steel is much larger than that of wood (the ratio is about 200). This setup makes contact duration between the ball and the bar short (Graff, 1975). This duration was shortened because the impact point was underneath the bar, maximizing the reversion force. After several experiments, a good compromise between speed, short duration, and lack of deformation of the material was obtained with a steel ball of 12 g and a 14 mm diameter, tightened by a 30-cm-long string. The impact point played an important role in the generation of sounds. To prevent the first modes from vanishing, the bar was struck close to one of its extremities (at 1 cm), allowing high frequency modes to develop. An omni-directional microphone (Neumann KM183mt) was placed in the close sound field at the opposite end of the impact location to measure the sound-radiated pressure. This configuration obviates the contribution of the spectral peak generated by the ball, peak which was at about 10 kHz. The sounds were digitally recorded at 48 kHz sampling frequency.

d. Signal characteristics. Figure 2 shows the temporal signal, the spectral representation, and the time-frequency representation of a typical sound obtained experimentally. The temporal signals are characterized by a short onset and a fast decay. Consequently, their durations generally do not exceed 1 s. Their spectra are composed of emergent resonances that do not overlap much. As shown by the time-frequency representation, the damping of these spectral components is frequency dependent, the high frequency components being more heavily damped than the low frequency ones.

III. DESIGN OF TUNED SOUND DATA BANK FOR TIMBRE STUDY

To facilitate comparison of the timbre of sounds generated striking different wood species, their pitch was equalized. In practice, this could have been possible using the same procedure adopted by percussive instrument makers, where the bar geometry is modified removing some substance around the center of the bar to be tuned (Fletcher and Rossing, 1998). This approach comes, however, with the risk of making irreversible mistakes, for example, removing an excessive amount of wood. As an alternative, we propose to digitally tune the pitch of sounds generated striking bars of equal length. Such an approach relies on the plausible assumption that the pitch of our recorded signals is primarily determined by the frequency of the first vibrational mode. In particular, we use a sound synthesis model which allows for sound transformations that are accurate relative to the physical phenomena, as compared to other signal processing approaches such as pitch shifting.

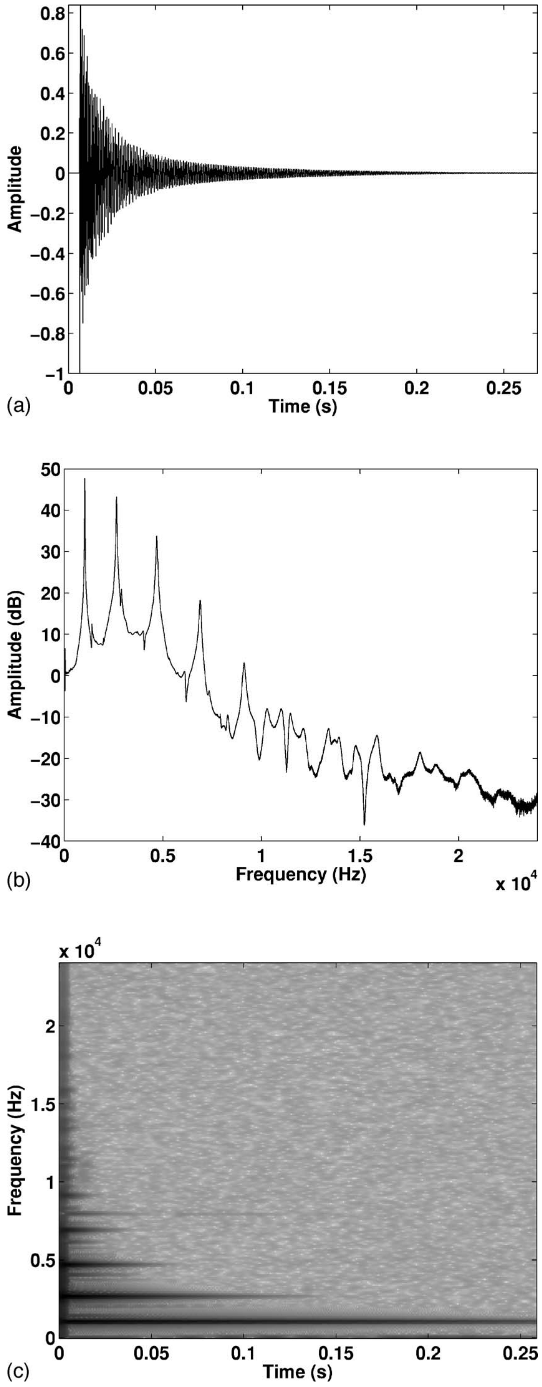


FIG. 2. (a) Wave form, (b) spectral representation, and (c) spectrogram (amplitude in logarithmic scale) of a typical sound obtained by impacting a wooden bar.

A. Synthesis model based on physical approach

To tune the sounds, we propose to use an additive synthesis model. This model simulates the main characteristics of the vibrations produced by an impacted bar to exhibit the principal properties of the radiated sound.

1. Simplified mechanical model

Numerous mechanical models of bar vibrations are available in the literature, but the relevant information can be pointed out using a simple model based on assumptions that are coherent with our experimental design. According to the manufacturing of the bars, one can assume that the fiber orientation follows the axis of the bar and that the ratio length/width is large. Consequently, one can neglect the anisotropy property of the wood and the contribution of the longitudinal and torsional modes (which are few, weak, and of little influence on the radiated sound). These assumptions allow for the consideration of a one-dimensional mechanical model depending only on the longitudinal Young modulus. Such a model can be described by the well-known Euler-Bernoulli equation

$$EI\frac{\partial^4 y(x,t)}{\partial x^4} + \rho S\frac{\partial^2 y(x,t)}{\partial t^2} = 0, \tag{1}$$

where E is the longitudinal Young modulus, I the quadratic moment, ρ the mass density, and S the cross section area. The general solution of the equation is given by

$$y(x,t) = \sum_n Y_n(x)e^{i\gamma_n t} \tag{2}$$

with

$$Y_n(x) = A\cosh(k_n x) + B\sinh(k_n x) + C\cos(k_n x) + D\sin(k_n x). \tag{3}$$

By injecting Eq. (2) and Eq. (3) into the Eq. (1), one obtains

$$\gamma_n = \pm\sqrt{\frac{EI}{\rho S}}k_n^2. \tag{4}$$

Our experimental setup corresponds to free-free boundary conditions written

$$\frac{\partial^2 Y(0)}{\partial x^2} = \frac{\partial^2 Y(L)}{\partial x^2} = \frac{\partial^3 Y(0)}{\partial x^3} = \frac{\partial^3 Y(L)}{\partial x^3} = 0$$

leading to

$$k_n = (2n+1)\frac{\pi}{2L}. \tag{5}$$

To take into account viscoelastic phenomena, E is considered as complex valued, see, for example (Valette and Cuesta, 1993)

$$E = E_d(1 + i\eta), \tag{6}$$

where E_d is the dynamical Young modulus, and η a dimensionless material loss factor. By injecting relations (5) and (6) into relation (4) and assuming that $\eta \ll 1$, one obtains the following important expressions:

$$\gamma_n = \omega_n + i\alpha_n \tag{7}$$

with

$$\begin{cases} \omega_n \approx \sqrt{\dfrac{E_d I}{\rho S}}(2n+1)^2\dfrac{\pi^2}{4L^2} \\ \alpha_n \approx \dfrac{\eta}{2}\omega_n \end{cases}. \tag{8}$$

Thus, one can rewrite the relation (2):

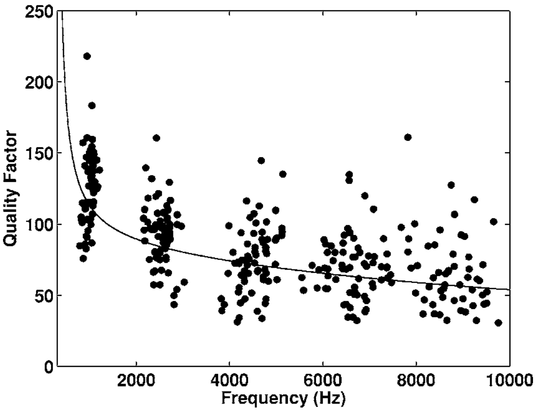


FIG. 3. Quality factor Q_n estimated on original sounds. The Q_n values are fitted, in a least squares sense, by a rational function (black curve) corresponding to Eq. (11).

$$y(x,t) = \sum_n Y_n(x)e^{i\omega_n t}e^{-\alpha_n t}. \tag{9}$$

It is accepted (Chaigne and Doutaut, 1997; McAdams *et al.*, 2004; Ono and Norimato, 1985) that the damping factors in case of wooden bars are described by a parabolic form:

$$\alpha(f) = a_0 + a_2 f^2 \tag{10}$$

where the constants a_0 and a_2 depend on the wood species. This corresponds to a quality factor Q_n given by

$$Q_n = \frac{\pi f_n}{\alpha_n} = \frac{\pi f_n}{a_0 + a_2 f_n^2}. \tag{11}$$

This behavior was experimentally verified, as shown in Fig. 3.

These expressions show that the vibrations of the bar, which are correlated with the radiated sound pressure, can be described by a sum of elementary components consisting of exponentially damped monochromatic signals. The frequency of these elementary components is inversely proportional to the square of the length of the bar, and their damping is proportional to the square of the frequency.

2. Additive synthesis model

The synthesis model aims at simulating the analytical solutions written in Eq. (9), which are expressed as a sum of exponentially damped sinusoids

$$s(x,t) = \theta(t)\sum_{n=1}^{N} A_n(x)\sin(\omega_n t)e^{-\alpha_n t}, \tag{12}$$

where N is the number of components, $\theta(t)$ the Heaviside function, A_n the amplitude, ω_n the frequency and α_n the damping coefficient of the nth component. The choice of either sine or cosine functions has no perceptual influence on the generated sounds but sine functions are often used in sound synthesis since they avoid discontinuities in the signal at $t=0$. Hence, the signal measured at a fixed location is considered to be well represented by the expression (12). Its spectral representation is given by

$$S(\omega) = \sum_{n=1}^{N} \frac{A_n}{2i} \left(\frac{1}{\alpha_n + i(\omega - \omega_n)} - \frac{1}{\alpha_n + i(\omega + \omega_n)} \right)$$

and the z transform by

$$S(z) = \sum_{n=1}^{N} \frac{A_n}{2i} \left(\frac{1}{1 - e^{(i\omega_n - \alpha_n)} z^{-1}} - \frac{1}{1 - e^{(-i\omega_n - \alpha_n)} z^{-1}} \right).$$

B. Estimation of synthesis parameters

Before the tuning process, the recorded sounds described in Sec. II are equalized in loudness, analyzed, and then resynthesized with the synthesis model described above. The loudness was equalized by listening tests. For that, the synthesis parameters are directly estimated from the analysis of the recorded sounds. The estimation of the parameters defining the sound is obtained by fitting the recorded signal with the expression given in relation (12). To do so, we used a signal processing approach that consists of identifying the parameters of a linear filter by auto regressive and moving average (ARMA) analysis. We model the original signal as the output of a generic linear filter whose z transform is written

$$H(z) = \frac{\sum_{m=0}^{M} a_m z^{-m}}{1 + \sum_{n=1}^{N} b_n z^{-n}} = a_0 z^{N-M} \frac{\prod_{m=1}^{M} (z - z_{0m})}{\prod_{n=1}^{N} (z - z_{pn})},$$

where z_{0m} are the zeros and z_{pn} are the poles of the system. Only the most prominent spectral components were modeled by $H(z)$. These spectral components were determined within a 50 dB amplitude dynamic, the reference being the amplitude of the most prominent spectral peak. Hence, the number of poles N and zeros M of the linear ARMA filter is determined by the number of spectral components taken into account. The coefficients a_m and b_n are estimated using classical techniques such as Steiglitz-McBride (Steiglitz and McBride, 1965). The synthesis parameters corresponding to the amplitudes, frequencies, and damping coefficients of the spectral components are thus determined:

$$\begin{cases} A_n = |H(z_{pn})|, \\ \omega_n = \arg(z_{pn}) f_s, \\ \alpha_n = \log(|z_{pn}|) f_s, \end{cases} \tag{13}$$

where f_s is the sampling frequency. In addition to the synthesis model described above, we have taken into account the attack time. Actually, even though the rising time of the sounds is very short, it does influence the perception of the sounds. These rising times were estimated on the original sounds and were reproduced by multiplying the beginning of the synthetic signal by an adequate linear function. Synthesis sounds were evaluated by informal listening tests confirming that their original sound qualities were preserved. The synthesis quality was further confirmed by results from the professional instrument maker showing a similar classification of original and synthetic sounds (classifications Cl and C2, see Sec. VI A 1).

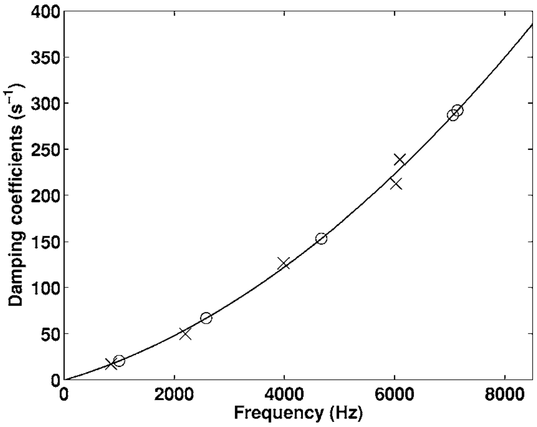


FIG. 4. The damping coefficients of the original sound (×) are fitted by a parabolic function (solid curve). The damping coefficients of the tuned sound (∘) are determined according to this damping law.

C. Tuning the sounds

The processing of tuning the sounds at the same pitch was based on some assumptions specific to the investigated stimulus set and consistent with the vibratory behavior of the bar. For the kind of sounds we are dealing with (impacted wooden bars), we assume the pitch to be related to the frequency of the first vibration mode, which is correlated with the length of the bar [cf. Eq. (8)]. Actually, if the length L changes to βL, then ω_n changes to ω_n / β^2. As a consequence, a change in pitch corresponds to a dilation of the frequency components. These assumptions made it possible to virtually equalize the pitches of the recorded bank of sounds. To minimize the pitch deviation, the whole set of sounds was tuned by transposing the fundamental frequencies to 1002 Hz, which is the mean fundamental frequency of all the sounds. The amplitude of the spectral components was kept unchanged by the tuning process. Once again, no precise listening test was performed, but our colleagues found the synthesis sounds preserved the specificity of the material.

According to the discussion in III A 1, the damping is proportional to the square of the frequency. Thus, from the expression (10), a damping law can be defined by a parabolic function that can be written in a general form:

$$\alpha(\omega) = D_A \omega^2 + D_B \omega + D_C. \tag{14}$$

As a consequence, when the pitch is changed, the damping coefficient of each tuned frequency component has to be evaluated according to the damping law measured on the original sound (cf. Fig. 4).

Figure 5 shows the comparison between the spectrum of a measured signal and the spectrum of a tuned signal after the resynthesis process. The entire sound data bank is available at http://www.lma.cnrs-mrs.fr/~kronland/JASA_Xylophone/sounds.html.

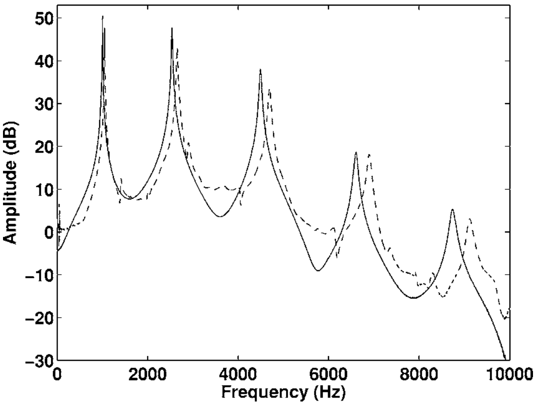


FIG. 5. Comparison between a spectrum of a measured signal (dashed trace) and the spectrum of the associated tuned signal (solid trace).

IV. EXPERIMENTAL PROTOCOL

Sounds from different wooden bars were evaluated focusing on the perceived musical quality of the wood samples. The participant was placed in front of a computer screen on which the sounds (all represented as identical crosses) were randomly distributed. The participant was asked to place the sounds on a bidimensional computer display. In particular, he was told that the horizontal dimension of the display represented an axis of musical quality so that sounds judged as having the worst/best quality were to be placed on the leftmost/rightmost part of the display. The participant could listen to the sounds as often as he wanted by simply clicking on the cross. The tests were carried on a laptop Macintosh equipped with a Sony MDR CD550 headset.

For this study, one instrument maker specialized in xylophone manufacture carried the task. For a complete perceptual study, more participants would, of course, be needed. As a first step we aimed at presenting a new methodology for an interdisciplinary approach uniting instrument makers and specialists within acoustics, signal processing, and wood sciences.

Three tests were conducted using this experimental protocol. The instrument maker carried the classification with the original sounds (recorded sounds with different pitches), called C1 (Brancheriau *et al.*, 2006a; Brancheriau *et al.*, 2006b). A second classification, called C2, using the synthesized sounds (resynthesis of the original sounds with different pitches) was done two years later. The comparison of C1 and C2 allowed us to check the quality of the resynthesis as well as the reliability of our experimental participant. The third test (C3) was carried on the signals tuned to the same pitch. The xylophone maker was not aware of the synthetic nature of sounds in C2 and C3. In particular, he was told that, for classification C3, the same pieces of wood had been sculpted in order to tune the sounds to the same fundamental frequency. Classification C3 is presented in Table III.

V. DESCRIPTORS

A. Mechanical descriptors

The wood species used for this study have been intensively examined at CIRAD and their anatomical and physical characteristics are well known. Thus, the mechanical descriptors are defined by the mass density, ρ, the longitudinal modulus of elasticity, E_ℓ, and the transverse shear modulus, G_t. The descriptors E_ℓ and G_t can be calculated using Timoshenko's model and the Bordonné solutions (Brancheriau and Baillères, 2002). We have also considered the specific longitudinal modulus, E_ℓ/ρ, and the specific shear modulus, G_t/ρ.

B. Signal descriptors

To characterize the sounds from an acoustical point of view, we calculated the following timbre descriptors (Caclin *et al.*, 2005; McAdams *et al.*, 1995): attack time, AT (the way the energy rises during the onset of the sound), spectral bandwidth, SB (spectrum spread), spectral centroid, SCG (brightness), and spectral flux, SF (the way the sound vanishes).

The attack time, AT, a temporal descriptor, characterizes the signal onset and describes the time it takes for the signal to reach its maximum. It is generally estimated as the time it takes the signal to deploy its energy from 10% to 90% of the maximum. The spectral timbre descriptors characterize the organization of the spectral peaks resulting from the modal behavior of the bar vibration. One of the most well known is the spectral centroid, SCG, which is correlated with the subjective sensation of brightness (Beauchamps, 1982):

$$\mathrm{SCG} = \frac{\sum_k f(k)|\hat{s}(k)|}{\sum_k |\hat{s}(k)|}, \tag{15}$$

where $\hat{s}$ is the discrete Fourier transform of the signal $s(t)$ and f the frequency. The spectral bandwidth, SB, measures the spread of the spectral components around the spectral centroid and is defined as (Marozeau, de Cheveigné, McAdams and Winsberg, 2003)

$$\mathrm{SB} = \sqrt{\frac{\sum_k |\hat{s}(k)|(f(k)-\mathrm{SCG})^2}{\sum_k |\hat{s}(k)|}}. \tag{16}$$

Finally, the fourth classical timbre descriptor called the spectral flux, SF, is a spectro-temporal descriptor that measures the deformation of the spectrum with respect to time. In practice, the spectral flux is given by a mean value of the Pearson correlation calculated using the modulus of local spectral representations of the signal (McAdams *et al.*, 1995):

$$\mathrm{SF} = \frac{1}{N}\sum_{n=1}^{N} \frac{\langle s_n, s_{n-1}\rangle}{s_n^2 s_{n-1}^2}, \tag{17}$$

where N represents the number of frames, s_n the modulus of the local spectrum at the discrete time n, and $\langle , \rangle$ the discrete scalar product.

In addition to these well-known timbre descriptors, we propose to consider various acoustical parameters chosen as function of the specificities of the impact sounds, i.e., the amplitude ratio between the first two frequency components of the sound, noted $A_{2/1}$, and the damping and the inharmonicity descriptors. The last two parameters are described below in more detail. The damping descriptor is defined from the Eq. (14) by the set of coefficients $\{D_A, D_B, D_C\}$ traducing the sound decrease. As the damping is the only parameter responsible for the variation of the spectral representation of the signal with respect to time, this descriptor is related to the spectral flux, SF. In addition, the damping coefficients α_1 and α_2 of components 1 and 2 have been included in the list of signal descriptors. The inharmonicity characterizes the relationship between the partials and the fundamental mode. This parameter is linked with the consonance, which is an important clue in the perceptual differentiation of sounds. For each spectral component, inharmonicity is defined by

$$I(n) = \frac{\omega_n}{\omega_0} - n. \tag{18}$$

From this expression, we propose an inharmonicity descriptor defined by a set of coefficients $\{I_A, I_B, I_C\}$ obtained by fitting $I(n)$ with a parabolic function, as suggested by the calculation $I(n)$ from Eq. (8):

$$I(n) = I_A n^2 + I_B n + I_C. \tag{19}$$

VI. RESULTS

Collected behavioral data could be considered as ordinal. Nevertheless, since the task consisted in placing the sounds on a quality axis "as a function of its musical quality," the relative position of the sounds integrates a notion of perceptual distance. Moreover, the classifications do not contain two sounds with the same position and do not show categories (see Table III). It was thus decided to consider the data as providing a quantitative estimate of perceived musical quality for the wood samples, the value associated with each species being given by its abscissa from 0 (worst quality) to 10 (best quality) on the quality axis. The main interest in using quantitative scales is the possibility of constructing an arithmetic model for perceived wood quality which can be easily used to estimate the musical quality of woods (and sounds) not considered in our experiments. All the statistical analyses were conducted with SPSS software (Release 11.0.0, LEAD Technologies).

A. Qualitative analysis—Choice of the variables

1. Resynthesis quality—Robustness of the classification

Only one participant performed the classifications on the basis of his professional skill, and his judgment of sound quality was used to build reference quality scales. The xylophone maker is thus considered as a "sensor" for measuring the acoustical wood quality. The raw classifications C1 and C2 were compared using the Wilcoxon signed rank test to evaluate the resynthesis quality of the model. Moreover, this comparison allowed us to evaluate the robustness of the xylophone maker classification. No particular distribution was assumed for the classifications. The Wilcoxon test is thus appropriate for comparing the distributions of the two classifications (C1, C2). The significance value of the Wilcoxon test (p=0.624) for (C1, C2) indicates that classification C1 equals classification C2. There was no significant difference in the xylophone maker responses between C1 and C2.

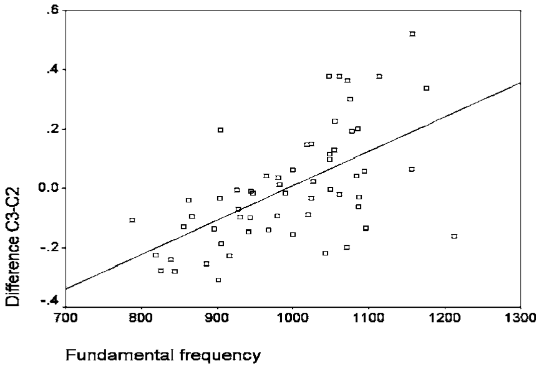


FIG. 6. Linear relationship between fundamental frequency and arithmetic difference C3-C2 (R=0.59, N=59).

2. Influence of the tuning process

The same Wilcoxon signed rank test was performed with classification C2 and classification C3 of tuned sounds. The hypothesis of equal distribution is rejected considering classifications C2 and C3. A significant difference between C2 and C3 (p=0.001) is due to the tuning process of sounds, which altered the sound perception of the xylophone maker. The arithmetic difference (C3-C2) was thus computed and related to the value of the fundamental frequency by using the Pearson correlation coefficient (Fig. 6). This coefficient value was found significant at the 1% level (R=0.59).

B. Quantitative analysis

1. Descriptor analysis

The 18 parameters presented in Table I were estimated for the tuned sounds and using standard mechanical calibrations. They are grouped into mechanical/physical descriptors and signal descriptors. In practice, for the spectral descriptors, the Fourier transform was estimated using a fast Fourier transform (FFT) algorithm. The length of the FFT was chosen so that it matches the longest sound, i.e., 2^{16} samples. For the SF calculation, the number of samples was 256 with an overlap of 156 samples. A Hamming window was used to minimize the ripples. Mechanical descriptors are linked with the intrinsic behavior of each sample but also linked with signal descriptors, as shown in Fig. 7. Indeed, the bivariate coefficients of determination matrix calculated on the basis of the 18 characteristic parameters revealed close collinearity between the parameters. Considering the strong relationship between the parameters, the statistical analyses were conducted by grouping the mechanical/physical descriptors and the signal descriptors in order to find those that best explain the classification C3.

A principal component analysis was thus conducted (Table II). Principal components analysis finds combinations of variables (components) that describe major trends in the data. This analysis generated a new set of parameters derived from the original set in which the new parameters (principal

TABLE I. Mechanical and signal descriptors computed from dynamic tests.

	No.	Variable	Signification
Mechanical descriptors	1	ρ	Mass density (kg/m^3)
	2	E_ℓ	Longitud. modulus of elasticity (MPa)
	3	G_t	Shear modulus (MPa)
	4	E_ℓ/ρ	Specific longitudinal modulus
	5	G_t/ρ	Specific shear modulus
Signal descriptors	6	$A_{2/1}$	Amplitude ratio of mode 2 and 1
	7	α_1	Temporal damping of mode 1 (s^{-1})
	8	α_2	Temporal damping of mode 2 (s^{-1})
	9	SCG	Spectral centroid (Hz)
	10	SB	Spectral bandwidth (Hz)
	11	SF	Spectral flux
	12	AT	Attack time (ms)
	13	D_A	Coefficient D_A of $\alpha(\omega)$
	14	D_B	Coefficient D_B of $\alpha(\omega)$
	15	D_C	Coefficient D_C of $\alpha(\omega)$
	16	I_A	Coefficient I_A of $I(n)$
	17	I_B	Coefficient I_B of $I(n)$
	18	I_C	Coefficient I_C of $I(n)$

components) were not correlated and closely represented the variability of the original set. Each original parameter was previously adjusted to zero mean and unit variance so that eigenvalues could be considered in choosing the main factors. In this case, the eigenvalues sum the number of variables, and eigenvalues can be interpreted as the number of original variables represented by each factor. The principal components selected thus corresponded to those of eigenvalue superior or equal to unity. Table II shows that six principal components accounted for 87% of all information contained in the 18 original parameters.

The relationships between original variables and principal components are presented in Figs. 8(a) and 8(b). These figures display the bivariate coefficient of determination between each principal component and each original parameter; the bivariate coefficient corresponds to the square loading coefficient in this analysis. The variance of the inharmonicity coefficients $\{I_A, I_B, I_C\}$ and the damping coefficients $\{D_A, D_B, D_C\}$ are captured by the first principal component and to a lesser degree by the third component [Fig. 8(a)]. The damping coefficients (α_1 and α_2), however, are mainly linked with the second component. This component is also linked with the amplitude ratio $A_{2/1}$ and with the timbre descriptors (SCG, SB, SF, AT). The variance of the mechanical/physical descriptors is scattered between all the principal components (parameter 1 is linked with PC1 and 2; parameter 2 with PC1 and 4; parameter 3 with PC3 and 5; parameter 4 with PC2, 3, and 4; and parameter 5 with PC3 and 5).

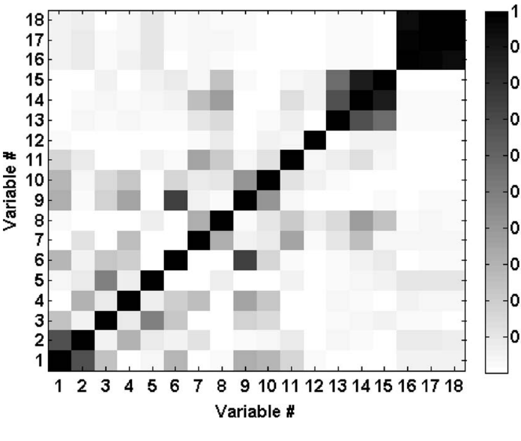


FIG. 7. Bivariate coefficients of determination for characteristic parameters (N=59).

2. Relationship between the descriptors and the acoustic classification of tuned sounds

a. Bivariate analysis. Figure 9 presents the results of bivariate analysis between characteristic parameters and classification C3. Assuming a linear relationship, the parameter α_1 (temporal damping of mode 1) appeared to be the best individual predictor with a R^2 value of 0.72. The second most significant predictor was the spectral flux, SF, with a R^2 value of 0.38. The other parameters were of minor importance considering classification C3. Note that the only mechanical parameter of interest was E_ℓ/ρ (specific longitudinal modulus) with a relatively low R^2 value of 0.25. Furthermore, the mass density, ρ, was not reflected in the acoustic classification (no significant R^2 value at the 1% level). Light woods and heavy woods were thus not differ-

TABLE II. Variance explained by the principal components (number of initial variables=18, number of samples=59).

Component	Eigen val.	% of Var.	Cumul. (%)
I	4.0	22.5	22.5
II	3.9	21.9	44.3
III	3.5	19.3	63.7
IV	1.8	10.1	73.8
V	1.2	6.7	80.5
VI	1.1	6.1	86.6

TABLE III. Botanical names of wood species (N=59), their density (kg/m^3), α_1 the temporal damping of mode 1(s^{-1}), SB the spectral bandwidth (Hz) and classification C3 by the xylophone maker (normalized scale from 0 to 10).

Botanical name	Density (kg/m^3)	α_1(s^{-1})	SB (Hz)	C3
Pericopsis elata Van Meeuw	680	21.76	2240	5.88
Scottellia klaineana Pierre	629	23.97	2659	6.38
Ongokea gore Pierre	842	26.07	2240	5.15
Humbertia madagascariensis Lamk.	1234	28.84	3820	0.48
Ocotea rubra Mez	623	23.47	2521	5.42
Khaya grandifoliola C.DC.	646	33.02	2968	0.95
Khaya senegalensis A. Juss.	792	33.98	3101	0.33
Coula edulis Baill.	1048	27.6	2674	2.1
Tarrietia javanica Bl.	780	20.33	2198	9.15
Entandrophragma cylindricum Sprague	734	30.6	2592	1.12
Afzelia pachyloba Harms	742	20.56	2048	8.24
Swietenia macrophylla King	571	20.99	1991	9.22
Aucoumea klaineana Pierre	399	32.17	2275	1.81
Humbertia madagascariensis Lamk	1277	23.36	3171	3.48
Faucherea thouvenotii H. Lec.	1061	20.18	2512	6.05
Ceiba pentandra Gaertn.	299	29.16	2396	2.57
Letestua durissima H. Lec.	1046	19.56	2770	3.87
Monopetalanthus heitzii Pellegr.	466	23.98	2344	5.57
Commiphora sp.	390	16.52	1269	9.77
Dalbergia sp.	916	14.29	2224	9.79
Hymenolobium sp.	600	20.58	2402	7.86
Pseudopiptadenia suaveolens Brenan	875	20.8	1989	6.53
Parkia nitida Miq.	232	26.86	1440	5.75
Bagassa guianensis Aubl.	1076	20.68	2059	6.82
Discoglypremna caloneura Prain	406	34.27	1506	1.38
Brachylaena ramiflora Humbert	866	21.85	2258	4.71
Simarouba amara Aubl.	455	21.26	1654	9.37
Gossweilerodendron balsamiferum Harms	460	35.26	1712	1.08
Manilkara mabokeensis Aubrev.	944	23.89	1788	3.25
Shorea-rubro squamata Dyer	569	23.9	1604	6.75
Autranella congolensis A. Chev.	956	38.97	3380	0.35
Entandrophragma angolense C. DC.	473	22.79	1612	7.67
Distemonanthus benthamianus Baill.	779	19.77	2088	8.75
Terminalia superba Engl. & Diels	583	21.89	2004	9.32
Nesogordonia papaverifera R.Cap.	768	27.96	2097	2.37
Albizia ferruginea Benth.	646	24.71	2221	4.32
Gymnostemon zaizou. Aubrev. & Pellegr.	380	30.15	2130	1.83
Anthonotha fragrans Exell & Hillcoat	777	24.87	1926	4.2
Piptadeniastrum africanum Brenan	975	22.41	3226	3.68
Guibourtia ehie J. Leon.	783	26.36	2156	4.05
Manilkara huberi Standl.	1096	35.11	2692	0.77
Pometia pinnata Forst.	713	25.5	1835	6.23
Glycydendron amazonicum Ducke	627	20.41	2292	7.91
Cunonia austrocaledonica Brong. Gris.	621	31.05	3930	0.59
Nothofagus aequilateralis Steen.	1100	37.76	3028	0.18
Schefflera gabriellae Baill.	570	28.16	1872	1.42
Gymnostoma nodiflorum Johnst.	1189	33	3013	1.26
Dysoxylum sp.	977	23.85	2106	4.49
Calophyllum caledonicum Vieill.	789	19.82	2312	8.66
Gyrocarpus americanus Jacq.	206	38.39	1982	0.6
Pyriluma sphaerocarpum Aubrev.	793	30.83	2318	1.23
Cedrela odorata L.	512	30.45	2070	3
Moronobea coccinea Aubl.	953	21.67	1781	4.92
Goupia glabra Aubl.	885	45.61	2525	0.22
Manilkara huberi Standl.	1187	22.6	2917	2.78
Micropholis venulosa Pierre	665	22.51	3113	7.12
Cedrelinga catenaeformis Ducke	490	22.5	1626	7.31
Vouacapoua americana Aubl.	882	23.18	1986	6.88
Tarrietia Densiflora Aubrev & Normand	603	29.76	2326	1.62

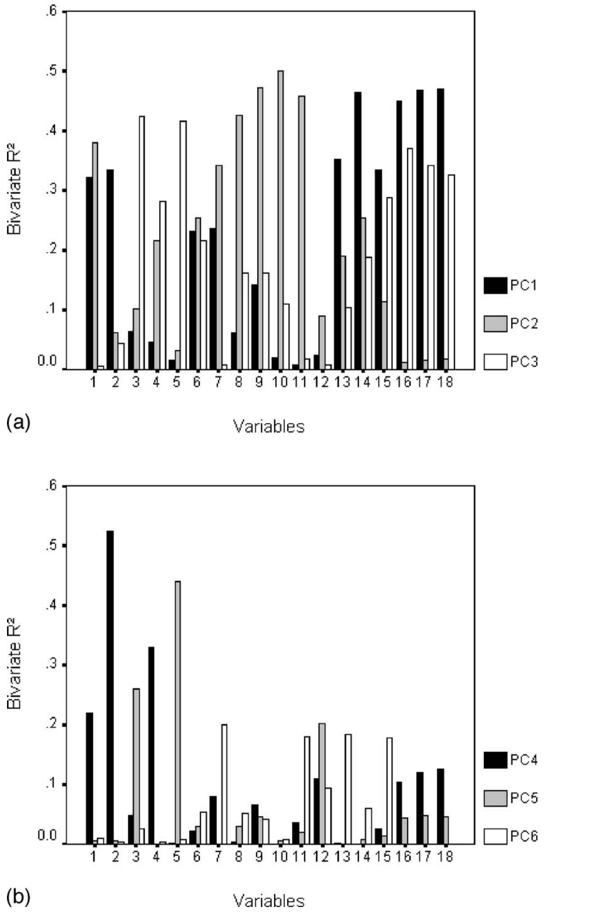


FIG. 8. Bivariate determination coefficient between original variables and principal components: (a) for PC1, PC2 and PC3; (b) for PC4, PC5 and PC6.

entiated by the xylophone maker in the acoustic classification.

b. Multivariate linear regression analysis. The second step of the analysis was to build a robust linear model to take into account the most significant predictors. The robustness of the model assumes that no multicollinearity among the variables exists (Dillon and Goldstein, 1984). The stepwise selection method was thus used to perform multivariate analysis. This method enters variables into the model one by one and tests all the variables in the model for removal at each step. Stepwise selection is designed for the case of correlations among the variables. Other automatic selection procedures exist (forward selection and backward elimination, for example), and the models obtained by these methods may differ, especially when independent variables are highly intercorrelated. Because of the high correlation between variables, several regression models almost equally explain classification C3. However, stepwise selection was used to build one of the most significant models with noncorrelated variables relating to different physical phenomena.

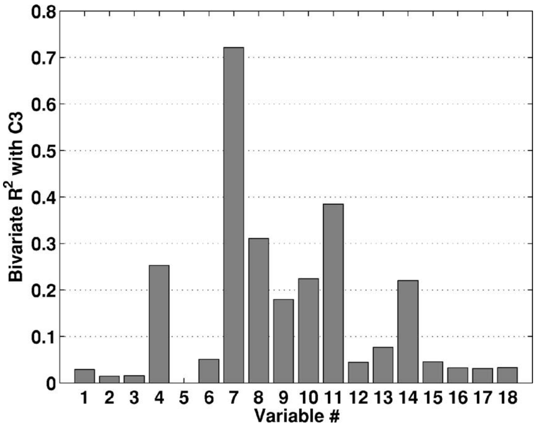


FIG. 9. Bivariate coefficients of determination between characteristic parameters and classification C3 (N=59).

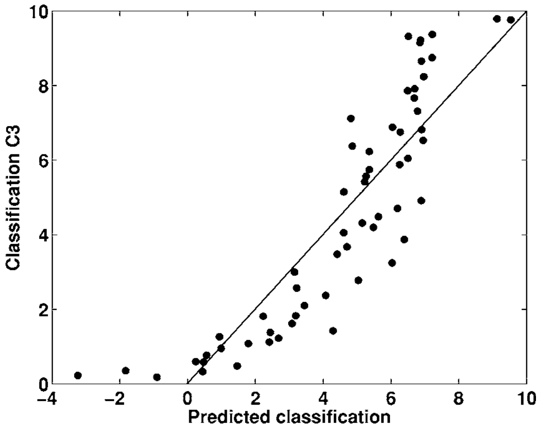


FIG. 10. Predicted vs observed C3 classification (linear predictors α_1 and SB, R^2=0.77, N=59).

The final linear model obtained by stepwise variable selection included the two predictors, α_1 and SB. The predicted classification is given by:

$$\hat{C}3_{\text{Linear}} = -3.82 \times 10^{-1}\alpha_1 - 1.32 \times 10^{-3} SB + 17.52. \tag{20}$$

The multiple coefficient of determination was highly significant (R^2=0.776 and Adjusted R^2=0.768, Fig. 10) and each regression coefficient was statistically different from zero (significance level: 1%). The predictor α_1 was predominant in the model with a partial coefficient value of $R_{\alpha 1}$= −0.84 (R_{SB}=−0.44). The negative sign of $R_{\alpha 1}$ showed that samples with high damping coefficients were associated with a poor acoustic quality.

Partial least squares regression showed that the damping coefficient α_1 was predominant in the model (Brancheriau *et al.*, 2006b). However, the physical significance of the partial least squares model was difficult to explain because the original variables were grouped in latent variables. The stepwise procedure was thus used to better understand the regression results.

The multivariate analysis differed from the bivariate analysis by the replacement of SF by SB, because the selected set of predictors was formed by noncorrelated variables. SB was thus selected because of the low correlation between α_1 and SB with a coefficient value of $R_{\alpha 1/\text{SB}}$=0.29 instead of SF with a value of $R_{\alpha 1/\text{SF}}$=−0.60.

Principal components regression (PCR) was another way to deal with the problem of strong correlations among the variables. Instead of modeling the classification with the variables, the classification was modeled on the principal component scores of the measured variables (which are orthogonal and therefore not correlated). The PCR final model was highly significant with a multiple R^2 value of 0.741 and Adjusted R^2 value of 0.721. Four principal components were selected and the resulting scatter plot was similar to the one in Fig. 10. Comparing the two multivariate models, we found the PCR model to be less relevant than the stepwise one. The R^2 of the PCR model was indeed lower than the R^2 of the stepwise model. Furthermore, the PCR model included four components while only two independent variables were included in the stepwise model. The difference between these two models was explained by the fact that the whole information contained in the characteristic parameters (Table I) was not needed to explain the perceptual classification. The PCR procedure found components that capture the greatest amount of variance in the predictor variables, but did not build components that both capture variance and achieve correlation with the dependent variable.

c. Multivariate nonlinear regression analysis. The configuration of points associated with the linear model (C3, α_1 and SB) in Fig. 10 indicated a nonlinear relationship. This was particularly true for samples of poor acoustic quality (negative values of the standardized predicted classification). As a final step of the analysis, we built a nonlinear model of the behavioral response. In particular, we transformed the values predicted by the linear model $\hat{C}3_{\text{Linear}}$ using a sigmoidal transform. Such transform was consistent with the relationship between C3 and $\hat{C}3_{\text{Linear}}$ (see Fig. 10). The fitting coefficients were extracted via the Levenberg-Marquardt optimization procedure by minimizing the residual sum of squares (dependent variable C3 and independent variable $\hat{C}3_{\text{Linear}}$: predicted classification with the linear modeling). The final equation is written as follows:

$$\hat{C}3_{\text{sigmoid}} = \frac{10}{1 + e^{-\frac{\hat{C}3_{\text{Linear}}-5}{1.64}}} \tag{21}$$

with $\hat{C}3_{\text{Linear}}$ defined by Eq. (20). The multiple coefficient of determination was highly significant (R^2=0.82) and each nonlinear regression coefficient was statistically different from zero (significance level: 1%). The nonlinear model provided a better fit than the linear model; moreover no apparent systematic feature appeared, indicating that residuals were randomly distributed (Fig. 11).

VII. DISCUSSION

In this section, we discuss the main results presented above, attempting to better understand the sound descriptors' influence on the xylophone maker classification. Further on, we discuss the influence of the pitch and the relationship between the wood anatomy and the produced sounds.

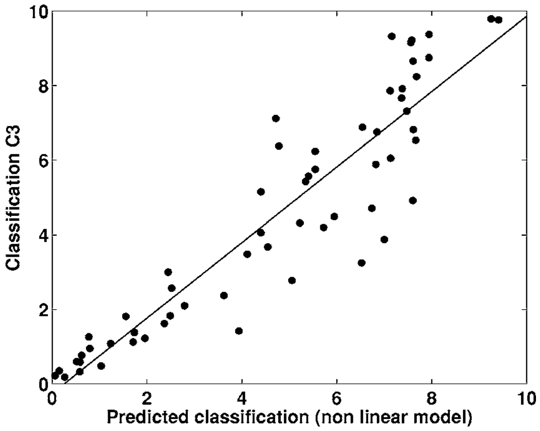


FIG. 11. Predicted vs observed C3 classification (nonlinear predictors α_1 and SB, R^2=0.82, N=59).

A. On the reliability of the xylophone maker

As we pointed out in the introduction, this paper does not aim to give categorical clues for choosing interesting species of wood for xylophone manufacturing. Nevertheless, note that these first conclusions probably accurately reflect what xylophone makers look for. Although we tested our methodology with only one renowned xylophone maker, the results show that:

- In accordance with the xylophone maker community, our maker chose *Dalbergia* sp. as the best species. Moreover, this choice was confirmed on both tuned and original sound classifications.
- The comparison of classifications C1 and C2 showed no significant differences according to the Wilcoxon test.

These observations confirm the good reliability of our xylophone maker and the accuracy of the results, which were further informally confirmed by both instrument makers and musicians.

B. Relation between descriptors and wood classification

The classification by the xylophone maker is correlated with several descriptors. Those that play an important role are three descriptors related to the time course of the sound (α_1, α_2 and SF) and two descriptors related to the spectral content of the sound (SCG and SB). Note that the physical descriptors linked with the wood properties do not explain by themselves the classification of the instrument maker, even though E_ℓ/ρ seems to be the most pertinent one. The relatively low importance of the specific modulus regarding classification C3 could be explained by its high correlation with the fundamental frequency (R^2=0.91) and its low correlation with the temporal damping coefficient α_1 (R^2=0.26). Most of the descriptors are correlated; these correlations are coherent with the physics and are shown in a qualitative way in Fig. 7. Both coefficients of the polynomial decomposition of $\alpha(\omega)$ are strongly correlated. So are the coefficients of the polynomial decomposition of $I(n)$. This finding points out the relative consistency in the behavior of the damping and the inharmonicity laws with respect to the species. Parameters α_1 and α_2 are also correlated, showing the monotonic behavior of the damping with respect to the frequency: the higher the frequency, the higher the damping. As a consequence, both α_1 and α_2 are correlated with the spectral flux, SF, since these descriptors are the only ones that relate to the time course of the sound.

Both global spectral descriptors, SCG and SB, are also correlated, showing that their increase is strongly related to the adjunction of high frequency energy. These descriptors are in addition correlated with the ratio $A_{2/1}$ and with the physical descriptors ρ and E_ℓ/ρ. This correlation can be explained by the way the energy is distributed through the excited modes. Actually, assuming that the bars are impacted identically (good reproducibility of the impact in the experimental setup), the initial energy injected depends on the impedance of each bar. Since the bars were impacted in the transversal direction, one can assume that the transversal Young modulus of elasticity together with the mass density are the main parameters in the difference of amplitudes of modes 1 and 2.

The multivariate linear regression analysis highlighted two main descriptors: α_1 and SB. These descriptors are noncorrelated and give rise to a linear predictor of the classification $\hat{C}3_{\text{Linear}}$ that explains 77% of the variance. This model is of great importance in the choice of species. Actually, it emphasizes the fact that the xylophone maker looks for a highly resonant sound (the coefficient of α_1 is negative) containing a few spectral components (the coefficient of SB is also negative). Such a search for a crystal-clear sound could explain the general choice of *Dalbergia* sp., which is the most resonant species and the most common in xylophone bars. Indeed, the predominance of α_1 agrees with the first rank of *Dalbergia* sp., for which $\alpha_1 = 14.28\ \text{s}^{-1}$ is the smallest in the data bank ($14.28\ \text{s}^{-1} < \alpha_1 < 45.61\ \text{s}^{-1}$) and SB $= 2224$ Hz is medium range in the data bank (1268 Hz $<$ SB $<$ 3930 Hz). Holz (1996) showed that the damping factor value α_1 should be lower than about 30 s^{-1} for a fundamental frequency value of 1000 Hz, which corresponds to the mean value of the study. The average value of α_1 is indeed 26.13 s^{-1} with a standard deviation of 6.18 s^{-1}. Actually, xylophone makers use a specific way of carving the bar by removing substance in the middle (Fletcher and Rossing, 1998). This operation tends to minimize the importance of partial 2, decreasing both the SCG and the SB. The importance of α_1 in the model is in line with several studies showing that the damping is a pertinent clue in the perception of impacted materials (Klatzky *et al.*, 2000; Wildes and Richards, 1988). Concerning parameter SB, the spectral distribution of energy is also an important clue, especially for categorization purposes.

The linear classification prediction has been improved by taking into account nonlinear phenomena. The nonlinear model then explains 82% of the variance. The nonlinear relationship between the perceptual classification and predictors (α_1 and SB) was explained by the instrument maker's strategy during the evaluation of each sample. The xylophone maker proceeded by first identifying the best samples and then the worst samples. This first step gave him the upper and lower bounds of the classification. The final step was to sort the samples of medium quality and place them between the bounds. One could deduce that three groups of acoustic quality (good, poor, and medium quality) were formed before the classification and that inside these groups the perceptual distance between each sample was different. The sigmoid shape indicated that the perceptual distance was shorter for good and poor quality groups than for medium quality groups. As a consequence, the nonlinear model is probably linked with the way the maker proceeded and cannot be interpreted as an intrinsic model for wood classification. Another explanation for the nonlinear relationship can also be found in the nonlinear transform relating physical and perceptual dimensions.

Note finally that there was no correlation between the classification and the wood density. However it is known that the wood density is of great importance for instrument makers. Holz (1996) suggested that the "ideal" xylophone wood bars would have density values between 800 and 950 kg/m^3. This phenomenon is due to the way we designed our experimental protocol, focusing on the sound itself and minimizing multi-sensorial effects (avoiding the access to visual and tactile information). Actually, in a situation where the instrument maker has access to the wood, bars with weak density are rejected for manufacturing and robustness purposes, irrespective of their sound quality.

C. Influence of the fundamental frequency (pitch) on the classification

As discussed previously, timbre is a key feature for appreciating sound quality and it makes it possible to distinguish tones with equal pitch, loudness, and duration (ANSI, 1973). Since this study aims at better understanding which timbre descriptor is of interest for wood classification, one expected differences in the classification of the tuned and the original sound data banks. The difference between classifications C2 (various pitches) and C3 (same pitches) shows a clear linear tendency; it is represented in Fig. 6 as a function of the original fundamental frequency of the bars. The difference is negative (respectively positive) for sounds whose fundamental frequencies are lower (respectively higher) than the mean frequency. The Pearson coefficient associated with the linear relationship between the arithmetic difference of the classification and the fundamental frequency leads to the important observation that *a wooden bar with a low fundamental frequency tends to be upgraded while a wooden bar with a high fundamental frequency tends to be downgraded*. This finding agrees with our linear prediction model, which predicts weakly damped sounds would be better classified than highly damped ones. Actually, sounds with low (respectively high) fundamental frequencies were transposed toward high (respectively low) frequencies during the tuning process, implying α_1 increase (respectively decrease), since the damping is proportional to the square of the frequency (cf. Sec. III C). As an important conclusion, one may say that the instrument maker cannot judge the wood itself independently of the bar dimensions, since the classification is influenced by the pitch changes, favoring wood samples generating low fundamental frequency sounds.

Once again, note the good reliability of our instrument maker, who did not change the classification of sounds whose fundamental frequency was close to the mean fundamental frequency of the data bank (i.e., sounds with nearly unchanged pitch). Actually, the linear regression line passes close to 0 at the mean frequency 1002 Hz. Moreover, the *Dalbergia* sp. was kept at the first position after the tuning process, suggesting that no dramatic sound transformations had been made. In fact, this sample was transposed upwards by 58 Hz, changing α_1 from 13.6 s^{-1} to 14.28 s^{-1}, which still was the smallest value of the tuned data bank.

D. Relationship between wood anatomy and perceived musical quality

The damping α_1 of the first vibrational mode was an important descriptor explaining the xylophone maker classification. Equation (11) shows that this descriptor is related to the quality factor Q, and consequently to the internal friction coefficient tan ϕ (inverse of the quality factor Q), which depends on the anatomical structure of the wood. An anatomical description of the best classified species has been discussed in a companion article (Brancheriau *et al.*, 2006b). We briefly summarize the main conclusions and refer the reader to the article for more information. A draft anatomical portrait of a good acoustic wood could be drawn up on the basis of our analysis of wood structures in the seven acoustically best and seven poorest woods. This portrait should include a compulsory characteristic, an important characteristic, and two or three others of lesser importance. The key trait is the axial parenchyma. It should be paratracheal, and not very abundant if possible. If abundant (thus highly confluent), the bands should not be numerous. Apotra-cheal parenchyma can be present, but only in the form of well-spaced bands (e.g., narrow marginal bands). The rays (horizontal parenchyma) are another important feature. They should be short, structurally homogeneous but not very numerous. The other characteristics are not essential, but they may enhance the acoustic quality. These include:

- Small numbers of vessels (thus large);
- A storied structure;
- Fibers with a wide lumen (or a high flexibility coefficient, which is the ratio between the lumen width and the fiber width; it is directly linked with the thickness of the fiber).

These anatomical descriptions give clues for better choosing wood species to be used in xylophone manufacturing. They undoubtedly are valuable for designing new musical materials from scratch, such as composite materials.

VIII. CONCLUSION

We have proposed a methodology associating analysis-synthesis processes and perceptual classifications to better understand what makes the sound produced by impacted wooden bars attractive for xylophone makers. This methodology, which focused on timbre-related acoustical properties, requires equalization of the pitch of recorded sounds. Statistical analysis of the classifications made by an instrument maker highlighted the importance of two salient descriptors: the damping of the first partial and the spectral bandwidth of the sound, indicating he searched for highly resonant and crystal-clear sounds. Moreover, comparing the classifications of both the original and processed sounds showed how the pitch influences the judgment of the instrument maker. Indeed, sounds with originally low (respectively high) fundamental frequency were better (lesser) classified before the tuning process than after. This result points to the preponderance of the damping and reinforces the importance of the pitch manipulation to better dissociate the influence of the wood species from that of the bar geometry. Finally, the results revealed some of the manufacturers' strategies and pointed out important mechanical and anatomical characteristics of woods used in xylophone manufacturing. From a perceptual point of view, the internal friction seems to be the most important characteristic of the wood species. Nevertheless, even though no correlation has been evidenced between the classification and the wood density, it is well known that this parameter is of great importance for instrument makers as evidence of robustness. As mentioned in the introduction, this work was the first step towards determining relations linking sounds and wood materials. Future works will aim at confirming the results described in this paper by taking into account classifications made by other xylophone makers in the statistical analysis. We plan to use this methodology on a new set of wood species having mechanical and anatomical characteristics similar to those well classified in the current test. This should point out unused wood species of interest to musical instrument manufacturers and will give clues for designing new musical synthetic materials.

ACKNOWLEDGMENTS

The authors thank Robert Hébrard, the xylophone maker who performed the acoustic classification of the wood species. They are also grateful to Pierre Détienne for useful advice and expertise in wood anatomy. They also thank Bloen Metzger and Dominique Peyroche d'Arnaud for their active participation in the experimental design and the acoustical analysis, and Jérémy Marozeau who provided the graphical interface for the listening test. We would also thank the reviewers for useful suggestions.

American National Standards Institute (**1973**). *American National Standard Psychoacoustical Terminology* (American National Standards Institute, NY).

Adrien, J. M. (**1991**). *The Missing Link: Modal Synthesis* (MIT Press, Cambridge, MA), Chap. 8, pp. 269–297.

Avanzini, F., and Rocchesso, D. (**2001**). "Controlling material properties in physical models of sounding objects," in *Proceedings of the International Computer Music Conference 2001*, 17–22 September 2001, Hawana, pp. 91–94.

Beauchamps, J. W. (**1982**). "Synthesis by spectral amplitude and "brightness" matching of analyzed musical instrument tones," J. Audio Eng. Soc. **30**(6), 396–406.

Blay, M., Bourgain , and Samson (**1971**). "Application des techniques électroacoustiques à la détermination du module d'élasticité par un procédé nondestructif (Application of electroacoustic techniques to determine the elasticity modulus by nondestructive procedure)," Technical Review to Advance Techniques in Acoustical, Electrical and Mechanical Measurement **4**, 3–19.

Bork, I. (**1995**). "Practical tuning of xylophone bars and resonators," Appl. Acoust. **46**, 103–127.

Brancheriau, L., and Baillères, H. (**2002**). "Natural vibration analysis of clear wooden beams: A theoretical review," Wood Sci. Technol. **36**, 347–365.

Brancheriau, L., Baillères, H., Détienne, P., Gril, J., and Kronland-Martinet, R. (**2006a**). "Key signal and wood anatomy parameters related to the acoustic quality of wood for xylophone-type percussion instruments," J. Wood Sci. **52**(3), 270–274.

Brancheriau, L., Baillères, H., Détienne, P., Kronland-Martinet, R., and Metzger, B. (**2006b**). "Classifying xylophone bar materials by perceptual, signal processing and wood anatomy analysis," Ann. Forest Sci. **62**, 1–9.

Bucur, V. (**1995**). *Acoustics of Wood* (CRC Press, Berlin).

Caclin, A., McAdams, S., Smith, B. K., and Winsberg, S. (**2005**). "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones," J. Acoust. Soc. Am. **118**(1), 471–482.

Chaigne, A., and Doutaut, V. (**1997**). "Numerical simulations of xylophones. I. Time-domain modeling of the vibrating bars," J. Acoust. Soc. Am. **101**(1), 539–557.

Dillon, W. R., and Goldstein, M. (**1984**). *Multivariate Analysis—Methods and Applications* (Wiley, New York).

Doutaut, V., Matignon, D., and Chaigne, A. (**1998**). "Numerical simulations of xylophones. II. Time-domain modeling of the resonator and of the radiated sound pressure," J. Acoust. Soc. Am. **104**(3), 1633–1647.

Fletcher, N. H., and Rossing, T. D. (**1998**). *The Physics of Musical Instruments*, 2nd ed. (Springer-Verlag, Berlin).

Giordano, B. L., and McAdams, S. (**2006**). "Material identification of real impact sounds: Effects of size variation in steel, wood, and Plexiglass plates," J. Acoust. Soc. Am. **119**(2), 1171–1181.

Graff, K. F. (**1975**). *Wave Motion in Elastic Solids* (Ohio State University Press), pp. 100–108.

Holz, D. (**1996**). "Acoustically important properties of xylophon-bar materials: Can tropical woods be replaced by European species?" Acust. Acta Acust. 82(6), 878–884.

Klatzky, R. L., Pai, D. K., and Krotkov, E. P. (**2000**). "Perception of material from contact sounds," Presence: Teleoperators and Virtual Environments **9**(4), 399–410.

Lutfi, R. A., and Oh, E. L. (**1997**). "Auditory discrimination of material changes in a struck-clamped bar," J. Acoust. Soc. Am. **102**(6), 3647–3656.

Marozeau, J., de Cheveigné, A., McAdams, S., and Winsberg, S. (**2003**). "The dependency of timbre on fundamental frequency," J. Acoust. Soc. Am. **114**, 2946–2957.

Matsunaga, M., and Minato, K. (**1998**). "Physical and mechanical properties required for violin bow materials II. Comparison of the processing properties and durability between pernambuco and substitutable wood species," J. Wood Sci. **44**(2), 142–146.

Matsunaga, M., Minato, K., and Nakatsubo, F. (**1999**). "Vibrational property changes of spruce wood by impregnating with water-soluble extractives of pernambuco (*Guilandina echinata Spreng.*)," J. Wood Sci. **45**(6), 470–474.

Matsunaga, M., Sugiyama, M., Minato, K., and Norimoto, M. (**1996**). "Physical and mechanical properties required for violin bow materials," Holzforschung **50**(6), 511–517.

McAdams, S., Chaigne, A., and Roussarie, V. (**2004**). "The psychomechanics of simulated sound sources: Material properties of impacted bars," J. Acoust. Soc. Am. **115**(3), 1306–1320.

McAdams, S., Winsberg, S., Donnadieu, S., Soete, G. D., and Krimphoff, J. (**1995**). "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," Psychol. Res. **58**, 177–192.

Obataya, E., Umewaza, T., Nakatsubo, F., and Norimoto, M. (**1999**). "The effects of water soluble extractives on the acoustic properties of reed (*Arundo donax* L.)," Holzforschung **53**(1), 63–67.

Ono, T., and Norimoto, M. (**1983**). "Study on Young's modulus and internal friction of wood in relation to the evaluation of wood for musical instruments," Jpn. J. Appl. Phys., Part 1 **22**(4), 611–614.

Ono, T., and Norimoto, M. (**1985**). "Anisotropy of Dynamic Young's Modulus and Internal Friction in Wood," Jpn. J. Appl. Phys., Part 1 **24**(8), 960–964.

Steiglitz, K., and McBride, L. E. (**1965**). "A technique for the identification of linear systems," IEEE Trans. Autom. Control **AC-10**, 461–464.

Sugiyama, M., Matsunaga, M., Minato, K., and Norimoto, M. (**1994**). "Physical and mechanical properties of pernambuco (*Guilandina echinata* Spreng.) used for violin bows," Mokuzai Gakkaishi **40**, 905–910.

Valette, C., and Cuesta, C. (**1993**). *Mécanique de la Corde Vibrante (Mechanics of Vibrating String)*, Traité des Nouvelles Technologies, série Mécanique (Hermès, Paris).

Wildes, R. P., and Richards, W. A. (**1988**). *Recovering Material Properties from Sound* (MIT Press, Cambridge, MA), Chap. 25, pp. 356–363.

Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2007, Article ID 30194, 13 pages
doi:10.1155/2007/30194

Research Article

Electrophysiological Study of Algorithmically Processed Metric/Rhythmic Variations in Language and Music

Sølvi Ystad,[1] Cyrille Magne,[2,3] Snorre Farner,[1,4] Gregory Pallone,[1,5] Mitsuko Aramaki,[2] Mireille Besson,[2] and Richard Kronland-Martinet[1]

[1] *Laboratoire de Mécanique et d'Acoustique, CNRS, Marseille, France*
[2] *Institut de Neurosciences Cognitives de la Méditerranée, CNRS, 13402 Marseille Cadex, France*
[3] *Psychology Department, Middle Tennessee State University, Murfreesboro, TN 37127, USA*
[4] *IRCAM, 1 Place Igor Stravinsky, 75004 Paris, France*
[5] *France Télécom, 22307 Lannion Cedex, France*

Received 1 October 2006; Accepted 28 June 2007

Recommended by Jont B. Allen

This work is the result of an interdisciplinary collaboration between scientists from the fields of audio signal processing, phonetics and cognitive neuroscience aiming at studying the perception of modifications in meter, rhythm, semantics and harmony in language and music. A special time-stretching algorithm was developed to work with natural speech. In the language part, French sentences ending with tri-syllabic congruous or incongruous words, metrically modified or not, were made. In the music part, short melodies made of triplets, rhythmically and/or harmonically modified, were built. These stimuli were presented to a group of listeners that were asked to focus their attention either on meter/rhythm or semantics/harmony and to judge whether or not the sentences/melodies were acceptable. Language ERP analyses indicate that semantically incongruous words are processed independently of the subject's attention thus arguing for automatic semantic processing. In addition, metric incongruities seem to influence semantic processing. Music ERP analyses show that rhythmic incongruities are processed independently of attention, revealing automatic processing of rhythm in music.

1. INTRODUCTION

The aim of this project associating audio signal processing, phonetics and cognitive neuroscience is twofold. From an audio point of view, the purpose is to better understand the relation between signal dilation and perception in order to develop perceptually ecological algorithms for signal modifications. From a cognitive neuroscience point of view, the aim is to observe the brain's reactions to modifications in duration of small segments in music and language in order to determine whether the perceptual and cognitive computations involved are specific to one domain or rely on general cognitive processes. The association of different expertise made it possible to construct precisely controlled stimuli and to record objective measures of the stimuli's impact on the auditor, using the event-related potential (ERP) method.

An important issue in audio signal processing is to understand how signal modification affects our perception when striving for naturalness and expressiveness in synthesized music and language. This is important in various applications such as designing new techniques to transcode audio tracks from cinema to video format and vice-versa. Specifically, the cinema format comprises a succession of 24 images per second, while the video format comprises 25 images per second. Transcoding between the two formats is realized by projecting the images at the same rate, inducing changes in the duration of the film. Consequently, the soundtrack duration needs to be modified to guarantee synchronization between sounds and images, thus requiring the application of time-stretching algorithms preserving the timbre content of the original soundtrack. A good understanding of how time-stretching can be used without altering perception, and how the quality of various algorithms can be evaluated, are thus of great importance.

A better understanding of how signal duration modifications influence our perception is also important for musical interpretation, since local rhythmic variations represent a key aspect of musical interpretation. A large number of authors (e.g., Friberg et al. [1]; Drake et al. [2]; Hirsh et al. [3]; Hoopen et al. [4]) have studied timing in acoustic communication and the just noticeable difference for small perturbations of isochronous sequences. Algorithms that act on the duration of a signal without modifying its properties are important tools for such studies. Such algorithms have been used in recent studies to show how a mixture between rhythm, intensity and timbre changes influence the interpretation (Barthet et al. [5]).

From a neuro cognitive point of view, recording the brain's reactions to modifications in duration within music and language is interesting for several reasons. First, to determine whether metric cues such as final syllabic lengthening in language[1] are perceived by the listeners, and how these modifications alter the perception (and/or comprehension) of linguistic phrases. This was the specific aim of the language experiment that we conducted. Second, to better understand how musical rhythm is processed by the brain in relation with other musical aspects such as harmony. This was the aim of the music experiment.

Since the early 1980's, the ERP method has been used to examine and compare different aspects of language and music processing. This method has the advantage of allowing to record changes in the brain electrical activity that are time-locked to the presentation of an event of interest. These changes are, however, small in amplitude (of the order of 10 μV) compared to the background EEG activity (of the order of 100 μV). It is therefore necessary to synchronize EEG recordings to the onset of the stimulation (i.e., event of interest) and to average a large number of trials (20 to 50) in which similar stimulations are presented. The variations of potential evoked by the event of interest (therefore called event-related potentials, ERPs) then emerge from the background noise (i.e., the EEG activity). The ERPs comprise a series of positive and negative deflections, called components, relative to the baseline, that is, the averaged level of brain electrical activity within 100 or 200 ms before stimulation. Components are defined by their polarity (negative, N, or positive, P), their latency from stimulus onset (100, 200, 300, 400 ms, etc.), their scalp distribution (location of maximum amplitude on the scalp) and their sensitivity to experimental factors.

So far, these studies seem to indicate that general cognitive principles are involved in language processing when aspects such as syntactic or prosodic processing are compared with harmonic or melodic processing in music (Besson et al. [6], Patel et al. [7]; Magne et al. [8]; Schön et al. [9]). By contrast, a language specificity seems to emerge when semantic processing in language is compared to melodic and harmonic processing in music (Besson and Macar [10], but see Koelsch et al. [11] for counter evidence). Until now, few electrophysiological studies have considered fine metric/rhythmic changes in language and music. One of these studies was related to the analysis of an unexpected pause before the last word of a spoken sentence, or before the last note of a musical phrase (Besson et al. [6]). Results revealed similar reactions to the pauses in music and language, suggesting similarities in rhythmic/metric processing across domain. However, since these pauses had a rather long duration (600 ms), such a manipulation was not ecological and results might reflect a general surprise effect. Consequently, more subtle manipulations are needed to consider rhythmic/metric processing in both music and language. This was the motivation behind the present study. In the language experiment, French sentences were presented, and the duration of the penultimate syllable of trisyllabic final words was increased to simulate a stress displacement from the last to the penultimate syllable. In the music experiment, the duration of the penultimate note of the final triplet of a melody was increased to simulate a rhythmic displacement.

Finally, it was of interest to examine the relationship between violations in duration and harmony. While several authors have used the ERPs to study either harmonic (Patel et al. [12]; Koelsch et al. [13]; Regnault et al. [14]) or rhythmic processing (Besson et al. [6]), to our knowledge, harmonic and rhythmic processing have not yet been combined within the same musical material to determine whether the effects of these fundamental aspects of music are processed in interaction or independently from one another. For this purpose, we built musical phrases composed of triplets, which were presented within a factorial design, so that the final triplet either was both rhythmically and harmonically congruous, rhythmically incongruous, harmonically incongruous, or both rhythmically and harmonically incongruous. Such a factorial design was also used in our language experiment and was useful to demonstrate that metric incongruities in language seems to hinder comprehension. Most importantly, we have developed an algorithm that can stretch the speech signal without altering its other fundamental characteristics (fundamental frequency/pitch, intensity and timbre) in order to use natural speech stimuli. The present paper is mainly devoted to the comparison of reactions to metric/rhythmic and semantic/harmonic changes in language and music, and to the description of the time-stretching algorithm applied to the language stimuli. A more detailed description of the behavioral and ERP data results of the language part is given in (Magne et al. [15]).

2. CONSTRUCTION OF STIMULI

2.1. Language experiment

Rhythm is part of all human activities and can be considered as the framework of prosodic organization in language (Astésano [16]). In French, rhythm (or meter, which is the term used for rhythm in language), is characterized by a final lengthening. Recent studies have shown that French words are marked by an initial stress (melodic stress) and a final stress or final lengthening (Di Cristo [17]; Astésano [16]). The initial stress is however secondary, and words or groups of words are most commonly marked by final lengthening. Similarly, final lengthening is a widespread musical phenomenon leading to deviations from the steady beat that is present in the underlying presentation. These analogies

[1] Final syllable lengthening is a widespread phenomenon across different languages by which the duration of the final syllable of the last word of sentences, or groups of words, is lengthened, supposedly to facilitate parsing/segmentation of groups of words within semantically relevant units.

between language and music led us to investigate rhythm perception in both domains.

A total of 128 sentences with similar number of words and durations, and ending with tri-syllabic words were spoken by a native male French speaker and recorded in an anechoic room. The last word of each sentence was segmented into syllables and the duration of the penultimate syllable was increased. As the lengthening of a word or a syllable in natural speech mainly is realized on the vowels, the artificial lengthening was also done on the vowel (which corresponds to the stable part of the syllable). Words with nasal vowels were avoided, since the segmentation of such syllables into consonants and vowels generally is ambiguous. The lengthening factor (dilation factor) was applied to the whole syllable length (consonant + vowel) for the following reasons:

(1) the syllable is commonly considered as the perceptual unit

(2) an objective was to apply a similar manipulation in both language and music, and the syllabic unit seems closer to a musical tone than the vowel itself. Indeed, musical tones consist of an attack and a sustained part, which may respectively be compared to the syllable's consonant and vowel.

The duration of the penultimate syllable of the last word was modified by a time-stretching algorithm (described in Section 2.1.2). Most importantly, this algorithm made it possible to preserve both the pitch and the timbre of the syllable without introducing audible artifacts. Note that the time-stretching procedure did not alter the F0 and amplitude contours of the stretched syllable, and simply caused these contours to unfold more slowly over time (i.e., the rate of F0 and amplitude variations differ between the metrically congruous and incongruous conditions). This is important to be aware of when interpreting the ERP effect, since it means that the syllable lengthening can be perceived soon after the onset of the stretched second syllable. Values of the mean duration of syllables and vowels in the tri-syllabic words are given in Table 1. The mean duration of the tri-syllabic words was 496 ms and the standard deviation was 52 ms.

Since we wanted to check possible cross-effects between metric and semantic violations, the tri-syllabic word was either semantically congruent or incongruous. The semantic incongruity was obtained by replacing the last word by an unexpected tri-syllabic word, (e.g., "Mon vin préféré est le karaté"—my favorite wine is the karate). The metric incongruity was obtained by lengthening the penultimate syllable of the last word of the sentence ("ra" in "kara̲té") by a dilation factor of 1.7. The choice of this factor was based on the work of Astésano (Astésano [16]), revealing that the mean ratio between stressed and unstressed syllables is approximately 1.7 (when sentences are spoken using a journalistic style).

2.1.1. *Time-stretching algorithm*

In this section, we describe a general time-stretching algorithm that can be applied to both speech and musical signals. This algorithm has been successfully used for cinema to video transcoding (Pallone [18]) for which a maximum of 20% time dilation is needed. We describe how this general algorithm has been adapted to allow up to 400% time dilation on the vowel part of speech signals.

Changing the duration of a signal without modifying its frequency is an intricate problem. Actually, if $s(\omega)$ represents the Fourier transform of a signal $s(t)$, then $(1/\alpha)s(\omega/\alpha)$ is the Fourier transform of $s(\alpha t)$. This obviously shows that compression (resp., lengthening) of a signal induces transposition to higher (resp., lower) pitches. Moreover, the formant structure of the speech signal—due to the resonances of the vocal tract—is modified, leading to an altered voice (the so-called "Donald Duck effect"). To overcome this problem, it is necessary to take into account the specificities of our hearing system.

Time-stretching methods can be divided into two main classes: frequency-domain and time-domain methods. Both methods present advantages and drawbacks, and the choice depends on both the signal to be modified and the specificities of the application.

2.1.2. *Frequency domain methods*

In the frequency domain approach, temporal "grains" of sound are constructed by multiplying the signal by a smooth and compact function (known as a window). These grains are then represented in the frequency domain and are further processed before being transformed back to the time domain. A well-known example of such an approach is the phase vocoder (Dolson [19]), which has been intensively used for musical purposes. The frequency-domain methods have the advantage of giving good results for high stretching ratios. In addition, they do not cause any anisochrony problems, since the stretching is equally spread over the whole signal. Moreover, these techniques are compatible with an inharmonic structure of the signal. They can however cause transient smearing since transformation in the frequency domain tends to smooth the transients (Pallone et al. [20]), and the timbre of a sound can be altered due to phase unlocking (Puckette [21]), although this has been improved later (Laroche and Dolson [22]). Such an approach is consequently not optimal for our purpose, where ecological transformations of sounds (i.e., that could have been made by human beings) are necessary. Nevertheless, they represent valuable tools for musical purpose, when the aim is to produce sound effects, rather than perfect perceptual reconstructions.

2.1.3. *Time-domain methods*

In the time-domain approach, the signal is time-stretched by inserting or removing short, non-modified segments of the original time signal. This approach can be considered as a temporal reorganization of non-modified temporal grains. The most obvious time-stretching method is the so-called "blind" method, which consists in regularly duplicating and inserting segments of constant duration (French and Zinn [23]). Such a method has the advantage of being very simple. However, even by using crossfades, synchronization discontinuities often occur, leading to a periodic alteration of the sound.

Table 1: Mean values (ms), and standard deviation (Sd) in brackets, of vowel(V) and syllable(S) lengths of the tri-syllabic words.

Segments	V1	V2	V3	V3/V2	S1	S2	S3	S3/S2
Meanva and Std	65 (24)	69 (17)	123 (36)	1.79(0.69)	150 (28)	145 (28)	202 (42)	1.39(0.45)

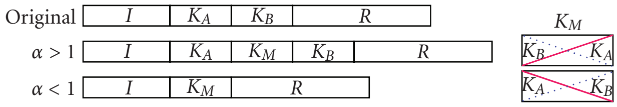


Figure 1: Insertion of a segment K_M to time-stretch a signal frame. The upper stripe represents the original signal. The second one illustrates how the signal is lengthened by adding an element K_M, and the third one illustrates how the signal can be shortened by replacing elements K_A and K_B by the element K_M. I is the initial delay, while R is the residual segment allowing to assure the correct dilation ratio before the next frame is processed.

Other time-domain approaches are based on adaptive methods aiming at matching the length of the inserted segments to the fundamental period (Roucos and Wilgus [24]). These methods give high quality sounds for dilation factors less than 20%. However, a doubling of transients might occur in this case as well as synchronization discontinuities on inharmonic and polyphonic sounds.

Finally, the problem of transient doubling has been addressed by Pallone [18]), whose work has been applied in a commercial product for real-time stretching of movie sound tracks between different playing speeds for instance between video (25 pictures/sec) and cinema (24 pictures/sec) format. The algorithm selects the best segment to insert, optimizes its duration and selects the best location for insertion. It was derived from so-called SOLA (WSOLA and SOLAFS) methods (Verhelst and Roelands [25], Hejna et al. [26]).

In our specific situation it was extremely important that the chosen signal processing method did not cause any audible sound quality modification. The algorithm used by Pallone [18] was found to be extensible to very strong dilation ratios, so we decided to adopt and optimize it for our purpose. We also foresee its usage on stretching of musical signals although we have settled on using MIDI in the music part of this study. In the following section, we briefly describe the algorithm in its completeness before presenting the optimizations that made us able to stretch vowels more than four times without audible defects.

2.1.4. A specific time-based algorithm

The principle of the time-domain algorithm is illustrated in Figure 1. The original signal is sequentially decomposed into a series of consecutive frames. Each frame is cut into 4 segments defined by 2 main parameters:

(1) the segment I, whose length I represents an initial delay, which can be adjusted in order to choose the best area of the frame for manipulation, and

(2) the segment K_M, whose length K is also the length of both K_A and K_B.

Letting α be the stretching factor, a lengthening of the signal ($\alpha > 1$) can be obtained by crossfading elements K_B and K_A, and inserting the resulting segment K_M between K_A and K_B. A similar procedure can be used to shorten the signal ($\alpha < 1$): by replacing K_A and K_B by a crossfaded segment K_M obtained from K_B and K_A. The crossfading prevents discontinuities because the transitions at the beginning and the end of K_M correspond to the initial transitions.

Each signal frame should be modified so that the dilation ratio is respected within the frame. The relation linking the length of R with the length of I, K_A, K_B, and K_M is thus given by the equation:

$$\alpha(I + K_A + K_B + R) = (I + K_A + K_M + K_B + R). \quad (1)$$

For $\alpha < 1$ (signal shortening), the segments K_A and K_B are set to zero at the right-hand side. Although this process seems simple and intuitive in the case of a periodic signal (as the length K should correspond to the fundamental period), the choice of the segments K_A and K_B is crucial and may be difficult if the signal is not periodic. The difficulty consists in adapting the duration of these segments-and consequently of K_M-to prevent the time-stretching process from creating any audible signal modifications other than the perceptual dilation itself. On one hand, a segment that is too long might, for instance, provoke the duplication of a localized energetic event (for instance a transient) or create a rhythmic distortion (anisochrony). Studies on anisochrony have shown that for any tempo, the insertion of a segment of less than 6 ms remains inaudible unless it contains an audible transient (Friberg and Sundberg [1]). On the other hand, a short segment might cause discontinuities in a low-frequency signal, because the inserted segment does not correspond to a complete period of the signal. This also holds for polyphonic and inharmonic signals in the case that a (long) common period may be found. Consequently, the length of the inserted segment must be adapted to the nature of the signal so that a long segment can be inserted when stretching a low-frequency signal and a short segment can be inserted when the signal is non-stationary.

To calculate the location and length of the inserted element K_M, different criteria were proposed for determining the local periodicity of the signal and the possible presence of transients. These criteria are based on the behavior of the autocorrelation function and of the time-varying energy of the signal, leading to an improvement of the sound quality obtained using WSOLA.

Choice of the length K of the inserted segment

The main issue here consists in determining the length K that gives the strongest similarity between two successive segments. This condition assures an optimal construction of the segment K_M and continuity between the inserted segment and its neighborhood. We have compared three different approaches for the measurement of signal similarities, namely the average magnitude difference function, the autocorrelation function, and the normalized autocorrelation function. Due to the noise sensitivity of the average magnitude function (Verhelst and Roelands [25] and Laroche [27]) and to the autocorrelation function's sensibility to the signal's energy level, the normalized autocorrelation function given by

$$CN(k) = \frac{\sum_{n=0}^{N_c-1} s(n)s(n+k)}{\sqrt{\sum_{n=0}^{N_c-1} s^2(n) \sum_{n=0}^{N_c-1} s^2(n+k)}} \qquad (2)$$

was applied. This function takes into account the energy of the analyzed chunks of signal. Its maximum is given by $k = K$, as for the autocorrelation function $C(k)$, and indicates the optimal duration of the segment to be inserted. For instance, if we consider a periodic signal with a fundamental period T_0, two successive segments of duration T_0 have a normalized correlation maximum of 1. Note that this method requires the use of a "forehand criterion" in order to compare the energy of the two successive elements K_A and K_B, otherwise, the inserted segment K_M might create a doubling of the transition between a weak and a strong sound level. Using a classical energy estimator easily allows to deal with this potential problem.

2.1.5. Modifications for high dilation factors

As mentioned in Section 2.1.1, our aim was to work with natural speech and to modify the syllable length of the second-last syllable of the last word in a sentence by a factor 1.7.

The described algorithm works very well for dilation factors up to about 20% ($\alpha = 1.2$) for any kind of audio signal, but for the current study higher dilation factors were needed. Furthermore, since vowels rather than consonants are stretched when a speaker slows down the speed in natural speech, only the vowel part of the syllable was stretched by the algorithm. Consequently, the local dilation factor applied on the vowel was necessarily greater than 1.7, and varied from 2 to 5 depending on the vowel to consonant ratio of the syllable. To achieve such stretching ratios, the above algorithm had to be optimized for vowels. Since the algorithm was not designed for dilation ratios above $\alpha = 1.2$, it could be applied iteratively until the desired stretching ratio was reached. Hence, applying the algorithm six times would give a stretching ratio of $\alpha = 1.2^6 \approx 3$. Unfortunately, we found that after only a few repetitions, the vowel was perceived as "metallic," probably because the presence of the initial segment I (see Figure 1) caused several consecutive modifications of some areas while leaving other ones unmodified.

Within a vowel, the correlation between two adjacent periods is high, so the initial segment I does not have to be estimated. By setting its length I to zero and allowing the next frame to start immediately after the modified element K_M, the dilation factor can be increased to a factor 2. The algorithm inserts one modified element K_M of length K between the two elements K_A and K_B, each of the same length K, and then lets K_B be the next frame's K_A. In the above described algorithm, this corresponds to a rest segment R of length-K for $\alpha = 2$.

The last step needed to allow infinite dilation factors, consists in letting the next segment start inside the modified element K_M (i.e., allowing for $-2K < R < -K$). This implies re-modifying the already modified element and this is a source for adding a metallic character to the stretched sound. However, with our stretching ratios, this was not a problem. In fact, as will be evident later, no specific perceptual reaction to the sound quality of the time-stretched signal were elicited, as evidenced by the typical structure of the ERP components.

Sound examples of speech signal stretched by means of such a technique can be found at http://www.lma.cnrs-mrs.fr/~ystad/Prosem.html, together with a small computer program to do the manipulations.

2.2. Music experiment

Rhythmic patterns like long-short alternations or final lengthening can be observed in both language and music (Repp [28]). In this experiment, we constructed a set of melodies comprising 5–9 triplets issued from minor or major chords. The triplets were chosen to roughly imitate the language experiment, since the last word in each sentence always was tri-syllabic. As mentioned above, the last triplet of the melody was manipulated either rhythmically or harmonically, or both, leading to four experimental conditions. The rhythmic incongruity was obtained by dilating the second-last note of the last triplet by a factor 1.7, like in the language experiment. The first note of the last triplet was always harmonically congruous with the beginning of the melody, since in the language part the first syllable of the last word in the sentences did not indicate whether or not the last word was congruous or incongruous. Hence, this note was "harmonically neutral," so that the inharmonicity could not be perceived before the second note of the last triplet was presented. In other words, the first note of an inharmonic triplet was chosen to be harmonically coherent with both the beginning (harmonic part) and the end (inharmonic part) of the melody.

A total of 128 melodies were built for this purpose. Further, the last triplet in each melody was modified to be harmonically incongruous (R+H−), rhythmically incongruous (R−H+), or both (R−H−). Figure 2 shows a harmonically congruous (upper part) and harmonically incongruous (lower part) melody. Each of these 4 experimental conditions comprised 32 melodies that were presented in pseudorandom order (no more than 4 successive melodies for the same condition) in 4 blocks of 32 trials. Thus, each participant listened to 128 different melodies. To ensure that each melody was presented in each of the four experimental conditions across participants, 4 lists were built and a total of 512 stimuli were created.

Piano tones from a sampler (i.e., prerecorded sounds) were used to generate the melodies. Frequencies and durations of the notes in the musical sequences were modified by altering the MIDI codes (Moog [29]). The time-stretching algorithm used in the language experiment could also have been used here. However, the use of MIDI codes considerably simplified the procedure and the resulting sounds were of very good quality (http://www.lma.cnrs-mrs.fr/∼ystad/Prosem.html, for sound examples). To facilitate the creation of the melodies, a MAX/MSP patch (Puckette et al. [30]) has been developed so that each triplet was defined by a chord (see Figure 3). Hereby, the name of the chord (e.g., C3, G4...), the type (minor or major), the first and following notes (inversions) can easily be chosen. For instance, to construct the first triplet of the melody in Figure 3 (notes G1, E1 and C2), the chord to be chosen is C2 with inversions −1 (giving G1 which is the closest chord note below the tonic), −2 (giving E1 which is the second closest note below the tonic) and 1 (giving C2 which is the tonic). A rhythmic incongruity can be added to any triplet. In our case, this incongruity was only applied to the second note of the last triplet, and the dilation factor was the same for all melodies ($\alpha = 1.7$). The beat of the melody can also be chosen. In this study, we used four different beats: 70, 80, 90, and 100 triplets/minute, so that the inter-onset-interval (IOI) between successive notes varied from 200 ms to 285 ms, with an increase of IOI, due to the rhythmic modifications, that varied from 140 ms to 200 ms.[2] Finally, when all the parameters of the melodies were chosen, the sound sequences were recorded as wave files.

2.3. Methods

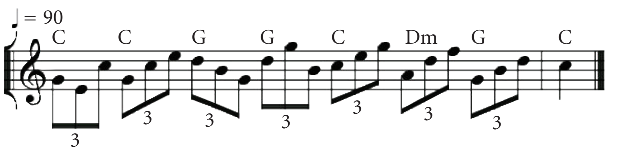
Subjects

A total of 14 participants (non-musicians, 23-years-old on the average) participated in the language part, of which 8 participated in the music part of the experiment. Volunteers were students from the Aix-Marseille Universities and were paid to participate in the experiments that lasted for about 2 hours. All were right-handed native French speakers, without hearing or neurological disorders. Each experiment began with a practice session to familiarize participants with the task and to train them to blink during the interstimulus interval.

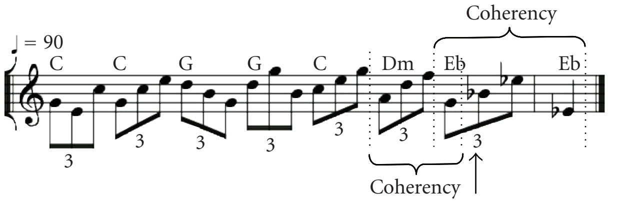
Procedure

In the present experiment, 32 sound examples (sentences or melodies) were presented in each experimental condition, so that each participant listened to 128 different stimuli. To make sure a stimulus was presented only once in the four experimental conditions, 512 stimuli were created to be used either in the language or in the music experiment. Stimuli were presented in 4 blocks of 32 trials.

The experiment took place in a Faradized room, where the participants, wearing an Electro Cap (28 electrodes), listened to the stimuli through headphones. Within two blocks of trials, participants were asked to focus their attention on the metric/rhythmic aspects of the sentences/melodies to decide whether the last syllable/note was metrically/rhythmically acceptable or not. In the other two blocks, participants were asked to focus their attention on the semantic/harmony in order to decide whether the last syllable/note was semantically/harmonically acceptable or not. The responses are given by pressing one of two response buttons as quickly as possible. The side (left or right hand) of the response was balanced across participants.



(a)



(b)

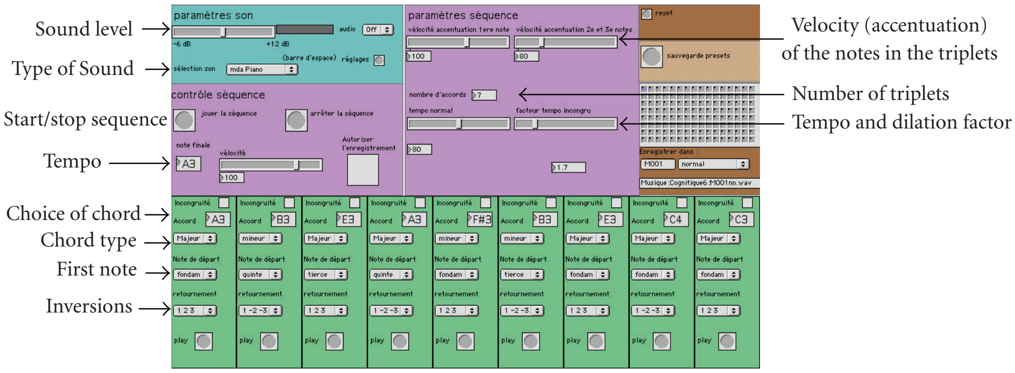
FIGURE 2: Upper part of the figure corresponds to a harmonically congruous melody, while the lower part corresponds to a harmonically incongruous melody. In the rhythmically incongruous conditions, the duration of the second last notes of the last triplet (indicated by an arrow in the lower part) was increased by a factor 1.7.

In addition to the measurements of the electric activity (EEG), the percentage of errors, as well as the reaction times (RTs), were measured. The EEG was recorded from 28 active electrodes mounted on an elastic head cap and located at standard left and right hemisphere positions over frontal, central, parietal, occipital and temporal areas (International 10/20 system sites; Jasper [31]). EEG was digitized at a 250 Hz sampling rate using a 0.01 to 30 Hz band pass. Data were rereferenced off-line to the algebraic average over the left and right mastoids. EEG trials contaminated by eye-, jaw- or head movements, or by a bad contact between the electrode and the skull, were eliminated (approximately 10%). The remaining trials were averaged for each participant within each of the 4 experimental conditions. Finally, a grand average was obtained by averaging the results across all participants.

Error rates and reaction times were analyzed using Analysis of Variance (ANOVAs) that included Attention (Rhythmic versus Harmonic), Harmonics (2 levels) and Rhythmic (2 levels) within-subject factors.

ERP data were analyzed by computing the mean amplitude in selected latency windows, relative to a baseline, and determined both from visual inspection and on the basis of previous results. Analysis of variance (ANOVAs) were used for all statistical tests, and all P-values reported below were adjusted with the Greenhouse-Geisser epsilon correction for non-sphericity. Reported are the uncorrected degrees

[2] A simple statistical study of syllable lengths in the language experiment showed that an average number of around 120 tri-syllabic words per minute were pronounced. Such a tempo was however too fast for the music part.



audio Off
(barre d'espace) réglages
sélection son: mda Piano
note finale: A3
vélocité: 100
vélocité accentuation 1ere note: 100
vélocité accentuation 2e et 3e notes: 80
nombre d'accords: 7
tempo normal: 80
facteur tempo incongru: 1.7
Enregistrer dans : M001 normal
Musique :Cognitique6 :M001nn.wav

Incongruité	□	□	□	□	□	□	□	□	□
Accord	A3	B3	E3	A3	F#3	B3	E3	C4	C3
	Majeur	mineur	Majeur	Majeur	mineur	mineur	Majeur	Majeur	Majeur
Note de départ	fondam	quinte	tierce	quinte	fondam	tierce	fondam	fondam	fondam
retournement	1 2 3	1 -2 -3	1 2 3	1 -2 -3	1 2 3	1 -2 -3	1 2 3	1 -2 -3	1 2 3

Figure 3: Real-time interface (Max/MSP) allowing for the construction of the melodies. In the upper left corner, the sound level is chosen (here constant for all the melodies) and underneath a sequence control allowing to record the melodies suitable for the experiment. In the upper right part, the tempo, number of triplets and the incongruity factor are chosen. Finally, the chords defining each triplet are chosen in the lowest part of the figure.

of freedom and the probability level after correction. Separate ANOVAs were computed for midline and lateral sites separately.

Separate ANOVAs were conducted for the Metric/Rhythmic and Semantic/Harmonic task. Harmony (2 levels), Rhythmic (2 levels) and Electrodes (4 levels) were used as within-subject factors for midline analysis. The factors Harmony (2 levels) and Rhythm (2 levels) were also used for the lateral analyses, together with the factors Hemisphere (2 levels), Anterior-Posterior dimension (3 regions of interest-ROIs): fronto-central (F3, Fc5, Fc1; F4, Fc6, Fc2), temporal (C3, T3, Cp5; C4, T4, Cp6) and temporo-parietal (Cp1, T5, P3; Cp2, T6, P4) and Electrodes (3 for each ROI), as within-subject factors, to examine the scalp distribution of the effects. Tukey tests were used for all post-hoc comparisons. Data processing was conducted with the Brain Vision Anayser software (Version 01/04/2002; Brain Products, Gmbh).

3. RESULTS

3.1. Language experiment

We here summarize the main results of the experiment conducted with the linguistic stimuli, mainly focusing on the acoustic aspects. A more detailed description of these results can be found in (Magne et al. [15]).

3.1.1. Behavioral data

Results of a three-way ANOVA on a transformed percentage of errors showed two significant effects. The meter by semantics interaction was significant ($F(1, 12) = 16.37$, $P < .001$): the participants made more errors when one dimension, Meter (19.5%) or Semantics (20%) was incongruous than when both dimensions were congruous (12%) or incongruous (16.5%). The task by meter by semantics interaction was also significant ($F(1, 12) = 4.74$, $P < .05$): the participants made more errors in the semantic task when semantics was congruous, but meter was incongruous (S+M−), (24%), than in the other three conditions.

The results of the three-way ANOVA on the RTs showed a main effect of semantics ($F(1, 12) = 53.70$, $P < .001$): they always were significantly shorter for semantically congruous (971 ms) than for incongruous words (1079 ms).

3.1.2. Electrophysiological data

Results revealed two interesting points. First, independently of the direction of attention toward semantics or meter, semantically incongruous (but metrically congruous) final words (M+S−) elicited larger N400 components than semantically congruous words (M+S+). Thus, semantic processing of the final word seems task-independent and automatic. This effect was broadly distributed over the scalp.

Second, some aspects of metric processing also seemed task independent because metrically incongruous words also elicited an N400-like component in both tasks (see Figure 4). As opposed to the semantically incongruous case, the meter by hemisphere interaction was almost significant ($P < .06$): the amplitude of the negative component was somewhat larger over the right hemisphere (metrically congruous versus incongruous: $F(1, 13) = 15.95$, $P = .001$; $d = -1.69\,\mu$V) than over the left hemisphere (metrically congruous versus incongruous: $F(1, 13) = 6.04$, $P = .03$; $d = -1.11\,\mu$V). Finally, a late positivity (P700 component) was only found for metrically incongruous words when participants focused their attention on the metric aspects, which may reflect the explicit processing of the metric structure of words.

No differences in low-level acoustic factors between the metrically congruous and incongruous stimuli were observed. This result is important from an acoustical point of view, since it confirms that no spurious effect due to a non-ecological manipulation of the speech signal has been created by the time-stretching algorithm described in Section 2.1.2.

3.2. Music experiment

3.2.1. Behavioral data

The percentages of errors and the RTs in the four experimental conditions (R+H+, R+H−, R−H+, and R−H−) in

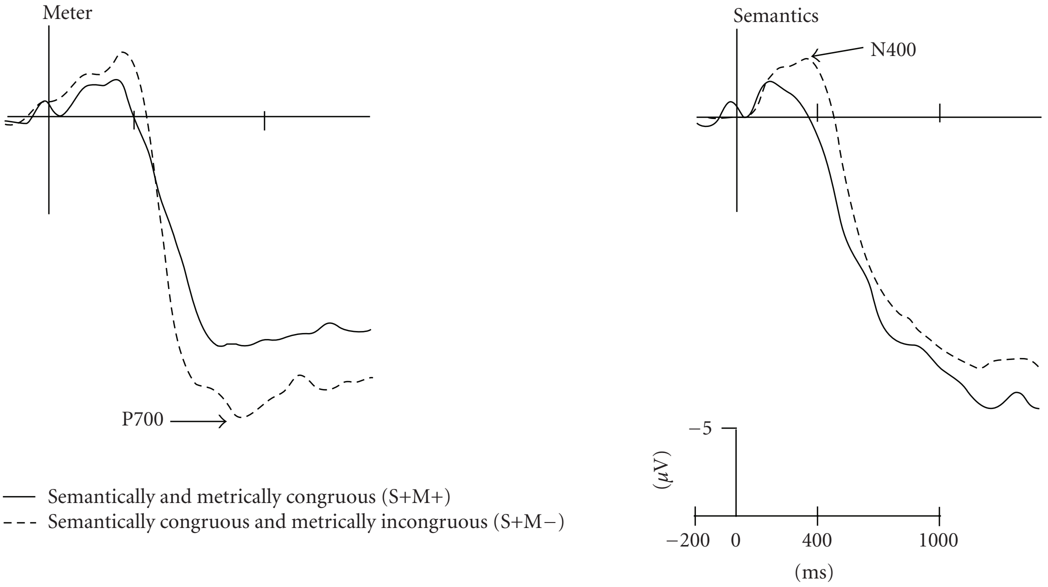


Figure 4: Event-related potentials (ERP) evoked by the presentation of the semantically congruous words when metrically congruous (S+M+) or metrically incongruous (S+M−). Results when participant focused their attention on the metric aspects are illustrated in the left column (Meter) and when they focused their attention on the semantic aspects in the right column (Semantic). The averaged electrophysiological data are presented for one representative central electrode (C_z).

the two attentional tasks (Rhythmic and Harmonic) are presented in Figures 5 and 6.

Results of a three-way ANOVA on the transformed percentages of errors showed a marginally significant main effect of Attention [$F(1,7) = 4.14$, $P < .08$]: participants made somewhat more errors in the harmonic task (36%) than in the rhythmic task (19%). There was no main effect of Rhythmic or Harmonic congruity, but the Rhythmic by Harmonic congruity interaction was significant [$F(1,7) = 6.32$, $P < .04$]: overall, and independent of the direction of attention, participants made more errors when Rhythm was congruous, but Harmony was incongruous (i.e., condition R+H−) than in the other three conditions.

Results of a three-way ANOVA on RTs showed no main effect of Attention. The main effect of Rhythmic congruity was significant [$F(1,7) = 7.69$, $P < .02$]: RTs were shorter for rhythmically incongruous (1213 ms) than for rhythmically congruous melodies (1307 ms). Although a similar trend was observed in relation to Harmony, the main effect of Harmonic congruity was not significant.

3.2.2. *Electrophysiological data*

The electrophysiological data recorded in the four experimental conditions (R+H+, R+H−, R−H+, and R−H−) in the two tasks (Rhythmic and Harmonic) are presented in Figures 7 and 8. Only ERPs to correct response were analyzed.

Attention to rhythm

In the 200–500 ms latency band, the main effect of Rhythmic congruity was significant at midline and lateral electrodes [Midlines: $F(1,7) = 11.01$, $P = .012$; Laterals: $F(1,7) = 21.36$, $P = .002$]: Rhythmically incongruous notes (conditions R−H+ and R−H−) elicited more negative ERPs than rhythmically congruous notes (conditions R+H+ and R+H−−). Moreover, the main effect of Harmonic congruity was not significant, but the Harmonic congruity by Hemisphere interaction was significant [$F(1,7) = 8.47$, $P = .022$]: Harmonically incongruous notes (conditions R+H− and R−H−) elicited more positive ERPs than harmonically congruous notes (conditions R+H+ and R−H+) over the right than the left hemisphere.

In the 500–900 ms latency band, results revealed a main effect of Rhythmic congruity at midline and lateral electrodes [midlines: $F(1,7) = 78.16$, $P < .001$; laterals: $F(1,7) = 27.72$, $P = .001$]: Rhythmically incongruous notes (conditions R−H+ and R−H−) elicited more positive ERPs

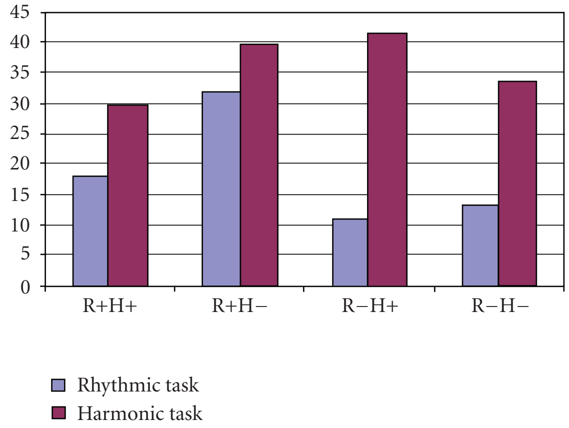


Figure 5: Percentages of error.

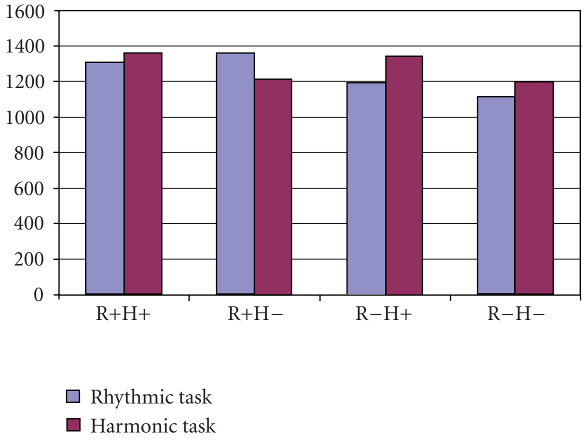


Figure 6: Reaction times (RTs).

than rhythmically congruous notes (conditions R+H+ and R+H−). This effect was broadly distributed over the scalp (no significant rhythmic congruity by Localization interaction). Finally, results revealed no significant main effect of Harmonic congruity, but a significant Harmonic congruity by Localization interaction at lateral electrodes [$F(2, 14) = 10.85$, $P = .001$]: Harmonically incongruous notes (conditions R+H− and R−H−) elicited more positive ERPs than harmonically congruous notes (conditions R+H+ and R−H+) at frontal electrodes. Moreover, the Harmonic congruity by Hemisphere interaction was significant [$F(1, 7) = 8.65$, $P = .02$], reflecting the fact that this positive effect was larger over the right than the left hemisphere.

Attention to Harmony

In the 200–500 ms latency band, both the main effects of Harmonic and Metric congruity were significant at midline electrodes [$F(1, 7) = 5.16$, $P = .05$ and $F(1, 7) = 14.88$, $P = .006$, resp.] and at lateral electrodes [$F(1, 7) = 5.55$, $P = .05$ and $F(1, 7) = 11.14$, $P = .01$, resp.]: Harmonically incongruous musical notes (conditions H−R+ and H−R−) elicited more positive ERPs than harmonically congruous notes (conditions H+R+ and H+R−). By contrast, rhythmically incongruous notes (conditions H+R− and H−R−) elicited more negative ERPs than Rhythmically congruous notes (conditions H+R+ and H−R+). These effects were broadly distributed over the scalp (no Harmonic congruity or Rhythmic congruity by Localization interactions).

In the 500–900 ms latency band, the main effect of Harmonic congruity was not significant, but the Harmonic congruity by Localization interaction was significant at lateral electrodes [$F(2, 14) = 4.10$, $P = .04$]: Harmonically incongruous musical notes (conditions H−R+ and H−R−) still elicited larger positivities than harmonically congruous notes (conditions H+R+ and H+R−) over the parieto-temporal sites of the scalp. Finally, results revealed a main effect of Rhythmic congruity at lateral electrodes [$F(1, 7) = 5.19$, $P = .056$]: Rhythmically incongruous notes (conditions H+R− and H−R−) elicited more positive ERPs than rhythmically congruous notes (conditions H+R+ and H−R+). This effect was broadly distributed over the scalp (no significant Rhythmic congruity by Localization interaction).

4. DISCUSSION

This section is organized around three main points. First, we discuss the result of the language and music experiments, second we compare the effects of metric/rhythmic and semantic/harmonic incongruities in both experiments, and finally, we consider the advantages and limits of the algorithm that was developed to create ecological, rhythmic incongruities in speech.

4.1. Language and music experiment

In the language part of the experiment, two important points were revealed. Independently of the task, semantically incongruous words elicited larger N400 components than congruous words. Longer RTs are also observed for semantically incongruous than congruous words. These results are in line with the literature and are usually interpreted as reflecting greater difficulties in integrating semantically incongruous compared to congruous words in ongoing sentence contexts (Kutas and Hillyard [32]; Besson et al. [33]). Thus participants seem to process the meaning of words even when instructed to focus attention on syllabic duration. The task independency results are in line with studies of Astésano (Astésano et al. [34]), showing the occurrence of N400 components independently of whether participants focused their attention on semantic or prosodic aspects of the sentences. The second important point of the language experiment is related to the metric incongruence. Independently of the direction of attention, metrically incongruous words elicit larger negative components than metrically congruous words in the 250–450 ms latency range. This might reflect the automatic nature of metric processing. Such early negative components have also been reported in the literature when controlling the influence of acoustical factors as prosody. In a study by Magne (Magne et al. [35]), a N400 component was observed when prosodically incongruous final sentence words were presented. This result might indicate that the violations of metric structure interfere with lexical access and thereby hinder access to word meaning. Metric incongruous words also elicited late positive components. This is in line with previous findings indicating that the manipulation of different acoustic parameters of the speech signal such as F0 and intensity, is associated with increased positivity (Astésano et al. [34], Magne et al. [35], Schön et al. [9]).

In the music part of the experience, analysis of the percentage of errors and RTs revealed that the harmonic task was somewhat more difficult than the rhythmic task. This may reflect the fact, pointed out by the participants at the end of the experiment, that the harmonic incongruities could be interpreted as a change in harmonic structure possibly continued by a different melodic line. This interpretation is coherent with the high error rate in the harmonically incongruous, but rhythmically congruous condition (R+H−) in both attention tasks. Clearly, harmonic incongruities seem

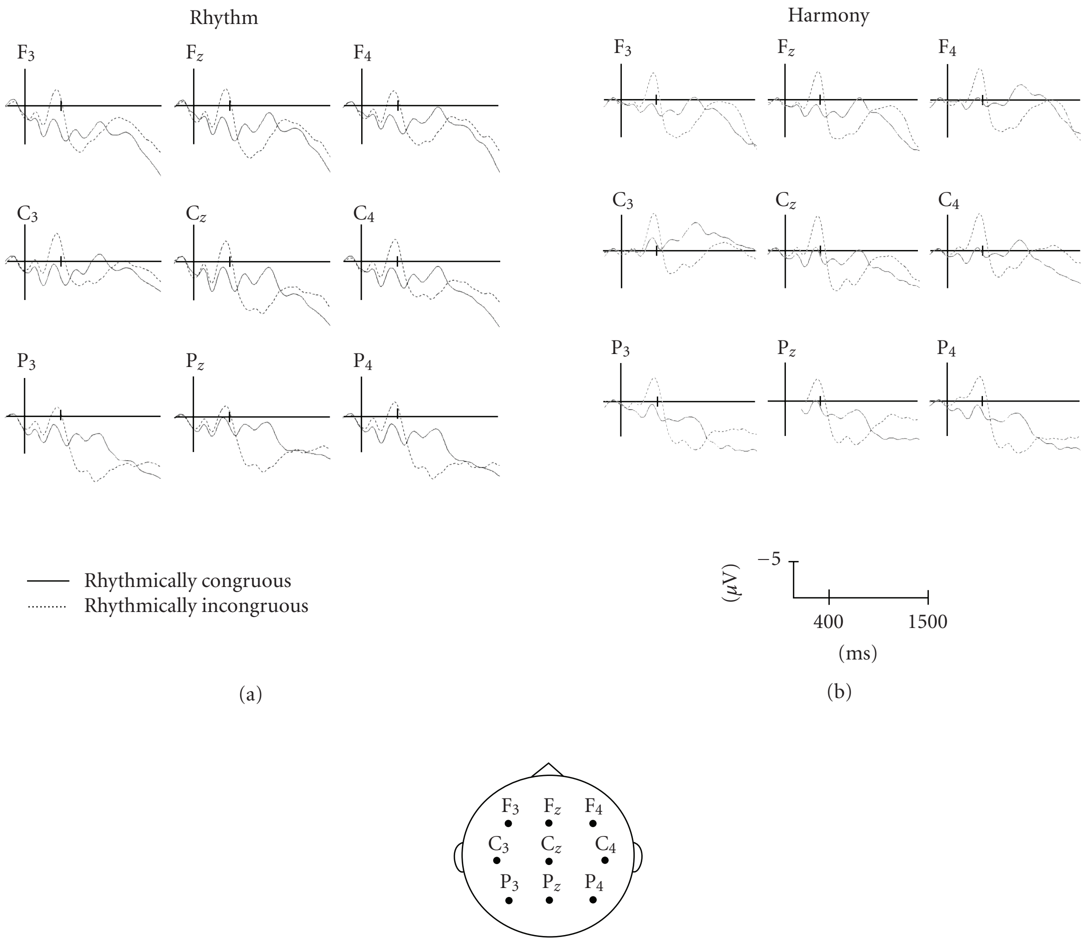


Figure 7: Event-related potentials (ERPs) evoked by the presentation of the second note of the last triplet when rhythmically congruous (solid trace; conditions H+R+ and H−R+) or rhythmically incongruous (dashed trace, conditions H+R− and H−R−). Results when participant focused their attention on the rhythmic aspects are illustrated in the left column (a) and when they focused their attention on the harmonic aspects in the right column (b). On this and subsequent figures, the amplitude of the effects is represented on the ordinate (microvolts, μV; negativity is up), time from stimulus onset on the abscissa (milliseconds, ms).

more difficult to detect than rhythmic incongruities. Finally, RTs were shorter for rhythmically incongruous than congruous notes, probably because participants in the last condition waited to make sure the length of the note was not going to be incongruous.

Interestingly, while rhythmic incongruities elicited an increased negativity in the early latency band (200–500 ms), harmonic incongruities were associated with an increased positivity. Most importantly, these differences were found independently of whether participants paid attention to rhythm or to harmony. Thus, different processes seem to be involved by the rhythmic and harmonic incongruities and these processes seem to be independent of the task at hand. By contrast, in the later latency band (500–900 ms) both types of incongruities elicited increased positivities compared to congruous stimuli. Again, these results were found independently of the direction of attention. Note, however, that the scalp distribution of the early and late positivity to harmonic incongruities differs depending upon the task: while it was larger over right hemisphere in the rhythmic task, it was largely distributed over the scalp and somewhat larger over the parieto-temporal regions in the harmonic task. While this last finding is in line with many results in the literature (Besson and Faïta [36]; Koelsch et al. [13, 37]; Patel et al. [12]; Regnault et al. [14]), the right distribution is more surprising. It raises the interesting possibility that the underlying process varies as a function of the direction of attention, a hypothesis that already has been proposed in the literature (Luks et al. [38]). When harmony is processed implicitly, because irrelevant for the task at hand (Rhythmic task), the right hemisphere seems to be more involved, which is in line with brain imaging results showing that pitch processing seems to be lateralized in right frontal regions (e.g., Zatorre et al. [39]). By contrast, when harmony is processed explicitly (Harmonic task), the typical centro-parietal distribution is found which may reflect the influence of decision-related processes. Taken together, these results are important because they show that different processes are responsible for the processing of rhythm and harmony when listening to the short musical sequences used here. Moreover, they open the

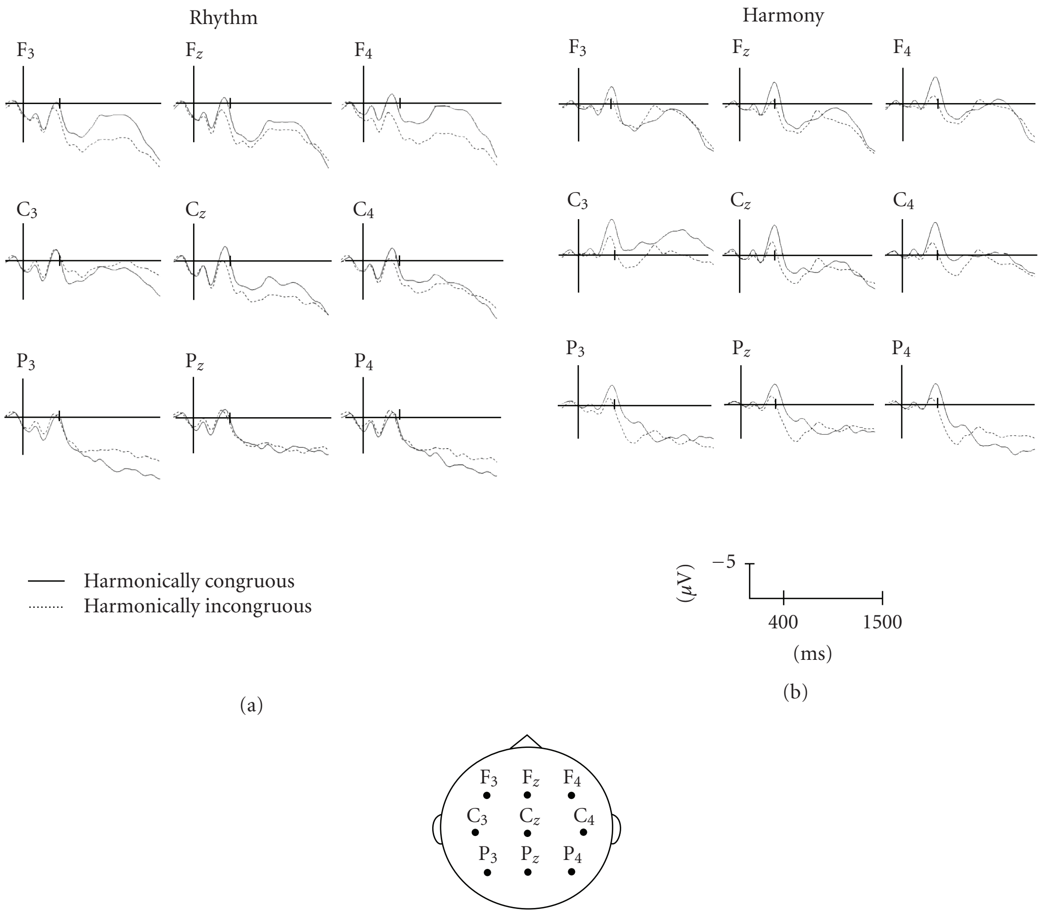


Figure 8: Event-related potentials (ERPs) evoked by the presentation of the second note of the last triplet when harmonically congruous (solid trace, conditions H+R+ and H−R+) or harmonically incongruous (dashed trace; conditions H+R− and H−R−). Results when participant focused their attention on the rhythmic aspects are illustrated in the left column (a) and when they focused their attention on the harmonic aspects in the right column (b).

intriguing possibility that the distribution of these processes vary as a function of attention. These issues clearly need to be pursued in future experiments.

4.1.1. Comparison between language and music

These results on the processing of rhythm in both speech and music point to interesting similarities and differences. Considering the similarities first, it is clear that the rhythmic structure is processed on-line in both language and music. Indeed, rhythmic incongruities elicited a different pattern of brain waves than rhythmically congruous events. Moreover, it is worth noting that rhythmic incongruities elicited increased positive deflections in similar latency bands in both language (P700) and music (P600). However, while these positive components were present in music under both attentional conditions (rhythmic and harmonic tasks), they were only present in the rhythmic task in the language experiment. Thus, while the processing of rhythm may be obligatory when listening to the short melodic sequences presented here, it seems to be modulated by attention when listening to the linguistic sentences.

Turning to the differences, it is interesting to note that while the rhythmic incongruity elicited positive components that were preceded by negative components in music, under both attentional conditions, no such early negativities were found in language in the rhythmic task. Thus, rhythmic violations seem to elicit earlier and larger effects in music than in language, which points to a larger influence on rhythm in music than in speech. Finally, one important difference between speech and music is that while semantic incongruities elicited clear N400 components (Kutas and Hillyard [32]) in the language experiment (Magne et al. [15]), as expected from a large body of results in the ERPs and language literature, no such N400s were associated to the presentation of harmonic incongruities in the music experiment reported here. While semantic processing may not be restricted to linguistic stimuli as demonstrated by (Koelsch et al. [11]) with the occurrence of an N400 component to words that did not match the musical content of the preceding musical phrase, it nevertheless “remains” that the type of harmonic violations used here did not seem to involve semantic processing. Therefore, the intriguing issue of musical semantic processing, remains open for future research.

4.1.2. Algorithm

The time-stretching algorithm that was adapted to dilation of syllables in the language experiment, allowed up to 400% time dilation on the vowel part of the speech signals. Most importantly for our purpose, and in spite of these high stretching ratios, the application of the algorithm did not induce any modifications in sound quality as evidenced by the typical structure of the electrophysiological data in this experiment. No spurious effect, due to a non-ecological manipulation of the speech signal (producing differences in low-level acoustic factors between the rhythmically congruous and incongruous stimuli) was observed in the ERPs. These results are important from an acoustic point of view, since they show the ecological validity of the time-stretching algorithm described in section 2.1.3.

Taken together, from an interdisciplinary perspective, our results demonstrate the influence of metrical structure in language and its important consequences for lexical access and word meaning. More generally, they highlight the role of lexical prosody in speech processing. These are important results from an audio signal processing point of view, since they imply that dilating signals can affect the comprehension of the utterances. Speech sequences in an audio signal must therefore be modified with care and the context should be taken into account, meaning that, for instance, sequences containing important information or sequences with a lot of expressiveness should be less modified than other sequences.

In the music part of the experiment, MIDI codes were used to modify the note duration. The algorithm could have been applied for this purpose, and results of this type are important for future applications of the algorithm on musical signals. Finally, note that our findings imply that the rhythmic modifications should be applied to music and language according to different rules. In the language part, semantics and context should be taken into account, as mentioned earlier, while in the music part interpretation rules according to instrument types and musical genre should determine eventual modifications of sound sequences.

ACKNOWLEDGMENTS

We would like to thank Thierry Voinier for developing the MAX/MSP patch used for constructing the melodies in the music part of the experiment. This research was partly supported by a grant from the Human Frontier Science Program (HSFP #RGP0053) to Mireille Besson and a grant from the French National Research Agency (ANR, JC05-41996, "sensons") to Sølvi Ystad. Mitsuko Aramaki was supported by a post-doctoral grant from the HFSP and the ANR.

REFERENCES

[1] A. Friberg and J. Sundberg, "Time discrimination in a monotonic, isochronous sequence," *Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2524–2531, 1995.

[2] C. Drake and M. C. Botte, "Tempo sensitivity in auditory sequences: Evidence for a multiple-look model," *Perception and Psychophysics*, vol. 54, pp. 277–286, 1993.

[3] I. J. Hirsh, C. B. Monahan, K. W. Grant, and P. G. Singh, "Studies in auditory timing : I, simple patterns," *Perception and Psychophysics*, vol. 74, no. 3, pp. 215–226, 1990.

[4] G. ten Hoopen, L. Boelaarts, A. Gruisen, I. Apon, K. Donders, N. Mul, and S. Aker-boom, "The detection of anisochrony in monaural and interaural sound sequences," *Perception and Psychophysics*, vol. 56, no. 1, pp. 210–220, 1994.

[5] M. Barthet, R. Kronland-Martinet, S. Ystad, and Ph. Depalle, "The effect of timbre in clarinet interpretation," in *Proceedings of the International Computer Music Conference (ICMC '07)*, Copenhagen, Denmark, August 2007.

[6] M. Besson, F. Faïta, C. Czternasty, and M. Kutas, "What's in a pause: event-related potential analysis of temporal disruptions in written and spoken sentences," *Biological Psychology*, vol. 46, no. 1, pp. 3–23, 1997.

[7] A. D. Patel and J. R. Daniele, "An empirical comparison of rhythm in language and music," *Cognition*, vol. 87, no. 1, pp. B35–B45, 2003.

[8] C. Magne, D. Schön, and M. Besson, "Prosodic and melodic processing in adults and children: behavioral and electrophysiologic approaches," *Annals of the New York Academy of Sciences*, vol. 999, pp. 461–476, 2003.

[9] D. Schön, C. Magne, and M. Besson, "The music of speech: music training facilitates pitch processing in both music and language," *Psychophysiology*, vol. 41, no. 3, pp. 341–349, 2004.

[10] M. Besson and F. Macar, "An event-related potential analysis of incongruity in music and other non-linguistic contexts," *Psychophysiology*, vol. 24, no. 1, pp. 14–25, 1987.

[11] S. Koelsch, E. Kasper, D. Sammler, K. Schulze, T. Gunter, and A. D. Friederici, "Music, language and meaning: brain signatures of semantic processing," *Nature Neuroscience*, vol. 7, no. 3, pp. 302–307, 2004.

[12] A. D. Patel, E. Gibson, J. Ratner, M. Besson, and P. J. Holcomb, "Processing syntactic relations in language and music: an event-related potential study," *Journal of Cognitive Neuroscience*, vol. 10, no. 6, pp. 717–733, 1998.

[13] S. Koelsch, T. Gunter, A. D. Friederici, and E. Schröger, "Brain indices of music processing: "nonmusicians" are musical," *Journal of Cognitive Neuroscience*, vol. 12, no. 3, pp. 520–541, 2000.

[14] P. Regnault, E. Bigand, and M. Besson, "Different brain mechanisms mediate sensitivity to sensory consonance and harmonic context: evidence from auditory event-related brain potentials," *Journal of Cognitive Neuroscience*, vol. 13, no. 2, pp. 241–255, 2001.

[15] C. Magne, C. Astésano, M. Aramaki, S. Ystad, R. Kronland-Martinet, and M. Besson, "Influence of syllabic lengthening on semantic processing in spoken French: behavioral and electrophysiological evidence," *Cerebral Cortex*, 2007, Oxford University Press, January 2007.

[16] C. Astésano, *Rythme et accentuation en français: Invariance et variabilité stylistique*, Collection Langue & Parole, L'Harmattan, Paris, France, 2001.

[17] A. Di Cristo, "Le cadre accentuel du français contemporain: essai de modélisation: premiére partie," *Langues*, vol. 2, no. 3, pp. 184–205, 1999.

[18] G. Pallone, *Dilatation et transposition sous contraintes perceptives des signaux audio: application au transfert cinéma-vidéo*, Ph.D. thesis, University of Aix-Marseille II, Marseilles, France, 2003.

[19] M. Dolson, "The phase vocoder: a tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[20] G. Pallone, P. Boussard, L. Daudet, P. Guillemain, and R. Kronland-Martinet, "A wavelet based method for audio-video synchronization in broadcasting applications," in *Proceedings of the 2nd COST-G6 Workshop on Digital Audio Effects (DAFx '99)*, pp. 59–62, Trondheim, Norway, December 1999.

[21] M. Puckette, "Phase-locked vocoder," in *Proceedings of IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 222–225, New Paltz, NY, USA, October 1995.

[22] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[23] N. R. French and M. K. Zinn, "Method of an apparatus for reducing width of trans-mission bands," US patent no. 1,671,151, May 1928.

[24] S. Roucos and A. Wilgus, "High quality time-scale modification for speech," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '85)*, vol. 10, pp. 493–496, Tampa, Fla, USA, April 1985.

[25] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '93)*, vol. 2, pp. 554–557, Minneapolis, Minn, USA, April 1993.

[26] D. J. Hejna, B. T. Musicus, and A. S. Crowe, "Method for time-scale modification of signals," US patent no. 5,175,769, December 1992.

[27] J. Laroche, "Time and pitch scale modification of audio signals," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds., pp. 279–309, Kluwer Academic Publishers, Norwell, Mass, USA, 1998.

[28] B. Repp, "Probing the cognitive representation of musical time: structural constraints on the perception of timing perturbations," *Haskins Laboratories Status Report on Speech Research*, vol. 111-112, pp. 293–320, 1992.

[29] B. Moog, "MIDI: musical instrument digital interface," *Journal of Audio Engineering Society*, vol. 34, no. 5, pp. 394–404, 1986.

[30] M. S. Puckette, T. Appel, and D. Zicarelli, "Real-time audio analysis tools for Pd and MSP," in *Proceedings of the International Computer Music Conference*, pp. 109–112, International Computer Music Association, Ann Arbor, Mich, USA, October 1998.

[31] H. H. Jasper, "The ten-twenty electrode system of the International Federation," *Electroencephalography and Clinical Neurophysiology*, vol. 10, pp. 371–375, 1958.

[32] M. Kutas and S. A. Hillyard, "Reading senseless sentences: brain potentials reflect semantic incongruity," *Science*, vol. 207, no. 4427, pp. 203–205, 1980.

[33] M. Besson, C. Magne, and P. Regnault, "Le traitement du langage," in *L'imagerie fonctionnelle électrique (EEG) et magnétique (MEG): Ses applications en sciences cognitives*, B. Renault, Ed., pp. 185–216, Hermés, Paris, France, 2004.

[34] C. Astésano, M. Besson, and K. Alter, "Brain potentials during semantic and prosodic processing in French," *Cognitive Brain Research*, vol. 18, no. 2, pp. 172–184, 2004.

[35] C. Magne, C. Astésano, A. Lacheret-Dujour, M. Morel, K. Alter, and M. Besson, "On-line processing of "pop-out" words in spoken French dialogues," *Journal of Cognitive Neuroscience*, vol. 17, no. 5, pp. 740–756, 2005.

[36] M. Besson and F. Faïta, "An event-related potential (ERP) study of musical expectancy: comparison of musicians with non-musicians," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 21, no. 6, pp. 1278–1296, 1995.

[37] S. Koelsch, T. Gunter, E. Schröger, and A. D. Friederici, "Processing tonal modulations: an ERP study," *Journal of Cognitive Neuroscience*, vol. 15, no. 8, pp. 1149–1159, 2003.

[38] T. L. Luks, H. C. Nusbaum, and J. Levy, "Hemispheric involvement in the perception of syntactic prosody is dynamically dependent on task demands," *Brain and Language*, vol. 65, no. 2, pp. 313–332, 1998.

[39] R. J. Zatorre, "Neural specializations for tonal processing," in *The Biological Foundations of Music*, R. J. Zatorre and I. Peretz, Eds., vol. 930 of *Annals of the New York Academy of Sciences*, pp. 193–210, New York Academy of Sciences, New York, NY, USA, June 2001.

Cerebral Cortex November 2007;17:2659-2668
doi:10.1093/cercor/bhl174
Advance Access publication January 30, 2007

Influence of Syllabic Lengthening on Semantic Processing in Spoken French: Behavioral and Electrophysiological Evidence

Cyrille Magne[1,2], Corine Astésano[1,2], Mitsuko Aramaki[3], Sølvi Ystad[3], Richard Kronland-Martinet[3] and Mireille Besson[1,2]

[1]Institut de Neurosciences Cognitives de la Méditerranée, Centre National de la Recherche Scientifique and Marseille Universités, Marseille, France, [2]Université de la Méditerranée, Marseille, France and [3]Laboratoire de Mécanique et d'Acoustique, Centre National de la Recherche Scientifique, Marseille, France

The present work investigates the relationship between semantic and prosodic (metric) processing in spoken language under 2 attentional conditions (semantic and metric tasks) by analyzing both behavioral and event-related potential (ERP) data. Participants listened to short sentences ending in semantically and/or metrically congruous or incongruous trisyllabic words. In the metric task, ERP data showed that metrically incongruous words elicited both larger early negative and late positive components than metrically congruous words, thereby demonstrating the online processing of the metric structure of words. Moreover, in the semantic task, metrically incongruous words also elicited an early negative component with similar latency and scalp distribution as the classical N400 component. This finding highlights the automaticity of metrical structure processing. Moreover, it demonstrates that violations of a word's metric structure may hinder lexical access and word comprehension. This interpretation is supported by the behavioral data showing that participants made more errors for semantically congruous but metrically incongruous words when they were attending to the semantic aspects of the sentence. Finally, the finding of larger N400 components to semantically incongruous than congruous words, in both the semantic and metric tasks, suggests that the N400 component reflects automatic aspects of semantic processing.

Keywords: late positivity, meter, N400, prosody, semantics

Introduction

One of the main differences between the written and spoken forms of language is that spoken language conveys prosodic information. Prosody comprises intonation, accentuation, and rhythmic patterns that are produced by variations in acoustic parameters such as fundamental frequency (F0), intensity, duration, and spectral characteristics. Prosody plays both an emotional (Buchanan et al. 2000; Besson et al. 2002; Schirmer et al. 2002) and a linguistic function in spoken language. Here we focused on the latter. The linguistic function of prosody operates both at the lexical level, to facilitate word recognition (e.g., Cutler and Van Donselaar 2001; Soto-Faraco et al. 2001; Copper et al. 2002; Friedrich et al. 2004), and at the structural levels (see Cutler et al. 1997; Hirst and Di Cristo 1998; Cutler and Clifton 1999; Di Cristo 1999 for comprehensive reviews of the functions of prosody). The structural function of prosody in utterance parsing and hierarchical organization can be seen both at the syntactic (e.g., Marslen-Wilson et al. 1992; Beckman 1996; Pynte and Prieur 1996; Warren 1996) and discourse levels (Most and Slatz 1979; Birch and Clifton 1995). Thus, prosody operates at every level of language organization. However, although it is an important feature of language comprehension, prosody has, until recently, received less attention in the psycho- and neurolinguistic literature than other aspects, such as syntax or semantics.

During the past 7 years, several studies have effectively used the event-related potential (ERP) method to study the online processing of prosody during language comprehension by exploring the electrophysiological correlates of different aspects of prosody. At the lexical level, the metrical function of prosody was first examined by Böcker et al. (1999) through the rhythmic properties (metrical stress) of spoken words in Dutch. Participants either listened passively to, or discriminated between, sequences of 4 bisyllabic Dutch words, which started with either weakly (12% of the Dutch lexicon) or strongly stressed syllables (i.e., weak-initial vs. strong-initial words). Results showed that weak-initial words elicited a larger frontal negativity (denoted as N325) than strong-initial words, particularly in the discrimination task. The authors concluded that the N325 may reflect the extraction of metrical stress from the acoustic signal.

Recent results by Friedrich et al. (2004) suggest that F0 also influences word recognition. The authors presented a prime syllable, accented or unaccented, followed by a bisyllabic German word or nonword target which, in half of the trials, started with the same syllable as the prime syllable. The accentuation was materialized through the manipulation of F0. The participants' task was to determine whether the target was a word or a pseudoword. Results demonstrated that reaction times (RTs) were shorter, and the P350 component's amplitude smaller, when the accentual pattern of the bisyllabic target matched with the prime syllable (*RE* followed by *REgel* or *re* followed by *reGAL*) than when it did not match (e.g., *RE* followed by *reGAL*). Thus, accentuation seems to facilitate word recognition by activating the relevant lexical representation. Interestingly, unexpected F0 manipulations of words in sentence contexts also elicit increased positivities in the 300- to 600-ms latency range over the posterior region of the scalp in both adults (Schön et al. 2004) and children (Magne et al. 2006; Moreno and Besson 2006).

At the structural level, a few studies have sought to examine the relationship between prosody, on one hand, and syntax, pragmatics, or semantics, on the other hand, using the ERP method. Steinhauer et al. (1999), for instance, were the first to demonstrate that prosody directs syntactic parsing in the initial analysis of syntactically ambiguous sentences. A positive component closure positive shift (CPS) was elicited by intonation phrase boundaries of cooperating sentences, reflecting the online processing of prosodic structure. Moreover, Pannekamp et al. (2005) used hummed sentences (i.e., without lexical information, but with preserved intonation) to show that prosodic boundaries still elicited CPS components, thus suggesting that the CPS is directly linked to prosodic processing. More recently, Eckstein and Friederici (2005) provided evidence for late interactions between prosodic and syntactic processing by manipulating the position of words that were prosodically marked as sentence-final or sentence-penultimate words in syntactically correct or incorrect sentences. They found that a right anterior negative (RAN) component was elicited by the prosodic manipulations, independently of the syntactic correctness of the sentences.

In order to examine the relation between prosody and pragmatics, and to determine whether listeners make online use of focal prominences/accents to build coherent representations of the informational structure of speech, Magne et al. (2005) used short French dialogues comprised of a question and an answer presented aurally. By manipulating the position of focal accents in the answer, it was possible to render the prosodic patterns either coherent or incoherent with regard to the pragmatic context introduced by the question (e.g., "Did he give his fiancée a ring or a bracelet? He gave a RING to his fiancée" vs. "*He gave a ring to his FIANCEE"; capitalized word bearing a focal accent). Results showed that incoherent prosodic patterns elicited different ERP components depending upon their position within the answer. Although sentence-medial incongruous prosodic patterns elicited a P300-like component that was interpreted as reflecting a prosodic surprise effect, sentence-final incongruous prosodic patterns elicited an N400-like component, possibly reflecting enhanced lexical, semantic, and pragmatic integration difficulties.

Finally, Astésano et al. (2004) investigated the relationship between semantic and prosodic processing by studying the modality function of prosody. Based on the typical findings of increased F0 patterns for questions and decreased F0 patterns for statements (see Hirst and Di Cristo 1998 for a review), prosodic incongruities were created by cross-splicing the beginning of statements with the end of questions, and vice versa. Participants had to decide whether the aurally presented sentences were semantically or prosodically congruous in 2 different attentional conditions (attention to semantics or to prosody). Results showed that a left temporoparietal positive component (P800) was associated with prosodic incongruities, whereas a right centroparietal negative component (N400) was associated with semantic incongruities. Moreover, the P800 component to prosodic incongruities was larger when the sentences were semantically incongruous than congruous, suggesting interactive effects of semantic and prosodic processing. Interestingly, results also showed that the semantic incongruities elicited an N400 component regardless of the orientation of participants' attention (prosody or semantics), thereby suggesting that at least some aspects of semantic processing rely on automatic processes. By contrast, prosodic incongruities elicited a P800 component only when participants focused their attention on the prosodic aspect (modality) of sentences, which may also be linked with the increased difficulty in the detection of prosodic incongruities.

The general aims of the present work were to study the processing of the rhythmic properties of words, through the manipulation of syllabic duration, and the consequences of this manipulation on semantic processing of French spoken sentences. Traditionally, French is described as having fixed accents located at the end of rhythmic groups. This accent is characterized by a lengthening of the last syllable of a word or group of words (Delattre 1966) and contributes to the rhythmic organization of French (Wenk and Wioland 1982; Bailly 1989), as it does in other languages (e.g., Frazier et al. 2004). For instance, Salverda et al. (2003) were able to demonstrate, by recording eye movements, that syllabic lengthening influences lexical interpretation. Items were more likely to be interpreted as monosyllabic words (e.g., ham) than as bisyllabic words (e.g., hamster) when the duration of the first syllable was longer than when it was shorter. Moreover, there is behavioral evidence that French listeners use final syllabic lengthening to speed up detection of a target syllable located at a rhythmic-group boundary in comparison to the same syllable at another location (Dahan 1996). Recently, Christophe et al. (2004) have shown in an elegant study that final lengthening at phonological phrases boundaries facilitates the resolution of local lexical ambiguity. Thus, for instance, although the phonological phrase "un chat drogué" was processed faster than "un *chat grin*cheux," because of the ambiguity linked with the competitor word "chagrin" in the second example, no difference was found for the phrase "son grand chat grimpait," supposedly, because the ambiguity straddled the phonological phrase boundary.

The specific aim of the present experiment was 3-fold. First, we wanted to determine whether we could find ERP evidence for the online processing of misplaced stress accents in French. To this end, we created prosodic incongruities by applying the duration of the last syllable to the penultimate syllable of the trisyllabic final words of sentences. We chose to manipulate the second syllable because, according to French metric structure (Astésano 2001), this syllable is never stressed. Based on previous results using different types of prosodic incongruities in sentence or word contexts (Astésano et al. 2004; Friedrich et al. 2004; Schön et al. 2004; Magne et al. 2005, 2006), we predicted that a metric violation of syllabic duration would produce increased late positivities. However, based on the results of Böcker et al. (1999) and Eckstein and Friederici (2005) mentioned above, such an irregular stress pattern may also elicit an increased negativity (e.g., N325 or RAN).

Second, we aimed to determine whether the metric and semantic aspects of spoken language are processed independently or in interaction. Thus, the semantic and metric aspects were manipulated orthogonally, so as to create 4 conditions in which sentences were 1) S+M+, both semantically and metrically congruous, 2) S+M–, semantically congruous and metrically incongruous, 3) S–M+, semantically incongruous and metrically congruous, and finally 4) S–M–, both semantically and metrically incongruous (see Table 1).

Finally, in different blocks of trials, participants were asked to focus their attention on the metrical structure or on the semantics of the final words of the sentence to decide whether the final words were metrically congruous or incongruous (metric task) or semantically congruous or incongruous (semantic task). Two questions were of main interest. First, would metric incongruities be associated with similar electrophysiological effects when attention is focused on meter as on semantics? Conversely, would semantic incongruities generate N400 components independently of the direction of attention? This is an important and difficult issue because contradictory results have been reported in the literature. For instance, although Chwilla et al. (1995) found evidence that the N400 component may reflect controlled aspects of the integration of word meaning, Astésano et al. (2004) reported that semantic processing may occur automatically, even when not relevant to the task at hand.

Table 1
Examples of stimuli used in the 4 experimental conditions

	Semantically congruous (S+)	Semantically incongruous (S−)
Metrically congruous (M+)	*Le concours a regroupé mille candidats* "The competition hosted a thousand candidates"	*Le concours a regroupé mille bigoudis* "The competition hosted a thousand curlers"
Metrically incongruous (M−)	*Le concours a regroupé mille candidats* "The competition hosted a thousand candidates"	*Le concours a regroupé mille bigoudis* "The competition hosted a thousand curlers"

Note: The lengthened syllable is underlined.

Methods

Participants

Fourteen participants (7 females, mean age 26, age range 23-31), gave their informed consent, and were paid to participate in the experiment, which lasted for about 2 h. All were right-handed native speakers of French, without hearing or neurological disorders.

Stimuli

A total of 512 experimental sentences were built in such a way that they all ended with a trisyllabic noun. Among the 512 sentences, 256 ended with semantically congruous words (S+) and 256 ended with semantically incongruous words (S–, see Table 1). Semantically incongruous sentences were built by replacing the final congruous word with a word that shared the same acoustic and phonological characteristics but that did not make sense in the sentence context. Moreover, semantically congruous and incongruous sentence-final words were matched for word frequency (92.38 and 91.36 occurrences per million, respectively), using the LEXIQUE2 French lexical database (New et al. 2001). Within each semantic condition, half of the sentences ended with an unmodified word with natural lengthening of the last syllable (M+). The other half of the sentences ended with an incongruous lengthening of the penultimate syllable of the final trisyllabic word (M–). This syllabic lengthening was created by increasing the duration of the vowel of the penultimate syllable, using a time-stretching algorithm which allows for the manipulation of the duration of acoustic signals without modifying their timbre or frequency (Pallone 1999 and see below).

Each of the 4 experimental conditions (S+M+, S–M+, S+M–, and S–M–) comprised 32 different sentences (sound examples illustrating stimuli in each experimental condition are available at http://www.lma.cnrs-mrs.fr/~ystad/CerebralCortex/CerebralCortex.html). Four different experimental lists of 128 sentences were built in order to present each sentence in each experimental condition across subjects, with no repetition within subjects.

Speech Signal

The 256 prosodically congruous sentences were spoken by a native male speaker of standard French and recorded in an anechoic chamber using a digital audiotape (sampling at 44.1 kHz). The mean duration of the sentences was 2.8 s (standard deviation [SD] = 0.9 s), and the mean speech rate was 5.35 syllables per second (SD = 0.65). All sentences were spoken in a declarative mode, and the pitch contour was always falling at the end of the sentence.

For all sentences, the final word comprised 3 syllables and was chosen according to the following constraints. The second and third syllables all possessed a consonant–vowel structure (X-CV-CV, where X could be CV or V). In addition, the second and third syllables never contained a nasal vowel (e.g., [ɔ̃], [ã], [ɛ̃], [œ̃]), which are known to have longer durations than the non-nasal vowels in French (e.g., [a], [a], [i], [e], [ɛ], [ə], [o], [u], [y], [ø], [œ], [ɔ]; Astésano 2001). We also avoided using trisyllabic words with liquid consonants ([l] or [r]) as much as possible, as they are more difficult to segment. All final words were segmented manually by a professional phonetician. The mean duration of final words was 496 ms (SD = 52 ms). On average, the first syllable was 150 ms long (SD = 28 ms), the second syllable was 145 ms long (SD = 28 ms), and the third syllable was 202 ms long (SD = 42 ms). Finally, and as determined from the French lexical database LEXIQUE2 (New et al. 2001), the mean number of phonemes is equal to 6.2, and the mean phonological unicity point (i.e., the rank of the phoneme from which the word can be identified without any ambiguity) is 5.4, which corresponds to an approximate duration of 400 ms (i.e., the unicity point is located, on average, between the onset and offset of the third syllable).

The choice of the lengthening factor that was used to create the metric incongruity was constrained by the necessity of remaining ecological (i.e., to keep the naturalness of speech). The ratio of lengthening for final syllables in French is known to vary considerably as a function of factors such as depth of the adjacent prosodic boundary and speech style (Astésano 2001). In our materials, the lengthening ratio used was equal to 1.7, which is within the range found for natural syllable lengthening of the last syllable in spoken French (1.7–2.5; Astésano 2001). The lengthening of the penultimate syllable creates the auditory impression of a misplaced accented syllable in the words while remaining ecological.

In fact, note that changing the duration of a signal without modifying its frequency is an intricate problem. It was therefore necessary to construct a time-stretching algorithm to modify the syllable length. We decided to use a time-domain approach, in which the signal is time stretched by accurately adding short, nonmodified segments of the original time signal. Consequently, the F0 and amplitude contours of the stretched syllable are identical to the F0 and amplitude contours of the unmodified syllable, but they unfold more slowly over time (i.e., the rate of F0 and amplitude variations differs between the metrically congruous and incongruous conditions; see Fig. 1*A*,*B*). The choice of the segments, together with the choice of the position of the insertion, is crucial for the quality of the resulting signal. To ensure that no audible rhythmic defaults or discontinuities occurred, we avoided duplicating the transient part of the signal and constructed segments containing a whole number of periods for periodic signals. To this aim, we built an algorithm, derived from Synchronous Overlap and Add (SOLA) methods (WSOLA and SOLAFS, Dattorro 1987; Laroche 1993), that calculated the ideal duration of the segment to be inserted (see Pallone [1999] for more details on the algorithm used in the present study).

Because the algorithm stretched the F0 and intensity contours from the beginning of the second syllable, we conducted statistical analyses to determine when these acoustical parameters started to be significantly different between lengthened words and their normal version. For both metrically congruous and metrically incongruous words, F0 and intensity values were extracted using 20 ms steps in between the beginning of the second syllable (considered as time 0 ms) until 100 ms later. Then, the differences between these F0/intensity values and the F0/intensity values at 0 ms were computed. Finally, paired *t*-tests were conducted between the values for the normal and for the lengthened words. Results revealed that F0 values started to differ significantly between metrically incongruous and metrically congruous words from 80 ms after second syllable onset ($t_{255} = 2.20$, $P < 0.03$) and intensity values from 60 ms after second syllable onset ($t_{255} = 3.28$, $P < 0.01$).

Procedure

Participants were presented with 128 short sentences that were semantically and/or metrically congruous or incongruous. Each experiment began with a practice session to familiarize participants with the task and to train them to blink during the interstimulus interval. The sentences were presented aurally, through headphones, in a pseudorandom order, within 4 blocks of 32 trials each. In 2 blocks, participants were asked to pay attention only to the semantic content in order to decide whether the last word of each sentence was semantically congruous or incongruous. In the other 2 blocks, participants were asked to pay attention only to the syllabic duration in order to decide

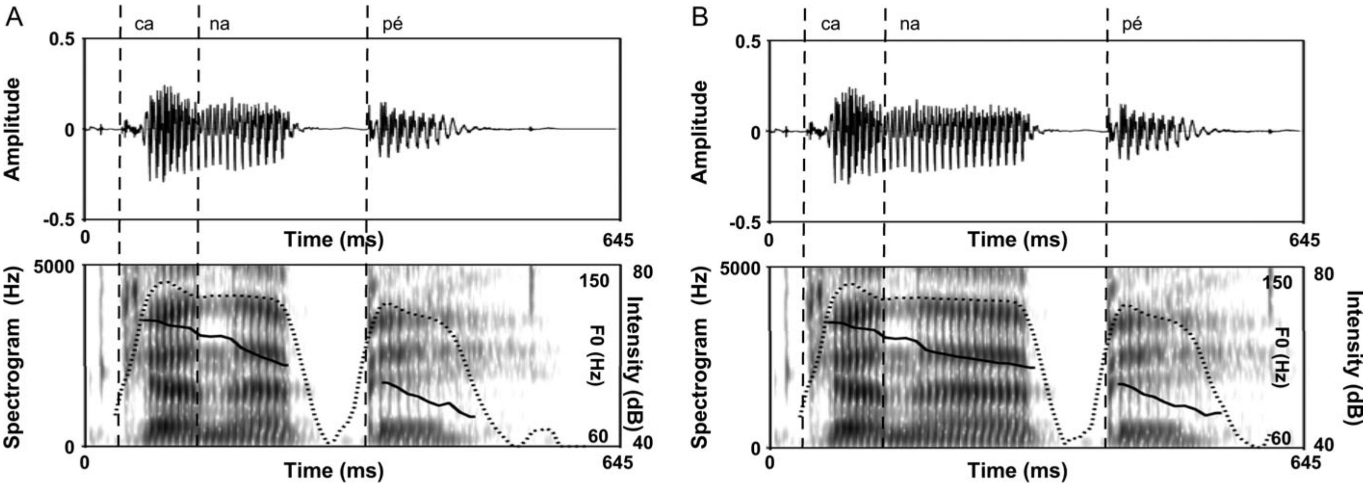


Figure 1. (*A*) Original version of the word "canapé." (*B*) Time-stretched version of the word "canapé." On both panels, the waveform is represented on the top part; the spectrogram (gray scale), the F0 (solid line), and the intensity (dashed line) are represented on the bottom part.

whether the last word of each sentence was well pronounced or not. Participants were required to press one of 2 buttons as quickly and accurately as possible to give their response. The side (right or left hand) of the response was balanced across participants. Furthermore, half of the participants began with the semantic task and the other half with the metric task.

ERP Recordings

EEG was recorded for 2200 ms, starting 200 ms before the onset of the last word from 28 scalp electrodes, mounted on an elastic cap, and located at standard left and right hemisphere positions over frontal, central, parietal, occipital, and temporal areas (International 10/20 system sites: Fz, Cz, Pz, Oz, Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fc5, Fc1, Fc2, Fc6, Cp5, Cp1, Cp2, Cp6). These recording sites plus an electrode placed on the right mastoid were referenced to the left mastoid electrode. The data were then rereferenced offline to the algebraic average of the left and right mastoids. Impedances of the electrodes never exceeded 3 kΩ. In order to detect horizontal eye movements and blinks, the horizontal electrooculogram (EOG) was recorded from electrodes placed 1 cm to the left and right of the external canthi, and the vertical EOG was recorded from an electrode beneath the right eye, referenced to the left mastoid. Trials containing ocular artifacts, movement artifacts, or amplifier saturation were excluded from the averaged ERP waveforms. The EEG and EOG were amplified by a SA Instrumentation amplifier with a band pass of 0.01–30 Hz and were digitized at 250 Hz by a PC-compatible microcomputer.

Data Analyses

Behavioral data (error rates and RTs) were analyzed using a 3-way analysis of variance (ANOVA), including task (metric or semantic), meter (congruous vs. incongruous), and semantics (congruous vs. incongruous) as within-subject factors. Moreover, mean amplitude ERPs to final words were measured in several latency bands (100–250, 250–450, 500–800, and 800–1200 ms) determined both from visual inspection and from the results of consecutive analyses of 50-ms latency widows. Results were analyzed using ANOVAs that included the same factors as above (task, meter, and semantics) plus electrodes (Fz, Cz, Pz, Oz) for midline analyses and hemispheres (left vs. right), regions of interest (3 ROIs: frontocentral, temporal, and parietotemporal), and electrodes (3 for each ROI: F3, F7, Fc1/F4, F8, Fc2; Fc5, C3, Cp5/Fc6, C4, Cp6; and Cp1, P3, T5/Cp2, P4, T6) for lateral electrodes. When interactions between 2 or more factors were significant, post hoc comparisons between relevant condition pairs were computed. All *P* values were adjusted with the Greenhouse–Geisser epsilon correction for nonsphericity when necessary.

Results

Behavioral Data

Results of a 3-way ANOVA on the transformed percentages of errors showed no significant main effect of task, meter, or semantics. The meter by semantics interaction was significant ($F_{1,12}$ = 16.37, $P < 0.001$): regardless of the direction of attention, participants made more errors when one dimension, meter (19.5%) or semantics (20%), was incongruous than when both dimensions were congruous (12%) or incongruous (16.5%; see Table 2). Finally, the task by meter by semantics interaction was also significant ($F_{1,12}$ = 4.74, $P < 0.05$): in the semantic task, participants made more errors when semantics was congruous, but meter was incongruous (S+M–) than in the other 3 conditions.

Results of a 3-way ANOVA on the RTs showed a main effect of semantics ($F_{1,12}$ = 53.70, $P < 0.001$): RTs were always shorter for semantically congruous (971 ms) than incongruous words (1079 ms; see Table 2). No other effect reached significance.

Electrophysiological Data

Results of the main ANOVAs in the different latency ranges are presented in Table 3. When the main effects or relevant interactions are significant, results of 2 by 2 comparisons are reported in the text.

Prior to 250 ms from final word onset, no significant differences were found either at midline or lateral electrodes. In the 250- to 450-ms range, and as can be seen on Figure 2, semantically incongruous words elicited larger negative components than semantically congruous words in both the semantic and the metric tasks and at both midline (difference [*d*] = –1.53 μV) and lateral electrodes (*d* = –1.36 μV; main effect of semantics and no task by semantics interaction, see Table 3). This N400 effect was broadly distributed over the scalp (no semantics by electrodes or by hemispheres/ROIs interactions, see Table 3 and Fig. 3). Interestingly, metrically incongruous words also elicited larger negative components than metrically congruous words in both tasks and at both midline (*d* = –1.13 μV) and lateral electrodes (*d* = –1.23 μV; main effect of meter and no task by meter interaction, see Table 3 and Fig. 4). Note,

Table 2

Mean error rates (% Err) in % and mean RTs in ms for each of the 4 experimental conditions (S+M+, S−M+, S+M−, and S−M−) in both the metric and semantic tasks

	Metric				Semantic			
	S+M+	S−M+	S+M−	S−M−	S+M+	S−M+	S+M−	S−M−
% Err	13 (16)	19 (16)	15 (14)	18 (15)	11 (6)	21 (18)	24 (18)	15 (11)
RTs	932 (152)	1055 (179)	966 (102)	997 (135)	975 (139)	1127 (184)	1011 (164)	1138 (146)

Note: The SD is indicated in parentheses.

Table 3

Results of ANOVAs computed on midline and lateral electrodes

Electrodes	Factors	df	Latency windows (ms)							
			100–250		250–450		500–800		800–1200	
			F	*P*	*F*	*P*	*F*	*P*	*F*	*P*
Midline	T	1,13	*F* < 1		*F* < 1		*F* < 1		*F* < 1	
	S	1,13	*F* < 1		**10.15**	**0.007**	**9.14**	**0.009**	*F* < 1	
	M	1,13	*F* < 1		**6.17**	**0.027**	2.05	0.17	1.76	0.20
	T × M	1,13	*F* < 1		*F* < 1		**9.68**	**0.008**	2.39	0.14
Lateral	T	1,13	*F* < 1		*F* < 1		*F* < 1		*F* < 1	
	S	1,13	*F* < 1		**4.64**	**0.05**	**4.82**	**0.04**	*F* < 1	
	M	1,13	*F* < 1		**10.01**	**0.007**	*F* < 1		*F* < 1	
	T × M	1,13	*F* < 1		*F* < 1		**9.47**	**0.008**	*4.3*	*0.06*
	M × H	1,13	*F* < 1		*4.53*	*0.06*	*F* < 1		*F* < 1	

Note: df, degrees of freedom; T, task; S, semantic congruity; M, metric congruity; H, hemisphere. Significant effects are marked in bold and marginally significant ($P = 0.06$) effects in italics. Values for interactions that were never significant in any of the latency bands considered are not reported.

however, that the meter by hemisphere interaction was almost significant ($P < 0.06$): the amplitude of the negative components was somewhat larger over the right hemisphere (metrically congruous vs. incongruous: $F_{1,13} = 15.95$, $P = 0.001$; $d = -1.69$ μV) than over the left hemisphere (metrically congruous vs. incongruous: $F_{1,13} = 6.04$, $P = 0.03$, $d = -1.11$ μV; see Fig. 5).

In the 500- to 800-ms range, and as can be seen on Figures 2 and 3, semantically incongruous words still elicited relatively larger negativities than semantically congruous words in both the semantic and metric tasks and at both midline ($d = -2.73$ μV) and lateral electrodes ($d = -2.5$ μV; main effect of semantics and no task by semantics interaction, see Table 3). By contrast, metrically incongruous words elicited larger positivities than metrically congruous words only in the metric task (no significant main effect of meter but significant task by meter interactions, see Table 3). In the metric task, the metric congruity effect was significant at both midline ($d = 3.07$ μV; $F_{1,13} = 8.53$, $P = 0.01$) and lateral electrodes ($d = 2.39$ μV; $F_{1,13} = 9.47$, $P = 0.008$, see Figs 4*A* and 5*A*).

Finally, in the 800- to 1200-ms range, metrically incongruous words were still somewhat more positive than metrically congruous words only in the metric task (marginally significant task by meter interaction) and at lateral electrodes ($d = 2.05$ μV; $F_{1,13} = 3.98$, $P = 0.06$, see Table 3 and Fig. 4*A*).

Discussion

Semantic Congruity Effect

In the semantic task, final semantically incongruous words with respect to the sentence context elicited larger N400 components than semantically congruous words in the 250- to 450-ms latency band. This semantic congruity effect showed a broad distribution over scalp sites, with a slight centroparietal maximum (see Fig. 3*B*). These differences extended in the 500- to 800-ms latency range with semantically incongruous words still being associated with relatively larger negativities than semantically congruous words. These results are in line with the literature (Kutas and Hillyard 1980; see Kutas and Federmeier 2000; Besson et al. 2004 for recent reviews) and have been interpreted as reflecting the greater difficulties encountered either in integrating semantically incongruous compared with congruous words in ongoing sentence contexts or in generating expectancies for semantically incongruous compared with congruous words, with recent results favoring the latter interpretation (DeLong et al. 2005). This interpretation is also in line with the behavioral data showing longer RTs to semantically incongruous than congruous words. Regarding the time course of the N400 effect, it is interesting to note that, according to acoustic analyses, the mean duration of the final words is 496 ms (SD = 52 ms). Moreover, the phonological unicity point is 5.4 phonemes (mean number of phonemes is 6.2), which corresponds to a word duration of approximately 400 ms (i.e., the isolation point—when the word can be recognized unambiguously—is located between the onset and offset of the third syllable). Clearly, the N400 effect starts before the isolation point, a result in line with the literature (Van Petten et al. 1999, van Berkum et al. 2003, van den Brink et al. 2006). Indeed, van den Brink et al. (2006) have recently shown that the onset of the N400 effect not only occurs prior to the isolation point of sentence-final words but also that the onset of the N400 effect is unaffected by the position (early vs. late) of the isolation point.

Interestingly, semantically incongruous words also elicited larger N400 components than congruous words in the metric task (no task by semantics interaction). Moreover, the scalp distribution of the semantic congruity effect was not significantly different in the metric and semantic tasks; this finding is taken to reflect the similarity of the semantic congruity effect in both tasks (see Fig. 3). Close inspection of Fig. 2 shows that at the F4 electrode, the N400 to semantically incongruous words

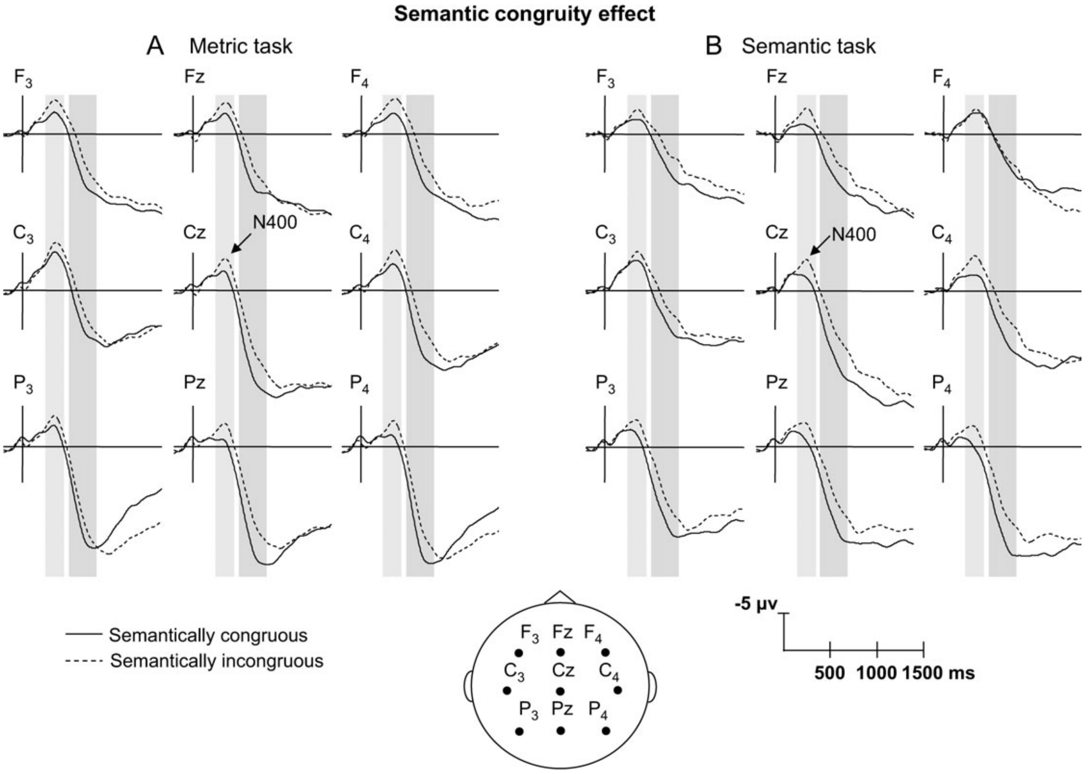


Figure 2. Averaged electrophysiological data time locked to the onset of semantically congruous (solid line) or semantically incongruous (dashed line) final words, in the metric (*A*) and semantic tasks (*B*). Selected traces from 9 electrodes are presented. The latency ranges within which statistical analyses revealed significant effects are shown in gray (250- to 450-ms and 500- to 800-ms latency ranges). In this figure, as in the following ones, the amplitude (in microvolts) is plotted on the ordinate (negative up) and the time (in milliseconds) is on the abscissa.

is larger in the metric than in the semantic task. Because this point is important for the discussion of the automaticity of the N400 component, we computed ANOVAs including F4 only: results showed that the task by semantics interaction was not significant (F1,13 = 1.32, P = 0.27). This finding is taken to reflect the similarity of the semantic congruity effect in both tasks (see Fig. 3). Thus, participants seem to process the meaning of words, even when instructed to focus attention on syllabic duration. These results are in line with those of Astésano et al. (2004) showing the occurrence of N400s to semantic incongruities independently of whether participants focused their attention on the semantic or prosodic aspects (modality contour, interrogative, or declarative) of the sentences.

The finding of a semantic congruity effect when participants were focusing attention on meter seems to argue in favor of the automaticity of semantic processing. This, however, is a complex issue, mainly because it is not clear whether the N400 reflects the automatic or controlled aspects of semantic processing. Indeed, quite mixed evidence can be found in the literature, even when experiments are based on similar designs. For instance, different results have been reported in studies using a semantic priming paradigm, in which a prime word and a target word, which are semantically related or not, are successively presented and listeners perform tasks that require focusing attention on the semantic, lexical, or physical characteristics of words. Besson et al. (1992) found larger N400s to semantically unrelated compared with related visually presented words when the task was to decide whether the prime and target words shared the same first and final letters. This was taken as evidence that the N400 reflects automatic aspects of semantic processing. By contrast, Chwilla et al. (1995) used a semantic priming paradigm and 2 discrimination tasks based on either the semantic aspects (lexical decision task) or the physical aspects (uppercase vs. lowercase letters) of words presented visually. Results showed an N400 priming effect only in the lexical decision task. In the physical task, a P300 effect was found for both related and unrelated targets. Thus, in contrast to the Besson et al. (1992) results above, these results suggest that the N400 effect reflects controlled aspects of semantic processing. This conclusion was in line with the results

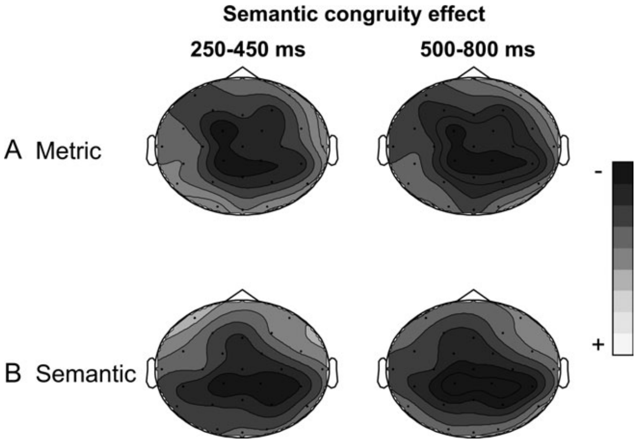


Figure 3. Topographic maps of the semantic congruity effect in the metric (*A*) and semantic tasks (*B*). Mean amplitude differences between semantically incongruous and congruous words in the 250- to 450-ms and 500- to 800-ms latency ranges.

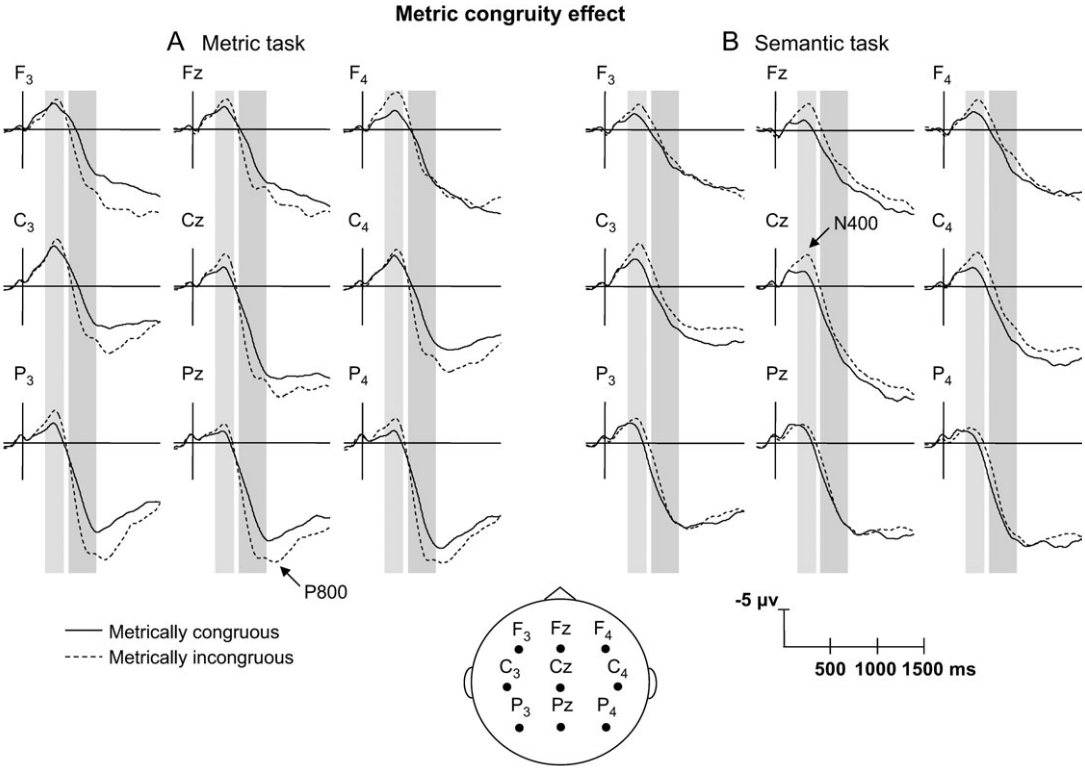


Figure 4. Averaged electrophysiological data time locked to the onset of metrically congruous (solid line) or metrically incongruous (dashed line) final words, in the metric (*A*) and semantic tasks (*B*).

of Brown and Hagoort (1993). Using a very well controlled masked priming design, they were able to demonstrate evidence for automatic semantic processing in RTs but found no ERP difference between semantically related and unrelated words (no N400 effect). However, Deacon et al. (2000) did recently investigate this issue further. Using a design very similar to the one used by Brown and Hagoort (1993), but within subjects, they did find evidence for semantic priming when participants could not consciously identify the prime words. Deacon et al. (2000) pointed out that because Brown and Hagoort (1993) used a between-subjects design, between-group differences may explain their different findings with masked and unmasked priming. This is in line with other recent results that also did not replicate the findings of Brown and Hagoort (1993) and suggested that automatic processes alone are sufficient to elicit N400 priming effects (Kiefer and Spitzer 2000; Heil et al. 2004). Finally, it should be noted that although the results reported above were found in the visual modality, the present results, together with those of Astésano et al. (2004), were obtained in the auditory modality. This opens the interesting possibility that semantic priming may be more automatic in the auditory than visual modality (Perrin and Garcia-Larrea 2003). This issue will be investigated in further experiments.

Metric Congruity Effect

In the metric task, metrically incongruous words also elicited larger negative components than metrically congruous words in the 250- to 450-ms range (see Fig. 4*A*), and this metric congruity effect was somewhat larger over the right than the left hemisphere (meter by hemisphere interaction, $P < 0.06$, see Fig. 5*A*). These results show that listeners are sensitive to the metric structure of words and perceive, in real time, the unusual but still ecological lengthening of the penultimate syllable. Regarding the time course of the metric congruity effect, it is striking that this effect starts around 100 ms after the onset of the second syllable (around 250 ms after final word onset, see Methods), which is earlier than the offset time of the stretched second syllable. This is to be expected, however, because statistical analyses conducted on F0 and intensity values of the last words revealed that F0 and intensity contours of the stretched second syllable and of the natural unstretched version

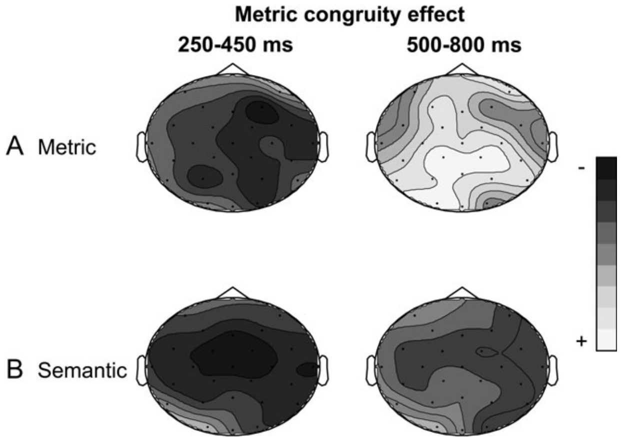


Figure 5. Topographic maps of the metric incongruity effect in the metric (*A*) and semantic tasks (*B*). Mean amplitude difference between metrically incongruous and metrically congruous words in the 250- to 450-ms and 500- to 800-ms latency ranges.

differed significantly as early as 80 ms and 60 ms, respectively, from the second syllable onset (one could then argue that participants were not sensitive to syllabic duration per se but to the stretched F0 and intensity contours. These aspects are, however, intrinsically linked because a natural slowing down necessarily induces a slowing down in F0 and intensity contours). Thus, because the second syllable varies in duration, F0 and intensity between metrically incongruous and metrically congruous words, the early negative effect may reflect the processing of low-level acoustic factors. However, although this may be the case, negative components have also been reported in the literature when controlling for the influence of such acoustic factors. For instance, in an experiment designed to examine prosodic focus, Magne et al. (2005) found a negativity, with a maximum amplitude around 400 ms (N400), to prosodically incongruous sentence-final words. Because these prosodically incongruous words were prosodically congruous in other discourse contexts, the N400 did not reflect the processing of low-level acoustic factors per se but rather the linguistic relevance of these acoustic cues.

Moreover, as mentioned in the Introduction, Eckstein and Friederici (2005) found a RAN to sentence-final words that were prosodically marked as penultimate words and were consequently prosodically incongruous. Again, Eckstein and Friederici controlled for low-level acoustic factors that, therefore, did not explain the occurrence of the RAN. Thus, the occurrence of the negative components rather seems to reflect the linguistic consequences of unexpected variations in the prosodic features of the words. Finally, the results recently reported by Strelnikov et al. (2006) are important for our understanding of the anatomo-functional basis of prosodic processing. They conducted experiments, using positron emission tomography (PET) and ERPs, that aimed at examining the influence of prosodic cues on sentence segmentation in Russian, a language in which word order is not as informative as it is in English or Romance languages. Interestingly, they found a right frontal negativity to words that preceded an intonational pause whose position in the sentence determined its meaning. Moreover, PET data revealed that the right dorsolateral prefrontal cortex and the right cerebellum were involved in prosodic segmentation. Thus, though it is difficult to infer the neural basis of the right anterior negativities directly from their scalp topography, it is interesting to note that, taken together, the results described above show converging evidence for the implication of right anterior brain structures in the processing of different aspects of prosody (see also Meyer et al. 2002, 2003, 2004).

Interestingly, the metric congruity effect was also significant when participants focused attention on the semantic aspects of the sentences (no task by metric congruity interaction). From these results, it may be concluded that participants did notice that the metric structure of words was incongruous independently of the direction of attention. Thus, as suggested above for the semantic congruity effect, the present results may be taken to reflect the automatic nature of metric processing. Although this may be the case, an alternative and possibly complementary interpretation should also be considered. A close inspection of Figures 4 and 5 indicates that the metric congruity effect shows a right hemisphere lateralization in the metric task but a more bilateral distribution in the semantic task. Such differences in scalp distribution may therefore reflect qualitative differences in the underlying processes. One interesting possibility is that violations of the metric structure interfere with lexical access and thereby hinder access to word meaning. The influence of metric properties of words on their lexical access has been shown in Spanish (Soto-Faraco et al. 2001), German (Friedrich et al. 2004), English (Copper et al. 2002), and Dutch (Cutler et al. 2001). Our present results show that metric properties also influence lexical access in French. Moreover, they suggest that such integration problems would be reflected by typically bilateral N400 components. Such an interpretation is in line with the behavioral data showing that semantically congruous but metrically incongruous words produced the highest error rates in the semantic task (see Table 2), as if metric incongruities disrupted semantic processing.

At present, we can only speculate about the possible mechanisms responsible for such effects. It may be that participants processed the metrically violated trisyllabic words as if they were metrically correct bisyllabic words with final syllabic lengthening, as shown by Salverda et al. (2003) for mono- and bisyllabic words. An N400 would be generated because such bisyllabic patterns are not real French words. An alternative interpretation is that the N400 is elicited by a prosodic mismatch between the expectation of finality caused by the lengthened second syllable and the following unexpected third syllable. More generally, and based on spoken word recognition models (e.g., Marslen-Wilson and Welsh 1978; McClelland and Elman 1986; McQueen et al. 1994), one may consider that none of the lexical candidates that were activated on the basis of the acoustic properties of the word fit with the violated metrical structure that was heard, thereby hindering access to word meaning and increasing integration costs. In any event, the most important point may be that results of several experiments converge in showing that syllabic lengthening (Salverda et al. 2003; Eckstein and Friederici 2005), correct accentuation (Cutler et al. 2001; Soto-Faraco et al. 2001; Copper et al. 2002; Friedrich et al. 2004), and initial phonological overlap (Connolly and Philips 1994; Van Petten et al. 1999; van den Brink et al. 2001; van den Brink and Hagoort 2004) influence lexical interpretation and lexical access.

Metrically incongruous words also elicited larger late positive components (although late positivities differ in amplitude across conditions, they were nevertheless always present. This is to be expected insofar as participants were asked to make a decision on the sentence final words [Kutas et al. 1977; McCarthy and Donchin 1979; Ragot and Renault 1981; Donchin and Coles 1988]) than metrically congruous words, but only in the metric task (see Figs 4*A* and 5*A*). These positivities developed in the 500- to 800-ms latency band and were still significant in the 800- to 1200-ms range. They were broadly distributed across scalp sites. This finding is in line with previous ones showing that the manipulation of different acoustic parameters of the speech signal, such as F0 and intensity, is also associated with increased positivity (Astésano et al. 2004; Friedrich et al. 2004; Schön et al. 2004; Eckstein and Friederici 2005; Magne et al. 2005, 2006; Moreno and Besson 2006). It has been proposed that these late positivities may belong to the P300 family of components elicited by surprising and task-relevant events (Donchin 1981; Donchin and Coles 1988). Indeed, manipulations of acoustic properties of the speech signal were unexpected and may have therefore produced a prosodic surprise effect in the listeners. Moreover, late positivities in sentence contexts are interpreted as reflecting the integration of syntactic, semantic, and pragmatic information (e.g., Kaan et al. 2000). The enhanced positivity reported here may therefore reflect the integration difficulties resulting from the violation of the typical metric pattern of final sentence words in French. Interestingly, no such increase in positivity to metrically incongruous words was observed when participants focused their attention on the semantics of the word (see Figs 4*B* and 5*B*). These results are in line with those reported by Astésano et al. (2004) and showing that violations of the intonation pattern of interrogative and declarative sentences elicit larger late positivities only when participants focused their attention on prosody. Thus, both types of prosodic violations (metric and modality) are associated with increased late positivities only when relevant to the task at hand.

Summary and Conclusion

In accordance with previous findings (e.g., Böcker et al. 1999; Salverda et al. 2003; Eckstein and Friederici 2005), the present results show evidence for the online processing of misplaced stress accents in French, through the occurrence of increased early negative and late positive components to metric incongruities in the metric task. In addition, in the semantic task, metrically incongruous words also elicited a negative component that may be considered as a member of the N400 family due to its N400-like latency, amplitude, and scalp distribution. This interpretation is supported by the present behavioral data. In particular, significant interactions found in statistical analyses of error rates showed that participants often (mistakenly) judged the metrically incongruous words as semantically incongruous even when they were attending to semantics and thus instructed to ignore the meter. Furthermore, previous findings showing enhanced negativities to prosodic incongruities when basic acoustic features of stimuli are controlled (e.g., Magne et al. 2005, Eckstein and Friederici 2005) lead us to believe that violations of prosodic expectancies, in this case inappropriately lengthened syllables, are manifested in negative ERP components that do not simply reflect differences in acoustic input but also represent significant interactions with lexical and/or semantic processes. Therefore, the present results highlight the importance of the metric structure of words for language comprehension in French. Finally, the similarity of the semantic congruity effects in both the semantic and metric tasks provides additional evidence for the N400 component's sensitivity to the automatic aspect of semantic processing.

Notes

This research was supported by a grant from the Human Frontier Science Program (HSFP #RGP0053) to Mireille Besson. Cyrille Magne benefited from a Cognitive Science fellowship from the French Ministry of Research, and Corine Astésano was a postdoctorate researcher supported by a grant from HFSP. The authors gratefully acknowledge Mario Rossi, Anne Lacheret, Michel Morel, and Valerie Pasdeloup for stimulating discussions, Monique Chiambretto and Reyna Leigh Gordon for their technical assistance, and 3 anonymous reviewers for their precious comments. *Conflict of Interest*: None declared.

Address correspondence to Cyrille Magne, Center for Complex Systems and Brain Sciences, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA. Email: magne@ccs.fau.edu.

References

Astésano C. 2001. Rythme et accentuation en français. Invariance et variabilité stylistique. Paris: L'Harmattan.

Astésano C, Besson M, Alter K. 2004. Brain potentials during semantic and prosodic processing in French. Brain Res Cogn Brain Res. 18:172-184.

Bailly G. 1989. Integration of rhythmic and syntactic constraints in a model of generation of French prosody. Speech Comm. 8: 137-146.

Beckman ME. 1996. The parsing of prosody. Lang Cogn Process. 11:17-67.

Besson M, Fischler I, Boaz T, Raney G. 1992. Effects of automatic associative activation on explicit and implicit memory tests. J Exp Psychol Learn Mem Cogn. 18:89-105.

Besson M, Magne C, Regnault P. 2004. Le traitement du langage. In: Renault B, editor. L'imagerie fonctionnelle électrique (EEG) et magnétique (MEG): Ses applications en sciences cognitives. Paris: Hermés. p. 185-216.

Besson M, Magne C, Schön D. 2002. Emotional Prosody: sex differences in sensitivity to speech melody. Trends Cogn Sci. 6:405-407.

Birch S, Clifton CE. 1995. Focus, accent, and argument structure: effects on language comprehension. Lang Speech. 38:365-391.

Böcker KBE, Bastiaansen MCM, Vroomen J, Brunia CHM, De Gelder B. 1999. An ERP correlate of metrical stress in spoken word recognition. Psychophysiology. 36:706-720.

Brown CM, Hagoort P. 1993. The processing nature of the N400: evidence from masked priming. J Cogn Neurosci. 5:34-44.

Buchanan TW, Lutz K, Mirzazade S, Specht K, Shah NJ, Zilles K, Jancke J. 2000. Recognition of emotional prosody and verbal components of spoken language: an fMRI study. Brain Res Cogn Brain Res. 9:227-238.

Christophe A, Peperkamp S, Pallier C, Block E, Mehler J. 2004. Phonological phrase boundaries constrain lexical access: I. adult data. J Mem Lang. 51:523-547.

Chwilla DJ, Brown CM, Hagoort P. 1995. The N400 as a function of the level of processing. Psychophysiology. 32:274-285.

Connolly JF, Phillips NA. 1994. Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. J Cogn Neurosci. 6:256-266.

Copper N, Cutler A, Wales R. 2002. Constraints of lexical stress on lexical access in English: evidence from native and non-native listeners. Lang Speech. 45:207-228.

Cutler A, Clifton C. 1999. Comprehending spoken language: a blueprint of the listener. In: Brown CM, Hagoort P, editors. The neurocognition of language. Oxford: Oxford University Press. p. 123-166.

Cutler A, Dahan D, van Donselaar W. 1997. Prosody in the comprehension of spoken language: a literature review. Lang Speech. 40:141-201.

Cutler A, Van Donselaar WA. 2001. Voornaam is not (really) a homophone: Lexical prosody and lexical access in Dutch. Language and Speech. 44:171-195.

Dahan D. 1996. The role of rhythmic groups in the segmentation of continuous French speech. Proceedings of the Fourth International Conference of Speech and Language Processing; 1996 Oct 3-6; Philadelphia. p. 1185-1188.

Dattorro J. 1987. Using digital signal processor chips in a stereo audio time compressor/expander. Journal of the Audio Engineering Society. 35:1062.

Deacon D, Hewitt S, Yang C, Nagata M. 2000. Event-related potential indices of semantic priming using masked and unmasked words: evidence that the N400 does not reflect a post-lexical process. Brain Res Cogn Brain Res. 9:137-146.

Delattre P. 1966. Studies in French and comparative phonetics. The Hague (The Netherlands): Mouton.

DeLong KA, Urbach TP, Kutas M. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nat Neurosci. 8:1117-1121.

Di Cristo A. 1999. Le cadre accentuel du français contemporain: essai de modélisation (1ère partie). Langues. 2:184-205.

Donchin E. 1981. Surprise!. . .Surprise? Psychophysiology. 18:493-513.

Donchin E, Coles MGH. 1988. Is the P300 component a manifestation of context-updating? Behav Brain Sci. 11:355-372.

Eckstein K, Friederici AD. 2005. Late interaction of syntactic and prosodic processes in sentence comprehension as revealed by ERPs. Brain Res Cogn Brain Res. 25:130–143.

Frazier L, Clifton C, Carlson K. 2004. Don't break, or do: prosodic boundary preferences. Lingua. 114:3–27.

Friedrich CK, Kotz SA, Friederici AD, Alter K. 2004. Pitch modulates lexical identification in spoken word recognition: ERP and behavioral evidence. Brain Res Cogn Brain Res. 20: 300–308.

Heil M, Rolke B, Pecchinenda A. 2004. Automatic semantic activation is no myth: N400 semantic priming effects in the letter search task in the absence of RT effects. Psychol Sci. 15:852–885.

Hirst DJ, Di Cristo A. 1998. A survey of intonation systems. In: Hirst DJ, Di Cristo A, editors. Intonation systems: a survey of twenty languages. Cambridge (UK): Cambridge University Press. p. 1–44.

Kaan E, Harris A, Gibson E, Holcomb PJ. 2000. The P600 as an index of syntactic integration difficulty. Lang Cogn Process. 15:159–201.

Kiefer M, Spitzer M. 2000. Time course of conscious and unconscious semantic brain activation. Neuroreport. 11:2401–2407.

Kutas M, Federmeier KD. 2000. Electrophysiology reveals semantic memory use in language comprehension. Trends Cogn Sci. 4:463–470.

Kutas M, Hillyard SA. 1980. Reading senseless sentences: brain potentials reflect semantic incongruity. Science. 207:203–205.

Kutas M, McCarthy G, Donchin E. 1977. Augmenting mental chronometry: the P300 as a measure of stimulus evaluation time. Science. 197:792–795.

Laroche J. 1993. Autocorrelation method for high quality time/pitch scaling. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New York: New Paltz. Conference date: 17-20 Oct 1993.

Magne C, Astésano C, Lacheret-Dujour A, Morel M, Besson M. 2005. On-line processing of "pop-out" words in spoken French dialogues. J Cogn Neurosci. 17:740–756.

Magne C, Schön D, Besson M. 2006. Musician children detect pitch violations in both music and language better than non-musician children: behavioral and electrophysiological approaches. J Cogn Neurosci. 18:199–211.

Marslen-Wilson WD, Tyler LK, Warren P, Grenier P, Lee CS. 1992. Prosodic effects in minimal attachment. Q J Exp Psychol. 45A:73–87.

Marslen-Wilson WD, Welsh A. 1978. Processing interactions during word recognition in continuous speech. Cogn Psychol. 10:29–63.

McCarthy G, Donchin E. 1979. A metric for thought: a comparison of P300 latency and reaction times. Science. 211:77–80.

McClelland JL, Elman JL. 1986. The TRACE model of speech perception. Cogn Psychol. 18:1–86.

McQueen JM, Norris D, Cutler A. 1994. Competition in spoken word recognition: spotting words in other words. J Exp Psychol Learn Mem Cogn. 20:621–638.

Meyer M, Alter K, Friederici AD. 2003. Towards the cerebral substrates of sentence-level syntactic and prosodic processing. J Neurolinguistics. 16:277–300.

Meyer M, Alter K, Friederici AD, Lohmann G, von Cramon Y. 2002. FMRI reveals brain regions mediating slow prosodic modulations in spoken sentences. Hum Brain Mapp. 17:73–88.

Meyer M, Steinhauer K, Alter K, Friederici AD, von Cramon DY. 2004. Brain activity varies with modulation of dynamic pitch variance in sentence melody. Brain Lang. 89:277–289.

Moreno S, Besson M. 2006. Musical training and language-related brain electrical activity in children. Psychophysiology. 43:287–291.

Most RB, Saltz E. 1979. Information structure in sentences: new information. Lang Speech. 22:89–95.

New B, Pallier C, Ferrand L, Matos R. 2001. Une base de données lexicales du français contemporain sur Internet: LEXIQUE. Année Psychologique. 101:447–462.

Pallone G, Boussard P, Daudet L, Guillemain P, Kronland-Martinet R. 1999. A wavelet based method for audio-video synchronization in broadcasting applications. In: Proceedings of the Conference on Digital Audio Effects (DAFx-99); 1999 Dec 9–11; Trondheim, Norway. p. 59–62.

Pannekamp A, Toepel U, Alter K, Hahne A, Friederici AD. 2005. Prosody-driven sentence processing: an ERP study. J Cogn Neurosci. 17:407–421.

Perrin F, Garcia-Larrea L. 2003. Modulation of the N400 potential during auditory phonological/semantic interaction. Brain Res Cogn Brain Res. 17:36–47.

Pynte J, Prieur B. 1996. Prosodic breaks and attachment decisions in sentence parsing. Lang Cogn Process. 11:165–191.

Ragot R, Renault B. 1981. P300 as a function of S-R compatibility and motor programming. Biol Psychol. 13:289–294.

Salverda AP, Dahan D, McQueen S. 2003. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition. 90:51–89.

Schirmer A, Kotz SA, Friederici AD. 2002. Sex differentiates the role of emotional prosody during word processing. Brain Res Cogn Brain Res. 14:228–233.

Schön D, Magne C, Besson M. 2004. The music of speech: electrophysiological study of pitch perception in language and music. Psychophysiology. 41:341–349.

Soto-Faraco S, Sebastian-Galles N, Cutler A. 2001. Segmental and suprasegmental mismatch in lexical access. J Mem Lang. 45:412–432.

Steinhauer K, Alter K, Friederici AD. 1999. Brain potentials indicate immediate use of prosodic cues in natural speech processing. Nat Neurosci. 2:191–196.

Strelnikov KN, Vorobyev VA, Chernigovskaya TV, Medvedev SV. 2006. Prosodic clues to syntactic processing—a PET and ERP study. Neuroimage. 29:1127–1134.

van Berkum JJ, Zwitserlood P, Hagoort P, Brown CM. 2003. When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. Brain Res Cogn Brain Res. 17:701–718.

van den Brink D, Brown CM, Hagoort P. 2001. Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. J Cogn Neurosci. 13:967–985.

van den Brink D, Brown CM, Hagoort. 2006. The cascaded nature of lexical selection and integration in auditory sentence processing. J Exp Psychol Learn Mem Cogn. 32:364–372.

van den Brink D, Hagoort P. 2004. The influence of semantic and syntactic context constraints on lexical selection and integration in spoken-word comprehension as revealed by ERPs. J Cogn Neurosci. 16:1068–1084.

Van Petten C, Coulson S, Rubin S, Plante E, Parks M. 1999. Time course of word identification and semantic integration in spoken language. J Exp Psychol Learn Mem Cogn. 25:394–417.

Warren P. 1996. Prosody and parsing: an introduction. Lang Cogn Process. 11:1–16.

Wenk BJ, Wioland F. 1982. Is French really syllabletimed? J Phonetics. 10:193–216.

Mitsuko Aramaki, Richard Kronland-Martinet, Thierry Voinier, and Sølvi Ystad
Laboratoire de Mécanique et d'Acoustique
Centre National de la Recherche Scientifique
(CNRS)
31, Chemin Joseph Aiguier
13402 Marseille Cedex 20 France
{aramaki,kronland,voinier,ystad}@lma.cnrs-mrs.fr

A Percussive Sound Synthesizer Based on Physical and Perceptual Attributes

Synthesis of impact sounds is far from a trivial task owing to the high density of modes generally contained in such signals. Several authors have addressed this problem and proposed different approaches to model such sounds. The majority of these models are based on the physics of vibrating structures, as with for instance modal synthesis (Adrien 1991; Pai et al. 2001; van den Doel, Kry, and Pai 2001; Cook 2002; Rocchesso, Bresin, and Fernström 2003). Nevertheless, modal synthesis is not always suitable for complex sounds, such as those with a high density of mixed modes. Other approaches have also been proposed using algorithmic techniques based on digital signal processing. Cook (2002), for example, proposed a granular-synthesis approach based on a wavelet decomposition of sounds.

The sound-synthesis model proposed in this article takes into account both physical and perceptual aspects related to sounds. Many subjective tests have shown the existence of perceptual clues allowing the source of the impact sound (its material, size, etc.) to be identified merely by listening (Klatzky, Pai, and Krotkov 2000; Tucker and Brown 2002). Moreover, these tests have brought to the fore some correlations between physical attributes (the nature of the material and dimensions of the structure) and perceptual attributes (perceived material and perceived dimensions). Hence, it has been shown that the perception of the material mainly correlates with the damping coefficient of the spectral components contained in the sound. This damping is frequency-dependent, and high-frequency modes are generally more heavily damped than low-frequency modes. Actually, the dissipation of vibrating energy owing to the coupling between the structure and the air increases with frequency (see, for example, Caracciolo and Valette 1995).

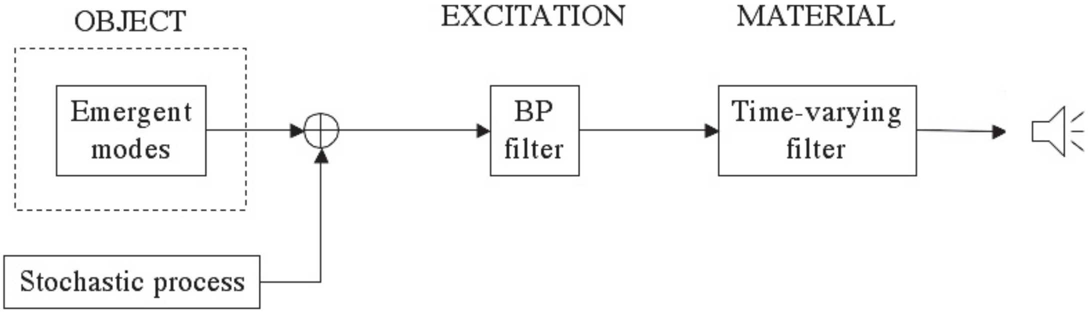
To take into account this fundamental sound behavior from a synthesis point of view, a time-varying filtering technique has been chosen. It is well known that the size and shape of an object's attributes are mainly perceived by the pitch of the generated sound and its spectral richness. The perception of the pitch primarily correlates with the vibrating modes (Carello, Anderson, and Kunkler-Peck 1998). For complex structures, the modal density generally increases with the frequency, so that high frequency modes overlap and become indiscernible. This phenomenon is well known and is described for example in previous works on room acoustics (Kuttruff 1991).

Under such a condition, the human ear determines the pitch of the sound from emergent spectral components with consistent frequency ratios. When a complex percussive sound contains several harmonic or inharmonic series (i.e., spectral components that are not exact multiples of the fundamental frequency), different pitches can generally be heard. The dominant pitch then mainly depends on the frequencies and the amplitudes of the spectral components belonging to a so-called dominant frequency region (Terhardt, Stoll, and Seewann 1982) in which the ear is pitch sensitive. (We will discuss this further in the Tuning section of this article.) With all these aspects in mind, and wishing to propose an easy and intuitive control of the model, we have divided it into three parts represented by an excitation element, a material element, and an object element.

The large number of parameters available through such a model necessitates a control strategy. This strategy (generally called a mapping) is of great importance for the expressive capabilities of the instrument, and it inevitably influences the way it can be used in a musical context (Gobin et al. 2004).

Computer Music Journal, 30:2, pp. 34–43, Summer 2006

Figure 1. Impact sound synthesis model. This model is divided in three parts representing the object, the excitation, and the material contributions.



In this article, we mention some examples of possible strategies, like an original tuning approach based on the theory of harmony. This approach makes it possible to construct complex sounds like musical chords, in which the root of the chord, its type (major or minor), and its inversions can be chosen by the performer. However, owing to the strong influence between mapping and composition, the choice of the strategy should, as far as possible, be available to the composer.

Theoretical Synthesis Model

The synthesis model we propose is shown in Figure 1. It is an extension of that proposed by Smith and Van Duyne (1995a, 1995b), developed to simulate the soundboard's influence on piano tones. This model is based on a time-varying subtractive synthesis process that acts on a noisy input signal. This sound-synthesis model reproduces two main contributions characterizing the perceived material (determined by the damping factors) and the perceived dimensions of the impacted object (determined by pitch and modal density). We decided to model these two contributions separately, even if they cannot be totally disconnected from a physical point of view. Actually, we believe this separation yields an easier and more intuitive control of the sounds.

Another important aspect of the model is its ability to resynthesize natural sounds, meaning that one can also reproduce a given impact sound that is perceptually identical to the original. Nevertheless, this aspect is not described here, and we refer the reader to a more theoretical article (Aramaki and Kronland-Martinet in press). In what follows, we give a more precise description of the three main elements contained in the model.

Material Element

From the literature, it is well known that damping is frequency-dependent, implying that high-frequency modes generally are more heavily damped than low-frequency modes (Caracciolo and Valette 1995). This is important from a perceptual point of view, because damping is a characteristic of the object's material and allows us to distinguish, for example, wood from steel. The damping of the modes is here simulated by a digital infinite-impulse-response (IIR) filter structure in which coefficients vary with time (here called a time-varying filter). Nevertheless, from a theoretical point of view, it is assumed that this variation is small enough for the filter to be considered stationary in a small time interval (Mourjopoulos, Kyriakis-Bitzaros, and Goutis 1990).

The filter used for the model generally is a low pass with a gain and cutoff frequency that decrease with time. In this way, by simply acting on the damping coefficients, we can reproduce the main perceptual features of an impacted material. In particular, if we strongly damp an initial white noise input, the sound will have wooden characteristics, whereas for the same initial white noise, it will have metallic characteristics when the damping is weak (Sound Examples 1 and 2, available online at www.lma.cnrs-mrs.fr/~kronland/CMJ/sounds.html). At this stage, this "material" model is adequate to reproduce perceptual effects reflecting the main characteristics of the impacted materials, even though the technique does not simulate modes.

Object Element

To provide a subjective notion of the size and shape of the sounding object, a few spectral components are added to the initial white noise. If, for example, we add one or a few low-frequency components, the sound will evoke a sense that the impacted structure is relatively big. Conversely, if we add high-frequency components, the sound will evoke a sense that the impacted structure is relatively small (Sound Examples 3 and 4, www.lma.cnrs-mrs.fr/~kronland/CMJ/sounds.html). From a physical point of view, these spectral components mainly correspond to the eigenmodes of the structures. These modes can be deduced for simple cases from the movement equation and can be generated simply by adding a sum of sinusoids to the white-noise input signal.

Nevertheless, the approach suffers from a lack of correlation between the stochastic part of the sound and the deterministic part, making the sounds unrealistic. To overcome this drawback, we generated the deterministic part from narrow bands of the initial white noise, improving the correlation between the two parts (Sound Examples 5 and 6, www.lma.cnrs-mrs.fr/~kronland/CMJ/sounds.html). Note that another method to generate resonances based on physical modeling, namely the banded digital waveguide approach (Essl et al. 2004), has been proposed in Aramaki and Kronland-Martinet (in press). Even though this method gives very satisfactory sounds, we have not used it in real-time applications, because it can lead to instability problems and increase the calculation time. By perceptually reproducing the most pertinent spectral components related to these eigenmodes, we can simulate sounds that evoke various structures like strings, plates, bells, etc. (Sound Examples 7–11, www.lma.cnrs-mrs.fr/~kronland/CMJ/sounds.html).

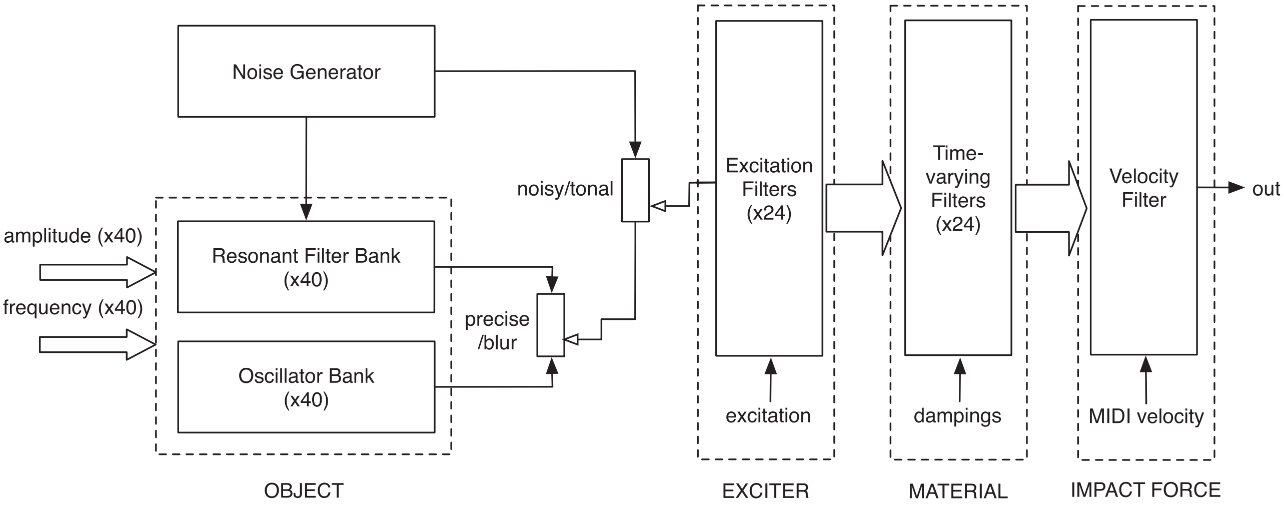
Excitation Element

Finally, to model the excitation, a band-pass filter is used to control the bandwidth of the generated spectrum. From a physical point of view, the response of this filter is strongly related to the strength of the impact, that is, the bandwidth increases as a function of the impact velocity. We can also add a time envelope that controls the attack time of the sound, thereby characterizing the collision between the exciter and the object. This possibility has been added in the real-time implementation of the model. (The slower the attack time, the smoother the excitation, as illustrated in Sound Example 12, www.lma.cnrs-mrs.fr/~kronland/CMJ/sounds.html.)

Tuning

In this section, we discuss the problem of tuning the pitch of the impact sounds. Even though we aim at designing an intuitive tool for musicians rather than a complete impact-sound tuning system, pitch tuning is not a trivial task. Actually, complex sounds often evoke several spectral pitches, because our hearing system tends to associate spectral components having consistent frequency ratios. Moreover, the perceived pitch of a series of spectral components, either harmonic or inharmonic, is not necessarily given by the frequency of the first component of the series. As Terhardt, Stoll, and Seewann (1982) explain, complex tones elicit both spectral and virtual pitches. Spectral pitches correspond to the frequency of spectral peaks contained in the sound spectrum, whereas virtual pitches are deduced by the auditory system from the upper partials in the Fourier spectrum, leading to pitches that may not correspond to any peak contained in the sound spectrum. A well-known example is the auditory generation of the missing fundamental of a harmonic series of pure tones. In addition, owing to the presence of a dominant frequency region situated around 700 Hz in which the ear is particularly pitch-sensitive, the perceived pitch depends on both the frequencies and the amplitudes of the spectral components. Hence, the pitch of complex tones with low fundamental frequencies (under 500 Hz)

Figure 2. Real-time implementation of the three main elements of the theoretical model (related to the object, material, and excitation). A low-pass filter is added to take into account the impact force on the drum interface (controlled by MIDI velocity).



depends on higher partials, while the pitch of tones with high fundamental frequencies is rather determined by the fundamental frequency, because it lies in the dominant region.

As a consequence, when a complex tone contains inharmonic partials, the perceived pitch is determined by the frequencies in the dominant region and might differ from the fundamental frequency. If, for example, a complex inharmonic tone has partials of 204, 408, 612, 800, 1,000, and 1,200 Hz, all with similar amplitudes, the first three partials yield a perceived pitch equal to the fundamental frequency (204 Hz). Nevertheless, the six partials together give a pitch of 200 Hz, because the higher partials determine the pitch, given that they lie in the dominant region (Terhardt, Stoll, and Seewann 1982).

Another aspect that can modify pitch perception is the masking effect between partials, because mental reconstruction of the fundamental frequency of a residue tone might be difficult or impossible if partials in the dominant region are masked by noise or other frequency components. Finally, in most musical situations, tones in context are less ambiguous, because the context normally suggests the pitch register in which the tone is most likely to be heard (Parncutt 1989). This might be of importance in future studies. For the time being, we only focus on the tuning of isolated complex tones.

We shall now see how the real-time model is implemented and adapted to give the user access to these parameters. As this synthesis model simulates percussive sounds, a drum interface is a natural choice for piloting the model.

Real-Time Implementation

Our real-time implementation using Max/MSP is based on the structure of the theoretical synthesis model: the "object" element, devoted to the simulation of the emergent modes; the "material" element, simulating the damping of the sounds; and the "excitation" element (Figure 2). A low-pass filter is added to take into account the impact force on the drum interface (controlled by MIDI velocity).

The input signal of the model consists of a stochastic contribution (limited here to a Gaussian noise generator) providing the broadband spectrum and a tonal contribution simulating the emergent modes. As mentioned earlier, the modes can be simulated by a sum of sinusoids, but the lack of correlation between the stochastic and the deterministic parts makes the sound unrealistic. The spectral peaks are therefore obtained by combining a sum of sinusoids (40 oscillator banks) and a narrow-band filtered white noise (40 resonant filter banks), enabling the creation of more or less "fuzzy" pitches. Indeed, fuzzy pitches are useful for adding reverberative effects to the sound when the stochastic part is weak.

The material element simulating the damping controls the evolution of the spectrum through 24 frequency bands, corresponding to the critical bands of hearing, known as the Bark bands (Zwicker and Fastl 1990). This configuration allows the reproduction of the frequency dependence of the damping, where the damping coefficients are taken as constant in each Bark band. Consequently, the damping is simulated by a time-varying gain for each Bark band.

The excitation part is reproduced by two contributions: the "exciter" element and the "impact force" element (Figure 2). The "exciter" element controls the spectral repartition of the initial energy given to the system and conveys the mechanical characteristics of the excitation element. For common cases, a band-pass filter generally is sufficient. Hence, this contribution is simulated by a static gain adjustment in each Bark band. In comparison to the theoretical model, a supplementary filter ("velocity filter") is added to take into account the player's gesture (MIDI velocity input) and is composed of a one-pole low-pass filter.

Control of the Synthesis Model

As seen in the previous sections, the synthesis model contains a large number of elements that make it possible to control different aspects of the sounds. Although the model has been divided into three parts for more intuitive control, its complexity necessitates the development of strategies to control the parameters. In this section, we first provide an overview of the basic control parameters of the model, and then we discuss how mapping strategies can be developed for musical purposes. Figure 3 shows the actual user interface and reveals the possibilities of control of the synthesis model.

Control of the Input Signal Parameters

As previously mentioned, the input signal is composed of tonal and noisy components. The user can define the relative gain of each contribution (the "tonal/noisy" slider in Figure 3). Concerning the tonal contribution, which consists of the sinusoids and narrow-band filtered white noise (the "object" element in Figure 2), the user can also define the ratio between these two parts (the "precise/blur" slider in Figure 3).

The object element contains 80 parameters (40 frequency and 40 amplitude values) that are associated with the control of the oscillator banks and the resonant filter banks. To minimize the high number of frequency parameters, a proposed tuning preset ("Chord Generator" in Figure 3) based on standard Western tonal definitions is constructed. Players can here choose whether they wish to construct the complex sound with a single pitch, with several pitches forming a specific four-note chord, or with arbitrary combinations of pitches.

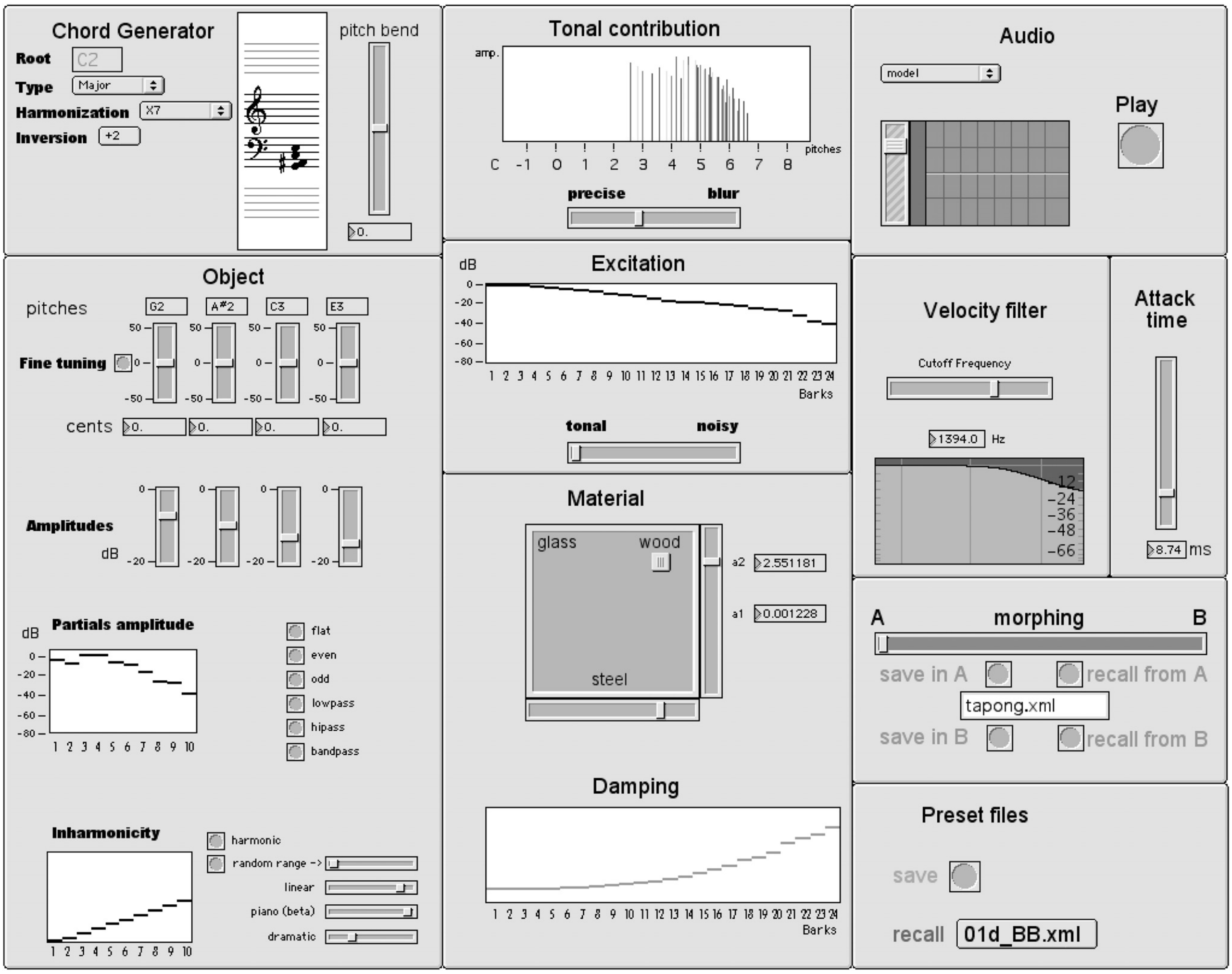
When the unison case is chosen, the four notes generated by the chord generator have the same pitch values. When the chord case is chosen, four different pitches are generated. In this case, the player selects the root of the chord, the type (major or minor), the harmonization (4th, 7th, 9th, diminished, etc.) and the inversion (four possible). (This is illustrated in Sound Example 13, www.lma.cnrs-mrs.fr/~kronland/CMJ/sounds.html.)

The inversions indicate the lowest note of the chord so that, for example, the first inversion ("inversion +1") corresponds to a chord built on the second note of the main chord, and the second inversion ("inversion +2") corresponds to a chord built on the third note of the main chord. Negative inversions are also possible. "Inversion –1" and "inversion –2" indicate that the lowest note of a four-note chord is taken as the fourth and third notes, respectively, the non-inverted chord. When the main chord contains four notes (e.g., "C7"), then "inversion +2" and "inversion –2" are identical (differing by one octave). This would not have been the case if the chords had contained more than four notes (e.g., "C4,7," "D7,9", etc.). As an example, Figure 4 shows some possible inversions of "Cminor7."

It is well known, for example, that a triad in root position played on a piano is more "stable" than its inversions, and that the relative stability of two chords is determined by the relative proximity to

Figure 3. User interface for control of the synthesis model.

Figure 4. Example of a C-minor seventh chord with some possible inversions.



Chord Generator

Root C2

Type Major

Harmonization X7

Inversion +2

pitch bend 0.

Object

pitches G2 A#2 C3 E3

cents 0. 0. 0. 0.

Audio: model

Material: a2 2.551181, a1 0.001228

Velocity filter: Cutoff Frequency 1394.0 Hz

Attack time 8.74 ms

morphing: tapong.xml

Preset files: recall 01d_BB.xml

Figure 3

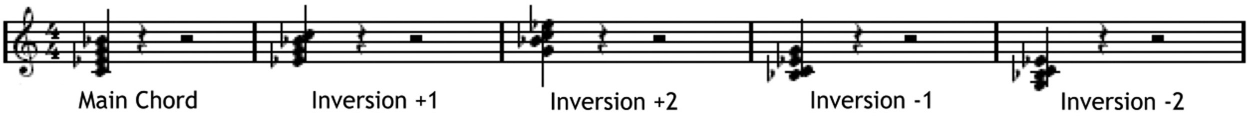
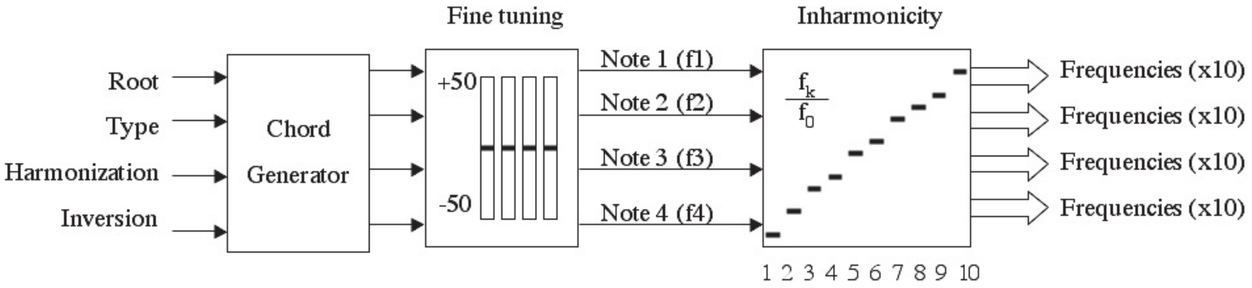


Figure 4

Figure 5. Control of the object element (i.e., of the 40 frequency values) with the chord generator, fine tuning, and inharmonicity.



the local tonic of their roots in the circle of fifths (Lerdahl and Jackendoff 1977). We will later investigate whether this is also the case for our impulsive sounds.

From a given chord preset, pitch deviations can also be created by moving sliders ("Fine tuning" in Figure 3). For each note, a set of ten oscillator banks and ten resonant filter banks are associated, meaning that a spectrum composed of ten partials is generated. The user can control the amplitude values globally by acting on each note's gain ("Amplitudes" in Figure 3) or more precisely by acting independently on the amplitudes of the ten spectral components associated with each note ("Partials amplitude" in Figure 3).

The player can also alter the relationship between the partials of each note by the control of inharmonicity ("Inharmonicity" in Figure 3). The inharmonicity relationship can either be chosen individually by defining the frequency ratio f_k/f_0 of each partial in the series as a function of the fundamental frequency, or by adjusting parameters of different presets. We have chosen the following presets for this purpose: harmonic ($f_k = k \times f_0$), linear ($f_k = a \times k \times f_0$) or piano-like ($f_k = k \times f_0 \times \sqrt{1 + \beta k^2}$, after Fletcher 1964 and Valette and Cuesta 1993). In each relationship, f_0 is the fundamental frequency, and $k \in \mathrm{N}+$, $a \in \mathbf{R}^*+$, and $\beta \in \mathbf{R}+$ are the control parameters. For example, when the signal is harmonic, the spectral components are integer multiples of the fundamental frequency, and the inharmonicity curve is given by a straight line, where the frequency of the kth component equals k times the frequency of the fundamental component. Figure 5 provides a more detailed view of the control of the "object" element (i.e., the 40 frequency values) using the chord generator, fine-tuning, and inharmonicity settings.

A schematic representation of the tonal contribution is given on the top of the user interface ("Tonal contribution" in Figure 3). The degree of inharmonicity of the spectral components together with their amplitudes makes it possible to alter the perceived pitch as a function of the dominant frequency region, as explained previously. This tool represents both an interesting musical interface to tune the synthesis model in a musical context and an important research tool to study the relationship between pitch perception and spectral components of complex sounds.

Control of the Material Element

The material part of the model is controlled by the damping parameters, with access to 24 values corresponding to the frequency bands on which the user can act independently ("Damping" in Figure 3). In addition, we have chosen to parameterize the set of damping values by an exponential function, defining a damping law $\alpha(\omega)$ that can be written

$$\alpha(\omega) = e^{a_1\omega + a_2} \qquad (1)$$

Hence, we reduce the damping control to only two parameters, a_1 and a_2. This damping law is a function of the frequency and can be directly estimated from physical considerations or from the analysis of natural sounds. Thus, a bi-dimensional space defined by $\{a_1; a_2\}$ in which the player can move a cursor is proposed ("Material" in Figure 3). As the damping values are strongly characteristic of the nature of the perceived material, this space can be considered as a "material space" where specific zones can be representative of different materials. Analysis of natural sounds of different materials (in particular wood, glass, and steel) allowed us to calibrate this bi-dimensional space and roughly determine the domains of each material, as shown in Figure 3. Hence, the player can move from one material to another, for example, from a wooden to a metallic sound (Sound Examples 14 and 15, www.lma.cnrs-mrs.fr/~kronland/CMJ/sounds.html).

Control of the Excitation Element

The excitation filter is controlled with 24 gains that can be modified graphically ("Excitation" in Figure 3). These gains act on 24 Bark-scale band-pass filters. We can also take into account the excitation point, which from a physical point of view causes envelope modulations in the corresponding spectrum. The player can change the attack time to vary the perceived excitation ("Attack time" in Figure 3). For instance, a slow attack simulates the perception of a rubbed object. More generally, one could propose an excitation space where the type of excitation could be chosen (e.g. plucking, striking, rubbing). Finally, the player's gesture is taken into account by a low-pass filter ("velocity filter" in Figure 3) which cutoff frequency depends on the force sensed by the trigger control interface. In this way, we can imitate the well-known non-linear effect that leads to an increase in the spectral width as a function of the force of the impact.

Other Mapping Strategies

We now propose possible mapping strategies for more intuitive control of the real-time synthesis model. These examples are intended to give a hint of many possible useful strategies. In addition to the material space proposed in the previous section, it would be of interest to define a space related to the size and the shape of the object. As already seen, the parameters related to such an object space would mainly be related to the pitch and the inharmonicity of the input signal. Actually, small objects are generally envisioned in the listener when high pitches are perceived, whereas big objects are envisioned when low pitches are perceived. Furthermore, a one-dimensional structure (e.g., a string) is perceived when a unison preset is chosen for the chord generator, whereas a multi-dimensional structure (e.g., a plate or a bell) is perceived when several pitches are chosen.

In this way, according to our mental representations of sounds, a more intuitive control can be proposed based on verbal input parameters such as "string-like" or "plate-like" that could be linked to a geometry dimension, and parameters such as "big" and "small" that could be linked to a size dimension in the object space. By proposing chord presets when multiple pitches are chosen, we believe the musician will have access to an interesting tool to control the combination of spectral components. The spectral content of complex sounds is often very dense and hence difficult to control intuitively. Being able to construct spectra from a musical approach (i.e., basic chord theory) attracts musicians and facilitates the complex task of structuring rich spectra.

Another possibility is to act directly on the sound quality, namely, on the timbre itself. In this case, we focus on the perceptual effects of the sound without taking into account physical aspects of the source. Thus, we can act directly on the timbre descriptors, such as the attack time, the spectral centroid, and the spectral flux (McAdams et al. 1995). The attack time is a measure for how quickly the energy envelope attains its maximum value, the spectral centroid is a measure of the spectral center of gravity and is directly related to the brightness of the sound, and the spectral flux is a measure for the variation of the spectral envelope over the duration of the note. Aspects of timbre can then be controlled by acting on a timbre space with two or three dimensions represented by different timbre descriptors. Such a control is not available currently, but it will be available in future versions.

In addition to the different control spaces linked to the material, size, shape, and timbre of the sounds, we have given the user the ability to morph between two different sounds ("morphing" in Figure 3). For this purpose, interpolations (linear or logarithmic) between the parameters of the two reference sounds are employed. This control possibility gives access to the creation of hybrid sounds, which for example makes it possible to simulate continuous transitions between different materials (Sound Examples 14 and 15, www.lma.cnrs-mrs.fr/~kronland/CMJ/sounds.html) or between different structures.

Conclusion

We have presented an efficient, hybrid synthesis technique for percussive sounds. This sound-synthesis model reproduces two main contributions characterizing the perceived material and the perceptual dimensions of the structure. A real-time implementation of the model showed its accuracy, allowing the generation of a wide variety of impact sounds. The system has been used in a musical context with a drum-like MIDI interface. As the drum interface itself offers limited controls, we have employed additional controllers (sliders, pedals, etc.) to act on different parameters, such as pitch and damping coefficients. To avoid additional controllers, the system has also been piloted by a MIDI keyboard that allows a direct control of the pitch and velocity and offers other control possibilities (e.g., aftertouch and pitch bend).

The adjustment of the model's parameters, however, is often difficult and necessitates the development of a mapping strategy. We have presented some mapping strategies, such as a morphing control, a material space, and an original approach to tune the complex sounds based on the standard-practice Western theory of harmony. This approach also makes the synthesis model an interesting research tool to investigate pitch perception of complex sounds. However, the choice of these strategies is left open to the composer, because it strongly influences the music that is to be written for this instrument.

This study is a first step toward a better understanding of the nature of percussive sounds, and especially toward a description of their most pertinent parameters from a perceptual and cognitive point of view. For this purpose, a larger, interdisciplinary project related to the semiotics of sounds associating sound modeling and neurosciences has been initiated.

In addition to the generation of synthesis sounds, we are also attempting to construct an analysis-synthesis platform. Actually, analysis-synthesis techniques that allow a given impact sound to be resynthesized have already been designed (Aramaki and Kronland-Martinet in press). The association of nonlinear analysis-synthesis processes can allow the resynthesis of sounds generated by source-resonance systems, while perceptual and cognitive approaches will be proposed to study the influence of each synthesis parameter on listeners.

References

Adrien, J. M. 1991. "The Missing Link: Modal Synthesis." In G. De Poli, A. Piccialli, and C. Roads, eds. *Representations of Musical Signals.* Cambridge, Massachusetts: MIT Press, pp. 269–297.

Aramaki, M., and R. Kronland-Martinet. In press. "Analysis-Synthesis of Impact Sounds by Real-Time Dynamic Filtering." *IEEE Transactions on Speech and Audio Processing* 14(2).

Caracciolo, A., and C. Valette. 1995. "Damping Mechanisms Governing Plate Vibration." *Acta Acustica* 3:393–404.

Carello, C., K. L. Anderson, and A. J. Kunkler-Peck. 1998. "Perception of Object Length by Sound." *Psychological Science* 9(3):211–214.

Cook, P. R. 2002. *Real Sound Synthesis for Interactive Applications.* Natick, Massachusetts: AK Peters.

van den Doel, K., P. G. Kry, and D. K. Pai. 2001. "FoleyAutomatic: Physically-Based Sound Effects for Interactive Simulation and Animation." *Proceedings of SIGGRAPH 2001.* New York: Association for Computing Machinery, pp. 537–544.

Essl, G., et al. 2004. "Theory of Banded Waveguides." *Computer Music Journal* 28(1):37–50.

Fletcher, H. 1964. "Normal Vibration Frequencies of a Stiff Piano String." *Journal of the Acoustical Society of America* 36(1):203–209.

Gobin, P., et al. 2004. "Designing Musical Interfaces with Composition in Mind." *Lecture Notes in Computer Science LNCS 2771.* Vienna: Springer, pp. 225–246.

Klatzky, R. L., D. K. Pai, and E. P. Krotkov. 2000. "Perception of Material from Contact Sounds." *Presence* 9(4): 399–410.

Kuttruff, H. 1991. *Room Acoustics.* New York: Elsevier.

Lerdahl, F., and R. Jackendoff. 1977. "Toward a Formal Theory of Tonal Music." *Journal of Music Theory* 21:110–171.

McAdams, S., et al. 1995. "Perceptual Scaling of Synthesized Musical Timbres: Common Dimensions, Specificities, and Latent Subject Classes." *Psychological Research* 58:177–192.

Mourjopoulos, J. N., E. D. Kyriakis-Bitzaros, and C. E. Goutis. 1990. "Theory and Real-Time Implementation of Time-Varying Digital Audio Filters." *Journal of the Audio Engineering Society* 38(7/8):523–536.

Pai, D. K., et al. 2001. "Scanning Physical Interaction Behavior of 3D Objects." *Proceedings of SIGGRAPH 2001.* New York: Association for Computing Machinery, pp. 87–96.

Parncutt R. 1989. *Harmony: A Psychoacoustical Approach.* Vienna: Springer.

Rocchesso, D., R. Bresin, and M. Fernström. 2003. "Sounding Objects." *IEEE Multimedia* 10(2):42–52.

Smith, J. O., and S. A. van Duyne. 1995a. "Commuted Piano Synthesis." *Proceedings of the 1995 International Computer Music Conference.* San Francisco, California: International Computer Music Association, pp. 335–342.

Smith, J. O., and S. A. van Duyne. 1995b. "Developments for the Commuted Piano." *Proceedings of the 1995 International Computer Music Conference.* San Francisco, California: International Computer Music Association, pp. 319–326.

Terhardt, E., G. Stoll, and M. Seewann. 1982. "Pitch of Complex Signals According to Virtual-Pitch Theory: Tests, Examples, and Predictions." *Journal of the Acoustical Society of America* 71:671–678.

Tucker, S., and G. J. Brown. 2002. "Investigating the Perception of the Size, Shape, and Material of Damped and Free Vibrating Plates." University of Sheffield, Department of Computer Science Technical Report CS-02-10.

Valette, C., and C. Cuesta. 1993. *Mécanique de la corde vibrante.* Lyon, France: Hermès.

Zwicker, E., and H. Fastl. 1990. *Psychoacoustics: Facts and Models.* Vienna: Springer.

Perceptive and cognitive evaluation of a piano synthesis model

Julien Bensa[1], Danièle Dubois[1], Richard Kronland-Martinet[2], and Solvi Ystad[2]

[1] Laboratoire d'Acoustique Musicale, Université Pierre et Marie Curie,
11 rue de Lourmel, Paris, France
{bensa, dubois}@lam.jussieu.fr
[2] Laboratoire de Mécanique et d'Acoustique, équipe S2M,
31 ch. Joseph Aiguier, Marseille, France
{kronland, ystad}@lma.cnrs-mrs.fr

Abstract. The aim of this work is to use subjective evaluations of sounds produced by a piano synthesis model to determine the perceptual influence on phenomena involved in sound production. The specificity of musical sounds is that they are intended for perception and judgments by human beings. It is therefore necessary, in order to evaluate the acoustic qualities of a musical instrument or a sound model, to introduce a research approach which takes into account the evaluation of the sound quality by human beings. As a first approach we synthesize a number of piano sounds. We then evaluate the quality of the perceived acoustic signal by questioning a group of persons. We hereby try to link the model's parameters to semantic descriptors obtained from these persons and to more classical perceptual signal descriptors. This approach should give a better understanding of how the model's parameters are linked to cognitive representations and more generally give new clues to cognitive descriptions of timbre of musical sounds.

1 Introduction

The piano is a complex instrument with a large number of mechanical elements, the majority of which contribute to the sound production. The physical characteristics of these elements together with their interaction influence the timbre of the piano sound. Thus, in order to give a precise description of the behavior of this instrument and effectuate a satisfactory sound synthesis, the totality of the physical phenomena that are part of the sound production ideally should be taken into account. However, due to the complexity of the sound production system, this is not possible. We therefore propose, thanks to sound modelling and synthesis, to determine the most important perceptual phenomena related to the sound production system and to evaluate the importance of each of them. This approach hopefully will give a better understanding of the relation between the physical behavior of the instrument and the perception of its sound quality, and hereby give clues to how the piano model can be simplified without loss of quality. This is crucial for the conception of high-quality synthesis models that are to be run in real-time.

As a contrast to non-intentional noises (or sounds) generated by different sources and generally considered as annoying, musical sounds are produced to be perceived and appreciated by human beings. Thus, in order to evaluate the sound quality, it is necessary to look for "subjective judgments" given by a number of subjects (professional musicians, amateurs or instrument makers) that listen to and compare the sounds. Although classical psychophysical approaches have been used for a long time to elaborate subjective evaluation methods, the analytic and parametric character of the sound samples that generally are implied in the tasks that the subjects are asked to perform, does not seem appropriate to the musical sounds that are studied here. The subjective evaluation used in psychoacoustics is generally based on stimuli which are analytically described within a multidimensional space given by physics. Such an approach mainly investigates low-level processing such as perceptual thresholds rather than high-level processing of complex sounds such as musical samples that cannot be reduced to a set of values identified by two or three physical dimensions. Although each stimulus is given by a unique physical description determined by the parameters of a physical model, there is a multitude of heterogeneous and poorly known principles of the organization of global perceptual judgments. In particular, the perception of timbre of a musical instrument is not only related to sensations linked to the characteristics of the physical signal, but also to criteria associated to interpretation processes and to knowledge achieved in a particular cultural community. We have therefore deliberately chosen to take these "high quality" characteristics into account when modelling the acoustic qualities of piano sounds. This means that global judgments are considered, referring not only to what is perceived in an analytical listening condition, but also to the sensation that the subjects approve as a result of their personal experience, their knowledge and their expertise.

To obtain this subjective evaluation, we used a methodology that has already been validated on visual and olfactory stimuli [1, 2]. This methodology relies on theoretical assumptions regarding cognitive categories and their relations to language [3]. The psycholinguistic analysis of verbal comments that subjects produce as answers to acoustic stimulations can be considered as an access to cognitive representations. We therefore processed a free categorization task: subjects were asked to freely classify the stimuli according to personal criteria, and to comment their final classification.

The free categorization method has a theoretical frame which is adapted to the rather unknown character of cognitive structures of musical objects [4–6]. As a contrast to estimation methods where subjects are asked to adapt their judgment to a pre-defined scale, the tasks of our method make it possible to induce properties of pertinent physical stimuli (not necessarily known) of the cognitive representations. In fact, the subjects are given the freedom to choose their own subjective measure to describe the stimuli. Thus, we are making the hypothesis that for the same class of rather complex sounds (for instance different sounds from the same instrument), subtle timbre variations can be identified from a semantic analysis of the descriptions given by the subjects although they are difficult to distinguish by more classical descriptors (obtained from the conceptualization of physical science). In fact, as Jensen mentions [7], there is a need to study subtle timbre variations within the same instrument since several timbre parameters often are more similar for sounds from different instruments with the same pitch than for sounds from the same instrument with a different pitch. Hence, the comments from the subjective evaluations will make it possible to discover different domains of knowledge involved in the evaluation of sound quality as a function of the subjects' different expertise.

In the first part of the paper we shortly describe the main physical phenomena involved in the piano sound production and the model we use to produce the set of stimuli. We here focus and confront two features of piano sounds: inharmonicity and "phantom" partials. We then show how classical timbre descriptors (centroid and spectral flux) may be used to characterize each stimulus. Finally, we describe the categorization task and give some preliminary results leading to a discussion on how the different cognitive categories are related to the model parameters.

2 The synthesis model and the control parameters

The piano consists of a large number of components, each one having a role that is more or less important to the sound production mechanism. The piano is a struck string instrument. Each note can use either one, two, or three strings with physical properties that differ from one note to another. The resulting timbre depends namely on the interaction between the hammer and the string, the coupling between the strings, and the way in which the sound radiates from the soundboard. Many studies (see i.e. [8, 9]) have described the behavior of the varied elements of the piano and we here refer to those publications for a more precise descriptions of the acoustics of this instrument. Different types of sound synthesis models of the piano simulating phenomena involved in sound production have been proposed [9–13]. Signal models are generally computationally efficient enough to run in real time and can be very accurate in reproducing the sound of an existing piano. However, these types of models fall short when it comes to incorporating the player into the control-instrument loop. Since they make no direct link between the player's actions and the physics of the instrument, important playing conditions have no effect on the produced sound. Physical models on the other hand, have the advantage of simulating the interaction between player and instrument, although this comes at a computational cost (though this cost is becoming less of an issue as computers become increasingly powerful). The parameters of physical models are difficult to accurately estimate from measured signals, and their sound quality often is poorer than for signal models. The quality of the synthesis depends on how accurately the acoustic system is taken into account in the physical model. Though algorithms are becoming more efficient and computer computation ever more powerful, the balance between sound quality and algorithmic complexity is still delicate. Thus, one of the first motivation of this study is to obtain a classification of the different phenomena involved in piano sound production. This classification is of great importance for the perceptual quality of real-time physical models.

Here, we would like to confront the perceptual effect of two phenomena involved in sound production: the string stiffness and the tension modulation, which respectively lead to the inharmonicity of the piano spectrum and the so-called "phantom" partials. Inharmonicity is a well-known characteristics of piano spectra. The "harmonics" of the piano sound are not exact integral multiples of the fundamental frequency. The whole spectrum is stretched and its components are called partials. This inharmonicity contributes to the piano tone and is mainly responsible for its specificity [16]. The stretched tuning of piano strings can be almost entirely attributed to their inherent stiffness [17], which leads to a dispersion of waves during the propagation. For small stiffness, the modal frequencies of the string are [16]:

$$f_n = nf_0\sqrt{1 + Bn^2} \tag{1}$$

where f_n is the modal frequency, f_0 the fundamental frequency, n the partial index and B the inharmonicity factor.

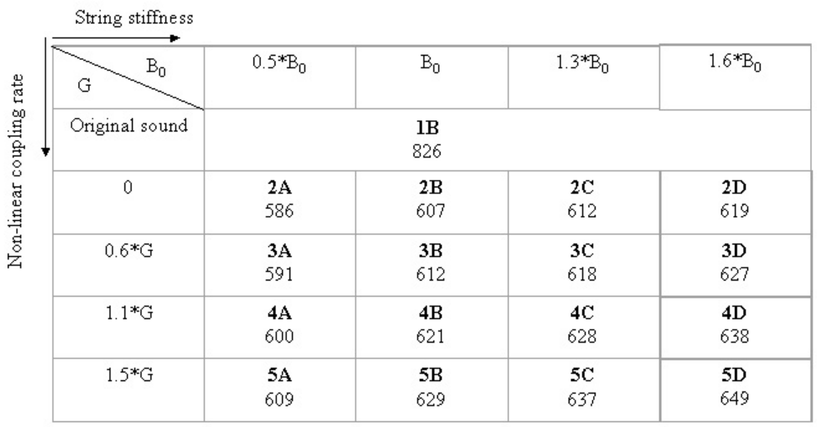
The models commonly used to simulate piano string vibrations take into account the propagation of transverse waves (including stiffness and losses), leading to a spectrum of inharmonic partials. However, a close inspection of Fourier spectra of recorded piano sounds shows that a number of partials cannot be related to the transverse modes of the string and are not foreseen by the linear theory. Moreover, those partials seem to contribute to the piano timbre, especially for low-pitched notes. Studies dealing with this phenomenon make the assumption that the appearance of those partials, also called "phantom" partials [18] (we will use this terminology in this article, even if this is maybe not the most appropriate term) are somewhat related to tension variation in the string. Due to transverse waves, the shape of the string changes during the motion, and the tension is modulated. This modulation introduces a coupling between the transverse and longitudinal modes of the string, giving rise to new partials [19, 20]. Both of the phenomena described previously affect the frequency, amplitude and damping of the spectral components of a piano sound. We here investigate how they contribute to the perception of the piano timbre and how their combination affects the perceptual judgement of the listeners.

The model we use is based on analysis and synthesis techniques described in [13]. For this test, we worked with only one piano note (B1), recorded in an anechoic room. This note corresponds to two strings tuned to a very close pitch and thus, the recorded sound exhibits beating phenomena. Moreover, we localized many "phantom" partials on the corresponding spectrum. We accurately analyze the original sound using technique based on time-frequency representation and parametric methods given in [13, 21]. As a first step, we estimate the modal parameters of each partial of the inharmonic spectrum, i.e. two frequencies, damping coefficients and initial amplitudes. In a second step, we localize the "phantom" partials (using an automatic algorithm we have developed) and estimate the corresponding modal parameters. The synthesis model is made of two different parts: a digital waveguide model [14] modelling the propagation of transverse waves in the string and an additive signal model allowing to introduce "phantom" partials [15]. This kind of signal model exhibits three main advantages for the needs of the listening test we would like to carry out. First, the synthesized sound is perceptually extremely close to the original sound: it is actually important that the subjects could not identify those synthesized sounds as "synthetic" stimuli. Secondly, the signal model offers the possibility of modifying independently the contribution of each phenomena and thus allows to regularly increasing different parameter values. Third it is possible, at least for partials coming from transverse waves, to modify their frequencies with respect to physical laws by taking into account inharmonicity law [16] and attenuation law (the physical descriptions of the "phantom" partials found in the literature are not accurate enough to allow a physical control of their modal parameters).

Using this model, we have synthesized sounds for different inharmonicity factor and different level of "phantom" partials. B is the inharmonicity factor (with $B_0 = 2.4176.10^{-4}$ its value for the original sound) and G is the global gain (G_0 is the value of the original sound) mixing "phantom" partials with "regular" partials. We obtained 17 stimuli as shown on Fig. 1 (labelled 1B, 2A, 2B...). The variation range of the two parameters has been chosen to cover a wide range of perceptual effects, from a sound with very weak inharmonicity and no "phantom" partials (2A) to a sound with exaggerated inharmonicity and a high level of "phantom" partials (5D). 4B is the closest sound to the original sound 1B from the model parameters point of view.

3 Timbre description using "classical" descriptors

Timbre is probably one of the the most well-known and least understood attribute of the quality of a sound. It refers to those aspects of a sound other than pitch, loudness, perceived duration, spatial location and reverberant environment [22] and can be regarded as the feature of an auditory stimulus that allows us to distinguish one source from another when the other five perceptual features are held constant. Thus timbre can be considered as the "tonal color and texture" [23] that allows us to distinguish two different instruments playing the same note. Although a lot of work has been done to describe timbre [24–26], it is not yet a fully understood component of auditory perception. One of the reasons for this is probably the fact that the cognitive processes of timbre classifications depend on the experience of each listener [3]. This is one of the motivations behind our attempt to link timbre to semantic descriptors by means of a piano synthesis model in order to approach a cognitive description of timbre of musical sounds. This is particularly important when studying subtle timbre variations within the same instrument since several timbre parameters often are more similar for sounds from different instruments with the same pitch than for sounds from the same instrument with a different pitch [7]. As a starting point to this investigation we will study some parameters that are traditionally used to describe timbre. Some of the most significant parameters commonly mentioned



String stiffness

Non-linear coupling rate

B_0 \ G	$0.5{*}B_0$	B_0	$1.3{*}B_0$	$1.6{*}B_0$
Original sound		**1B** 826		
0	**2A** 586	**2B** 607	**2C** 612	**2D** 619
0.6*G	**3A** 591	**3B** 612	**3C** 618	**3D** 627
1.1*G	**4A** 600	**4B** 621	**4C** 628	**4D** 638
1.5*G	**5A** 609	**5B** 629	**5C** 637	**5D** 649

Fig. 1. The stimuli as a function of the inharmonicity factor B (with $B_0 = 2.4176.10^{-4}$ the original value) and the global gain G of the spectral "phantom" components ($G = 0$ means no "phantom" components, $G = G_0$ means the same level of "phantom" components as for the original sound). Spectral centroid in Hz for the sixteen synthesized sounds and the original sound.

in the literature are the spectral centroid, the attack time, the spectral flux and the spectral irregularity [25]. We have chosen to calculate two of these timbre descriptors from the 17 synthesized piano sounds that have been made for this particular study, namely the spectral centroid and the spectral flux. Since the attack time of the synthesized sounds used in this study is the same for all the piano tones, and since the irregularity of the spectrum (variations between the amplitude of the spectral components) is constant for the 17 sounds, we have here been considering the spectral centroid (SC) and the spectral flux (SF). The spectral centroid is defined as [24]

$$SC = \frac{\sum_k kA_k}{\sum_k A_k}, \tag{2}$$

(with k the partial index and Ak the spectral partial amplitude) and is often said to be related to the brightness of the sound. The spectral flux is a mean value of the variation of the spectral components as a function of time [25]. This means that it describes the attenuation of the spectral components which is of great importance for the perception of the sound. In this article we have chosen to calculate the spectral flux from successive centroids estimated at different times, meaning that this parameter can be considered as a time varying brightness. As mentioned in Sect. 2, the 17 different sounds are obtained for the same pitch, by varying the inharmonicity B and the phantom components represented by the gain factor G.

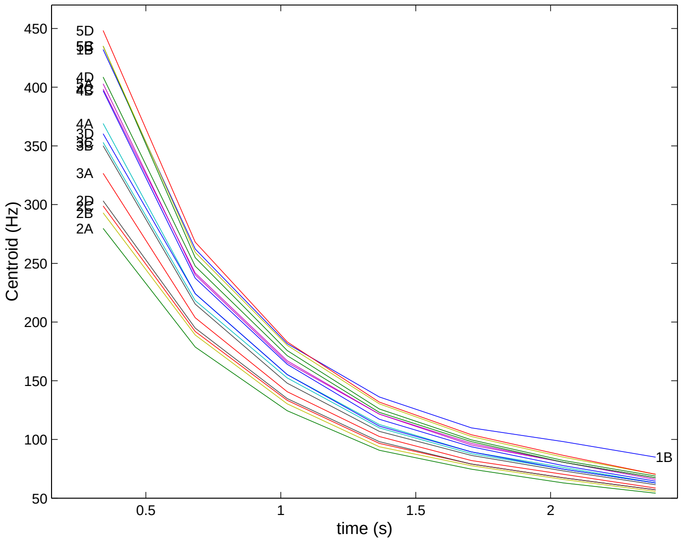
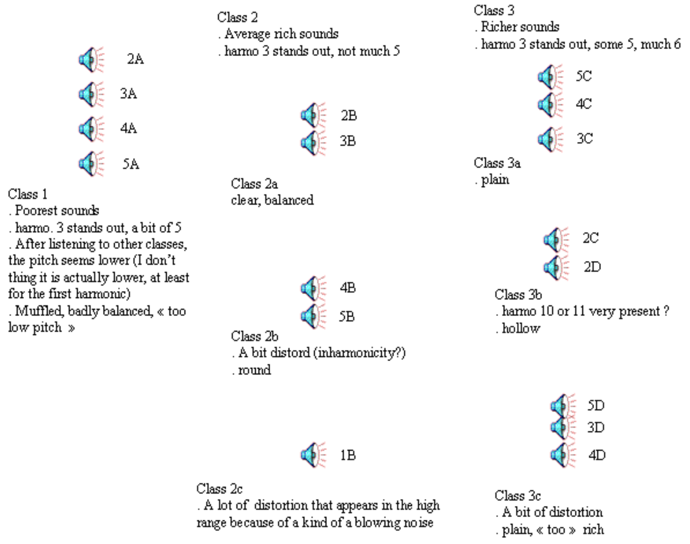


Fig. 2. Spectral flux represented by the spectral centroid as a function of time for the 17 different sounds.

Fig. 1 shows the values of the spectral centroid for the different synthesized sounds (mean value for 6 ms of the sound). As expected the centroid increases for increasing inharmonicity since the spectral components get more and more separated, and it also increases when the gain factor increases, since there are more and more "phantom components" for higher frequencies. We can also notice that stimuli 1B has a higher value of its spectral centroid. Fig. 2 shows the spectral flux represented by the spectral centroid as a function of time. One can observe that during the attack of the sounds (1st second), five different groups of sounds can be distinguished, namely 5D-5C-5B-1B,4D-5A-4C-4B, 4A-3D-3C-3B, 3A, and 2D-2C-2B-2A. These groups are rather coherent with the physical parameters since they more or less correspond to lines in the table (Fig. 1). This seems to indicate that the spectral flux varies with G_0 (gain of the phantom partials), but does not depend on the inharmonicity factor. After one second there is no specific group of sounds to observe and the spectral flux almost becomes the same for all of the synthesized sounds. The spectral flux of the original sound is higher than those of the synthesized sounds in the last part of the sound, but this doesn't seem to be perceptually important since listeners tend to pay more attention to the attack of the sound (as we will see in the next section).

4 Free categorization task: protocol and first results

The subjective evaluation was carried out on subjects using a free categorization task of the stimuli described above. Subjects were asked to freely sort the 17 piano sounds, symbolized by a schematic speaker on a screen (see Fig. 3) into groups and subgroups, and when the task was completed, to write down the reasons of their choice. When clicking on the speaker, the subjects could listen to the sound (through two "real" speakers) and interrupt it at their convenience by clicking again. The subjects could form as many categories as they wanted. One example of such a categorization is given in Fig. 3 below (results for subject SX, a pianist).



Class 1
. Poorest sounds
. harmo. 3 stands out, a bit of 5
. After listening to other classes, the pitch seems lower (I don't thing it is actually lower, at least for the first harmonic)
. Muffled, badly balanced, « too low pitch »

Class 2
. Average rich sounds
. harmo 3 stands out, not much 5

Class 2a
clear, balanced

Class 2b
. A bit distord (inharmonicity?)
. round

Class 2c
. A lot of distortion that appears in the high range because of a kind of a blowing noise

Class 3
. Richer sounds
. harmo 3 stands out, some 5, much 6

Class 3a
. plain

Class 3b
. harmo 10 or 11 very present ?
. hollow

Class 3c
. A bit of distortion
. plain, « too » rich

Fig. 3. Results of the free categorization task given by SX, a pianist.

The experiment is presently running and we here present the results obtained on 21 subjects, 7 pianists, 8 musicians non-pianists and 6 non-musicians. The full experiment will include at least 30 subjects, in order to contrast the different principles of categorization involved in the task and to elicit the characteristics that structure the cognitive specificities of these diverse populations. We therefore present the categories and comments of a pianist and a non-pianist (SX and SY as reported in Figs. 4 and 5) mapped on the matrix of controlled physical parameters involved in the construction of the experimental stimuli. As shown in Fig. 4, the first level of cognitive categorization for the pianist subject SX fits the physical criterion of inharmonicity, and the categorization that can be attributed to "phantom" partials is subordinate at a second level of categorization. However, such a categorical structure is not observed on the non-pianist subject SY (Fig. 5) where, except for the lowest value of the inharmonicity factor, the 3 categories (2B, 2C), (2D, 5B, 5C and 5D), (3C, 3D, 4C and 4D) integrate the two physical dimensions with respect to their values on a single level. These two preliminary results favor the hypothesis that criteria of categorization highly depend on subjects' experience and therefore that different cognitive categories can emerge from the same set of stimuli univocally described in terms of physics.

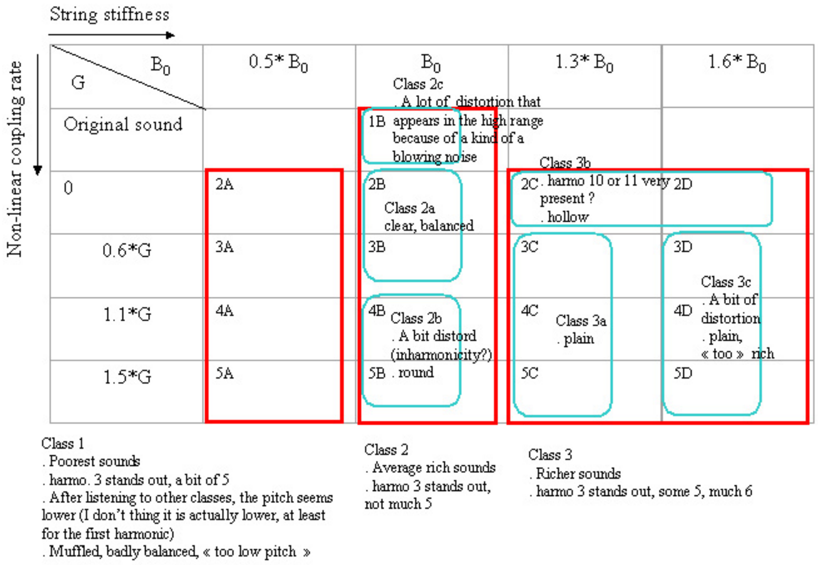
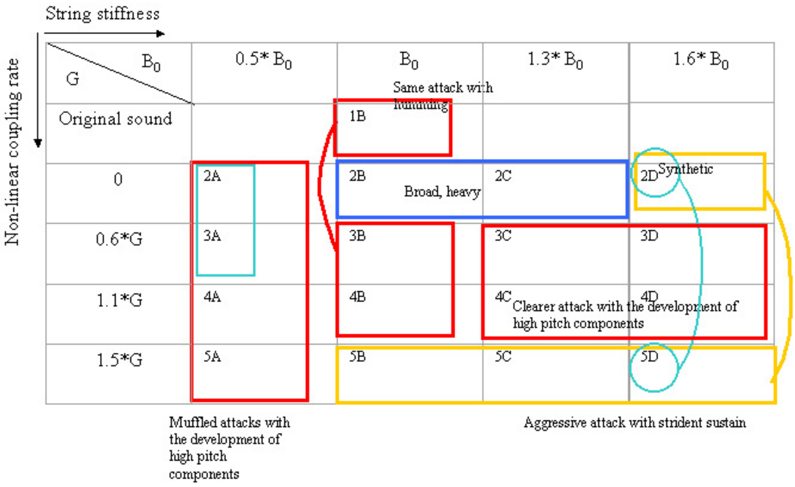

String stiffness
Non-linear coupling rate
G
B_0
$0.5* B_0$
B_0
$1.3* B_0$
$1.6* B_0$
Original sound
0
0.6*G
1.1*G
1.5*G
1B
2A
2B
2C
2D
3A
3B
3C
3D
4A
4B
4C
4D
5A
5B
5C
5D
Class 2c
. A lot of distortion that appears in the high range because of a kind of a blowing noise
Class 2a
clear, balanced
Class 2b
. A bit distord (inharmonicity?)
. round
Class 3b
. harmo 10 or 11 very present ?
. hollow
Class 3a
. plain
Class 3c
. A bit of distortion
. plain,
« too » rich
Class 1
. Poorest sounds
. harmo. 3 stands out, a bit of 5
. After listening to other classes, the pitch seems lower (I don't thing it is actually lower, at least for the first harmonic)
. Muffled, badly balanced, « too low pitch »
Class 2
. Average rich sounds
. harmo 3 stands out, not much 5
Class 3
. Richer sounds
. harmo 3 stands out, some 5, much 6

Fig. 4. SX (pianist) categorization projected on the representation of the sound samples as a function of the model parameters.

We can also already notice (as shown on Fig. 5) that the mapping of the cognitive categorization onto the physical description is not continue. 2D and 5D are grouped together within the same subcategory even if they are constructed on two extreme values of the "phantom" partial level. Considering the verbal comments, this "incoherence" means that subjects categorize the two stimuli together because they are "synthetic" or "bizarre", that is, they are different from what can be considered by the subjects as "common" sounds of a "good piano". In other words, cognitive categorization not only relies on "intrinsic" properties of the stimulus but also on similarities or discrepancies from "common" stimuli that the subjects frequently encountered and therefore experienced and memorized. These observations indicate that, for some values of a given physical parameter (here from 1,3*$B0$), the cognitive categorization is not related in a monotonous manner to a physical parameter, but operates according to mental representations elaborated from the subject's previous experiences. Moreover, the traditional timbre descriptors of the signal proposed in Sect.3 can not explain the categorization given by the subjects. We can thus find stimuli of very different centroids or spectral flux within the same category. The non-pianist subject sometimes seems to choose to group stimuli with respect to their inharmonicity (group 2A, 3A, 4A, 5A), sometimes with respect to the similarities of their centroids (group 3C, 3D 4C, 4D). The relation between groups given by the first seconds of the spectral flux and the subject categorization is not trivial. Thus, 1B and 4B have been grouped whereas their centroids as well as their spectral flux are different. These observations indicate that cognitive categorization can give us new ideas when looking for meaningful timbre descriptors to distinguish similar sounds. As already mentioned the cognitive categorization partly relies on similarities or discrepancies from "common" stimuli, meaning that one possible timbre descriptor could be related to a predefined distance from a reference sound.



B_0 / G	0.5* B_0	B_0	1.3* B_0	1.6* B_0
Original sound		1B		
0	2A	2B	2C	2D
0.6*G	3A	3B	3C	3D
1.1*G	4A	4B	4C	4D
1.5*G	5A	5B	5C	5D

String stiffness
Non-linear coupling rate
Same attack with humming
Broad, heavy
Synthetic
Clearer attack with the development of high pitch components
Muffled attacks with the development of high pitch components
Aggressive attack with strident sustain

Fig. 5. SY (musician non-pianist) categorization projected on the representation of the sound samples as a function of the model parameters.

For the full experiment on a higher number of subjects, we will have to identify the semantic properties attributed to the same physical descriptions by subjects with different expertise and knowledge. In particular, we could like to verify the hypothesis stipulating that the categorization given by the experts (pianists) are more closely linked to the structures described by the physical parameters, and especially when it comes to the inharmonicity, while the non-experts used a different categorization strategy associating physical parameters of the model with a semantic interpretation integrating other criteria like the spectral centroid. The analysis of the verbal comments will be used to complete our interpretation.

As opposed to free audio categorization tests which have been done before, where subjects tend to classify stimuli regarding to the nature of the supposed source which produced them (door shock, engine noise [27]), stimuli are here in general grouped with respect to their perceptive properties ("muffled", "broad", "round"...). This is due to the fact that timbre variations, and consequently categorization levels, are here extremely subtle. The qualitative properties of stimuli coming from a same source (piano), are treated in an "intensionnal" way. Thus, the aim is here to listen to the sound itself as a contrast to an indicial way of listening [28]. Certain subjects sometimes can refer to a particular kind of source like "synthetic piano" as a qualification of a stimulus, meaning that they base the criteria of extensional classification of the sound producing system rather than on the sound itself.

If we look closer at the results of the free categorization task, for different values of our two "physical dimensions", we can state that

– when subjects judge the sound as a function of the inharmonicity, they use expressions like "muffled", "poor" and "mate" for the least inharmonic sounds, "between muffled and clear", "average rich sound" for the medium inharmonicity and "clear" or "distorded", "richer sound" for a maximum value of inharmonicity. An increase in inharmonicity therefore seems to "brighten" or "clear" the sounds from a perceptual point of view.
– when subjects judge the sound as a function of the "phantom" partials (which corresponds to the second categorization level for pianist subjects), the sounds with few partials are classified as "hollow", "damped", while for higher amounts of "phantom" partials they are classified as "clear", "round", "plain", "balanced" and finally for the maximum value of "phantom" partials as "slightly metallic", "aggressive". An increase in the number of "phantom" partials therefore seems to converge with an increase in the inharmonicity, constructing categories of more and more bright sounds.

The second criterion (presence of "phantom" partials) contributes to the construction of categories in which the tendency "muffled" or "dark" given by the weak levels of inharmonicity is corrected and renormalized. Finally, for the category constructed with a strong inharmonicity, the extreme stimuli are considered as "synthetic". Again these terms demonstrate the important gap between the norm of the human category jugdment.

5 Conclusion and perspectives

This study shows the major difference between physical and cognitive descriptions, i.e. the dimensional character of the first one and the categorical character of the other. This difference can be seen by the deviations from the value which is considered as "normal" (sound close to the original one) within the same category of values in different registers. Thus, from a certain "level", the categorization induces a discontinuity compared to the physical parameters. From this separation in their categorical belonging, the stimuli are classified as a function of their distance to a prototype sound (either previously memorized from the experiences of the subjects, or constructed as a mean element within the group of stimuli). In other words, different categorical structures can correspond to the unique description of the stimuli in the physical space. These structures depend on different strategies that supposedly rely on the variations in expertise and experience of the different subjects. Thus, the non-pianist subjects construct their categories differently from one part to the other of the inharmonicity level. At this level the categorization is effectuated from the stimulus sharing the values of the two parameters of the model taken simultaneously. When the subjects possess a very high degree of expertise, the processes of categorization are no longer related to the qualitative properties of the signal, but on extensional criteria belonging to categories of objects with a close differentiation (here different piano types). It is therefore necessary to pay attention to the characterization of different groups of subjects. We have also noticed that the use of this diversity of categorization processes was sensitive to the characteristics of the stimuli. Thus, certain stimuli are for instance too far from the prototype, too atypical, have a construction that is too distant for the subjects to be able to constitute a representation of a “real” sound and correspond to an indication of a source. In such cases they are treated analytically, directly on the properties of the signal (already observed in visual experiments [3]). When a larger number of subjects are to be tested, it will therefore be necessary to take care of the validity of the representations of the different stimuli related to the “natural” sounds. We finally will strive to find new descriptors of the acoustic signal which represent categorical structures of the cognitive analysis, regrouping all of the parameter values that do not necessarily correspond to the same physical dimension. Finally, the coupling between the categorical analysis and the verbal comments will make it possible to extract semantic descriptors qualifying the perceptual effects of the two phenomena studied here. This will further give new clues to the definitions of new timbre descriptors adapted to subtle timbre variations of piano sounds.

References

1. Rosch, E.: Principles of categorization. In: Lloyd, B.B., Erlbaum , L. (eds.): Cognition and categorization. Hillsdale (1978) 27–47
2. Dubois, D. (ed.): Sémantique et Cognition. Editions du CNRS, Paris (1991)
3. Dubois, D. (ed.): Catégorisation et cognition : de la perception au discours. Kimé, Paris (1997)
4. Guyot, F.: Etude de la perception sonore en termes de reconnaissance et d’appréciation qualitative : une approche par la catégorisation. Thèse de doctorat, Université du Maine, Le Mans (1996)
5. Maffiolo, V.: De la caractérisation sémantique et acoustique de la qualité sonore de l’environnement sonore urbain. Thèse de doctorat, Université du Maine, Le Mans (1999)
6. Gaillard, P.: Etude de la perception des transitoires d’attaque des sons de steeldrums : particularités acoustiques, transformation par synthèse et catégorisation. Th‘ese de doctorat, Université de Toulouse II, Toulouse (2000)

7. Jensen, K.: Timbre models of musical sounds. Ph.D. thesis, DIKU press, Copenhagen, Denmark (1999)
8. Askenfelt, A. (ed): Five Lectures on the Acoustics of the Piano. Royal Swedish Academy of Music, Stockholm (1990)
9. Chaigne, A., Askenfelt, A.: Numerical simulations of struck strings. I. A physical model for a struck string using finite difference methods. J. Acoust. Soc. Amer, Vol. 95(2). (1994) 1112–1118
10. Smith III, J. O., Van Duyne, S.A.: Commuted piano synthesis. Proc. Int. Computer Music Conf., Banff (1995)
11. Avanzini, F., Bank, B., Borin, G., De Poli, G., Rocchesso, D.: Musical instrument modeling: the case of the piano. Proc. of the worshop on current research directions in computer music. MOSART Research training network (2001)
12. Bank, B.: Physics-based sound synthesis of the piano. Master thesis, Budapest University of Technology and Economics, Hungary (2000). Published as Report 54 of HUT Laboratory of Acoustics and Audio Signal Processing. URL:http://www.mit.bme.hu/ bank
13. Bensa, J.: Analysis and synthesis of piano sounds using physical and signal models. Ph.D. thesis, Université de la méditerranée (2003). URL:http://www.lma.cnrs-mrs.fr/∼bensa
14. Smith, J. O.: Digital Waveguide Modeling of Musical Instruments. Available online at http://ccrma.stanford.edu/˜jos/waveguide
15. Bensa, J., Daudet L.: Efficient modeling of “phantom” partials in piano tones. Proc. of the International Symposium on Musical Acoustics, Nara, Japan (2004).
16. Fletcher, H., Blackham E.D., Stratton S.: Quality of Piano Tones. J. Acoust. Soc. Amer, Vol. 34(6). (1961) 749–761
17. Young, R. W: Inharmonicity of plain wire piano strings. J. Acoust. Soc. Amer, Vol. 21. (1952) 267–273
18. Conklin Jr., H.A.: Generation of partials due to non-linear mixing in stringed instrument. J. Acoust. Soc. Amer, Vol. 105(1). (1999) 536–545
19. Valette, C., Cuesta, C.: Mécanique de la corde vibrante. Traité des nouvelles technologies. Série Mécanique, ed. Hermès (1993)
20. Bank, B., Sujbert, L.: Modeling the longitudinal vibration of piano strings. Proc. Stockholm Music Acoustics Conf. (2003)
21. Aramaki, M., Bensa, J., Daudet, L., Guillemain, Ph., Kronland-Martinet, R.: Resynthesis of coupled piano strings vibrations based on physical modeling. J. of New Music Research, Vol. 30(3). Swets & Zeitlinger, (2001) 213–226
22. McAdams, S.: Recognition of sound sources and events. In: McAdams, S., Bigand, E. (eds.): Thinking in Sound: The Cognitive Psychology of Human Audition, Oxford Univ. Press. (1993) 146–198
23. Menon, V., Levitin, D.J., Smith, B.K., Lembke, A., Krasnow, B.D., Glazer, D., Glover, G.H., McAdams, S.: Neural Correlates of Timbre Change: In Harmonic Sounds NeuroI-mage, Vol. 17. (2002) 1742–1754
24. Beauchamp, J.: Synthesis by spectral amplitude and “Brightness” matching of analyzed musical instrument tones. J. of the Audio Engenering Society, Vol. 30(6). (1982) 396–406
25. McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., Krimphoff, J.: Perceptual scaling of synthesized musical timbres: Common dimensions, specificties, and latent subject classes. Psychol. Res. Vol. 58. (1995) 177–192
26. Grey, J. M.: Multidimensional perceptual scaling of musical timbres. J. Acoust. Soc. Am., Vol. 61. (1977) 1270–1277

27. Dubois, D.: Categories as acts of meaning: the case of categories in olfaction and audition. Cognitive Science Quarterly Vol. 1. (2000) 35–68
28. Castellengo, M.: Perception auditive des sons musicaux. In : Zenatti, A. (ed): Psychologie de la musique. PUF Paris (1994) 55–86
29. Dubois, D., Resche-Rigon, P., Tenin, A.: Des couleurs et des formes : catégories perceptives ou constructions cognitives. In : Dubois, D. (ed.): Catégorisation et cognition : de la perception au discours. Kim, Paris (1997) 7–40

Sølvi Ystad and Thierry Voinier
Laboratoire de Mécanique et d'Acoustique
Centre National de la Recherche Scientifique (CNRS)
31 Chemin Joseph Aiguier
13402 Marseille Cedex 20, France
[ystad, voinier]@lma.cnrs-mrs.fr

A Virtually Real Flute

Since the first keyboard-controlled digital synthesizers became available, several new synthesis interfaces have been developed (e.g., Mathews 1991a, 1991b; Cook 1992; De Laubier 1998). As most of these digital instruments differ considerably from traditional instruments, musicians must learn new techniques to play them (Kronland-Martinet, Voinier, and Guillemain 1997). Here, we propose overcoming this difficulty by designing a digital flute using a traditional instrument form factor to control a synthesis model. The digital flute was assumed to extend the technical scope of the traditional flute, but we also wanted to be able to use the instrument in the traditional way. To connect the instrument to a computer, we added sensors to its key pads and placed a microphone inside the mouthpiece. The synthesis model to be controlled by this interface had to take the physical characteristics of the instrument into account. A physical model was therefore developed to simulate the propagation of waves inside the flute.

The system of excitation involved in flute-playing is highly complex from a physical point of view. To construct a real-time model with parameters that can be measured while the instrument is being played, we used a signal model to simulate the source excitation. By injecting this model into the physical one, we constructed a hybrid model which accounts for both the physical and perceptual aspects of the sound produced.

Design of the Interface

Playing a wind instrument involves two main factors. The first of these is the player's finger position, which is correlated with an equivalent length of the instrument (Nederveen 1998) and thus with the pitch, and the way the instrument is excited by the air jet (Fletcher and Rossing 1990; Verge 1995). This information must be detected and combined in order to control a synthesis model in real time or to produce standard MIDI messages.

Computer Music Journal, 25:2, pp. 13–24, Summer 2001

Pitch Detection

The pitch is determined by both the player's finger position and the way the instrument is blown. Signal processing methods can be used to analyze the sound emitted by the instrument and accurately estimate the pitch. Since the flute is a monophonic instrument, a pitch extractor of this kind can be used to directly perform the MIDI encoding of musical sounds without having to solve the problems associated with polyphonic instruments. The `fiddle~` MSP object (Puckette, Apel, and Zicarelli 1998) is a good example of an available tool which is well suited to applications of this kind. In this case, the instrument only needs to be equipped with a microphone connected to the sound input of a computer running an MSP program.

In our case, we wanted to be able to control the synthesis model with the real instrument, even when the flute is not blown. The state of the key pads therefore must be detected to obtain information about the player's finger position. In addition, the key pad noise is of musical relevance, and we therefore had to collect information of another type: the speed at which the key is pressed. To detect the time-varying position of the keys, we used a combination of magnets and Hall effect sensors. A Hall effect sensor gives an output voltage which is a function of the magnetic field received. If the magnetic field is generated by a permanent magnet, its intensity will depend on the square of the distance between the sensor and the magnet. The output voltage of the sensors is then correlated with the spatial distance between the key pads and

Figure 1. Close-up view of the flute equipped with magnets and sensors.

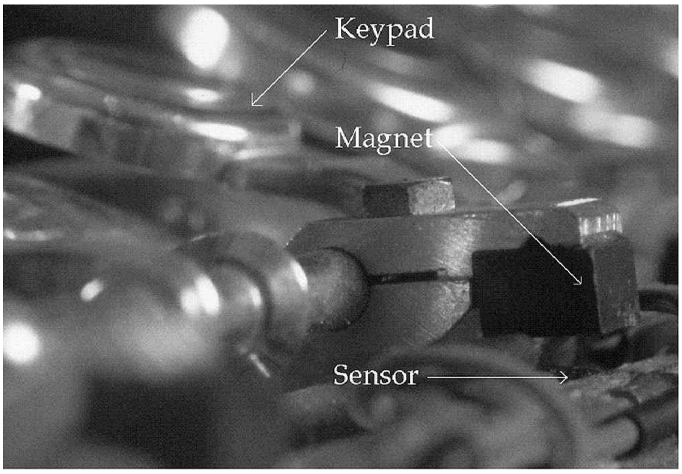
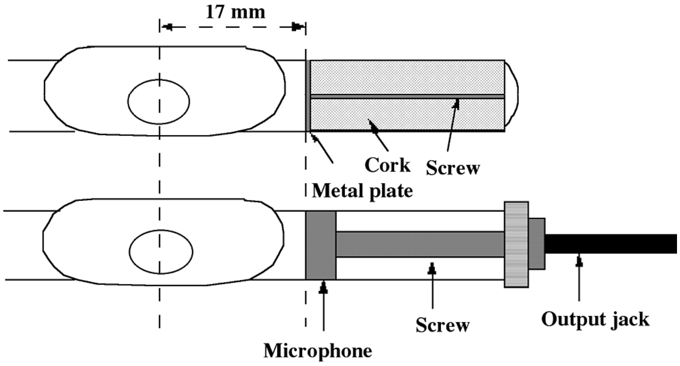


Figure 2. Modification of the mouthpiece of the traditional instrument.



the corresponding holes, which makes it possible to detect whether the hole is opened or closed. By regularly sampling this voltage, we can estimate the speed at which each key is pressed or released. For practical reasons, the magnets were connected to each finger key on the instrument, while the Hall effect sensors were placed in front of each magnet on an aluminum rail placed parallel to the instrument, as shown in Figure 1.

The magnetic field generated by the magnets had to be strong enough to obtain a suitable output signal from the sensors. The two main states of the holes (opened and closed) had to be clearly distinguishable. The magnets were chosen so that the neighboring sensors were not influenced by the magnetic field.

An Infusion Systems I-Cube System was used to digitize data from the sensors. This system digitizes the sensors' output voltages and sends this data to a Macintosh computer running Max. With this interface, it is possible to sample 14 sensors at a rate of 50 Hz with a resolution of 7 bits, which suffices for this application. (A Max object, iCube, is provided with the hardware, and a Max patch can easily be created in order to process the data from the sensors.) The processing consists mainly in checking whether each aperture is in the open or closed state, and then finding the corresponding pitch in a lookup table. In this case, wrong fingering would not be recognized, and the last valid pitch detected would, for example, remain activated.

Measurement of the Excitation

The way the instrument is excited is a highly complex process. It depends on several parameters, such as the player's lip position and the angle at which the air jet hits the labium of the embouchure. These important features are difficult to measure, thus we decided to concentrate on detecting the internal pressure, which depends on the way the instrument is being played.

To measure the acoustic pressure produced by the instrument, a microphone was placed at the embouchure level (near the mouthpiece). More specifically, the original cork with which the instrument is fitted was removed and replaced by a custom assembly containing the microphone, enabling the instrument to be finely tuned. This device is shown in Figure 2.

A moisture-resistant electrodynamic microphone able to handle high acoustic pressure (approximately 140 dB SPL) was placed inside the flute pipe, and the microphone signal was delivered to the hardware sound input on the Macintosh. The signal was sampled at the audio rate and processed by a peak detector (McNally 1984) providing the pressure envelope. The pressure envelope was then sampled at a lower rate (50 Hz) and used to trigger note-on and note-off MIDI messages in which the associated pitch is given by the state of the holes. A schematic description of the MIDI generator is given in Figure 3.

MIDI compatibility makes it possible to connect the flute interface to other MIDI instruments and to control them in different ways. A flautist uses

Figure 3. Overview of the flute's MIDI component.

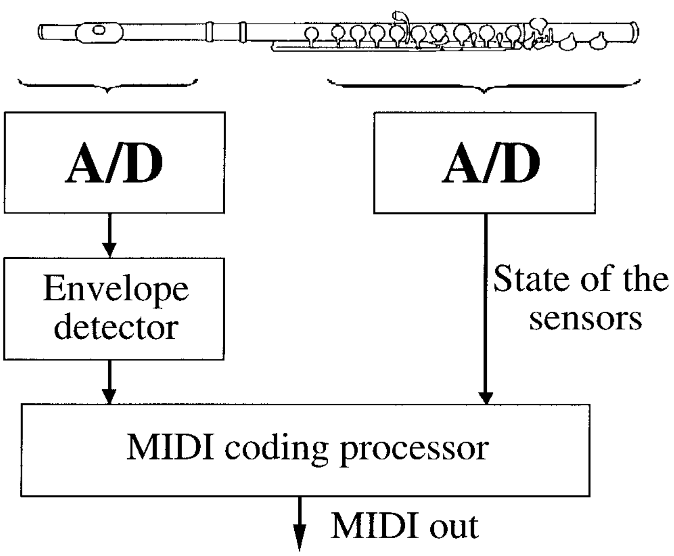
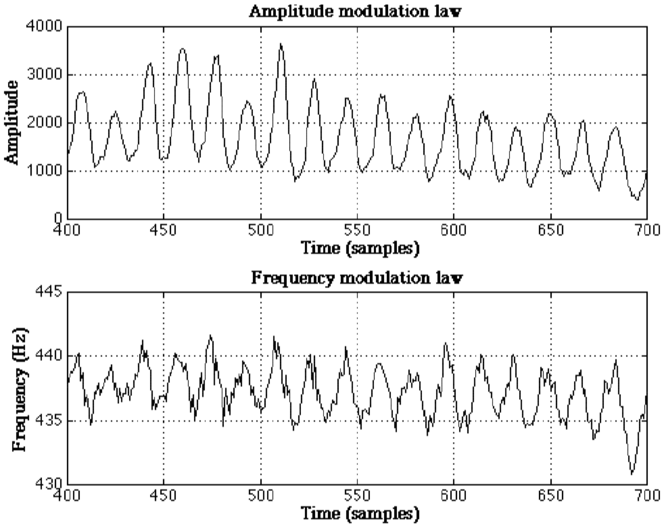


Figure 4. Amplitude and frequency modulation laws of the fundamental component of a flute sound (A4) with vibrato.



pressure variations, for instance, to produce effects such as tremolo and vibrato (Fletcher 1975). To investigate the relationship between pressure variations (tremolo) and frequency variations (vibrato), amplitude and frequency modulation laws can be estimated using time–frequency techniques (Kronland-Martinet and Grossmann 1991). In Figure 4, a close view of the amplitude and frequency modulation laws (fundamental component) of A4 (440 Hz) flute sound with vibrato is shown. We can see that the vibrato is clearly in phase with the amplitude variation (Ystad 1998), which means that the frequency modulation can be controlled by variations in the air jet pressure. With the new instrument, effects of this kind can be used to produce MIDI messages such as aftertouch and pitch bend.

At this stage, the traditional instrument can convert the fingering and even the sounds produced into MIDI information. This method could be used to drive any MIDI synthesizer. It could also be used to study performance and interpretation—or, after some programming—to accompany the flute player with another MIDI instrument. Since the I-Cube system is able to manage 32 inputs, other analog inputs can be used for additional devices that either trigger MIDI messages or control synthesis parameters. Some of these possibilities will be described in the last section of this article.

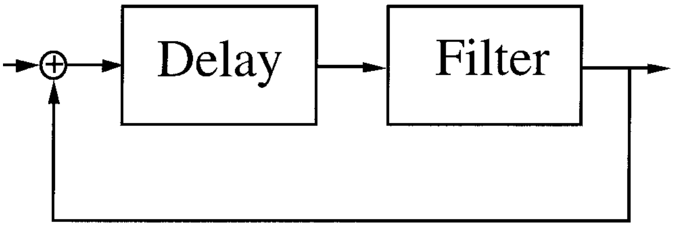
The Synthesis Model

Although the MIDI part of this interface is of musical interest, its main advantage is the fact that it can be used to control sound synthesis models for wind instruments in a natural way. These synthesis models must be constructed with the specific interface in mind. This means that the model must be able to operate in real-time and that its parameters must be linked to the interface in a natural way. We decided to implement a source-resonance model in which both physical and signal models are combined (Ystad 1998, 2000). This synthesis model takes into account many physical features of the sound-producing system as well as many perceptual features captured in the spectral representation.

Physical Model

The propagation of waves inside the flute can be simulated using physical models. These models can either be constructed from the equations describing the behavior of the waves propagating in the structure and their radiation in air (Chaigne 1995) or from the behavior of the solution of the same equations (Karjalainen et al. 1991; Cook 1992; Smith 1992). We opted for the latter alternative by constructing a waveguide model consisting

Figure 5. The waveguide model.

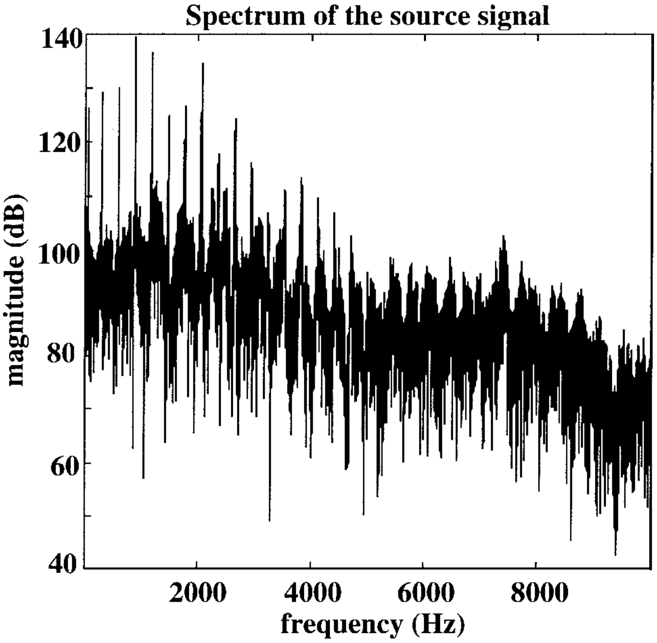


of a looped system with a delay line simulating the propagation time of the waves and a filter simulating both dissipation and dispersion phenomena. The waveguide model is shown in Figure 5.

We have proposed a new way of constructing the loop filter (Ystad 1998) related to the inverse of a time–frequency representation for a transient sound (Guillemain and Kronland-Martinet 1996). To construct the filter, the damping factors and the frequencies of the eigenmodes in the tube must be found. These parameters can either be calculated from theoretical equations describing waves propagating inside a tube (Kergomard 1981; Fletcher and Rossing 1990) or be obtained by performing time–frequency analysis on real sounds (Kronland-Martinet 1988). In a performance situation, the player can control these parameters via sliders. This gives the user the possibility of changing the characteristics of the resonator of the instrument during the performance. One can thus make cross synthesis by combining the flute source excitation with the resonator of any instrument (or conversely), or doing continuous morphing—for instance between a flute-like sound and a string-like sound. The finger position (which is detected by the sensors on the flute) is used to compute the value of the delay line.

The physical processes involved in the excitation of a flute instrument are much more complex. The air jet from the player's mouth hits the labium of the embouchure, and this interaction transfers energy to the acoustic standing waves in the resonator (Coltman 1968; Verge 1995). Flow visualizations of the jet/labium interactions have shown the occurrence of vortical structures on each side of the jet (Fabre, Hirschberg, and Wijmands 1996). This means that a complete physical model of the excitation system would be very difficult to implement in real-time. In addition, the parameters involved in the equations would be difficult to measure while the flute is being played. Therefore we decided to use signal models to simulate the source.

Figure 6. Spectrum of the source signal of a flute sound (D1 with a dynamic level corresponding to mf) obtained by deconvolution.



Signal Model

To construct a signal model simulating the excitation of a flute, we first had to extract the source from the rest of the flute sound. From a physical point of view, the source and the resonator cannot be separated because they interact constantly while the instrument is being played. In our case, however, the separation of the source and the resonator turned out to be a good approximation. As mentioned, we previously developed a model simulating the resonator of the instrument. By removing the contribution of the resonator from the total signal, we obtained the source signal. Because the transfer function of the physical model corresponds to a recursive all-pole filter, we know that its inverse exists. This means that we can extract the source signal from the total flute sound by deconvolution (Ystad 1998). Figure 6 shows the deconvolved signal extracted from a flute sound.

This figure shows that the spectrum of the source signal contains both spectral lines (harmonics) and a broadband noise (which will be called in what follows the *deterministic* and the *stochastic* contributions, respectively). We proposed to separate these two contributions to model them independently. Among the many methods available for this purpose (e.g., Serra 1989), the Least Mean Square (LMS) algorithm, which uses an adaptive filter, was found to be the most suitable for dealing with the problem (Widrow and Stearns 1985). This method involves removing all the components from an input signal that are correlated with a reference signal. By using an estimation of the deterministic signal (harmonics) as a reference signal and the source signal as an input signal, we obtained the stochastic signal (Ystad 1998, 2000).

Modeling the Deterministic Part

The deterministic part of the source signal was found to have nonlinear behavior, because the amplitudes of the spectral components evolve differently from each other as the excitation energy increases. This is the case for most musical instruments, whose timbres depend greatly on the dynamic nature of the sound excitation. In most cases, the nonlinear behavior is correlated with the excitation, and we assume this to be the case here. To model these nonlinearities, we used a global synthesis method, namely the waveshaping method (LeBrun 1979; Arfib 1979), because it provides a useful means to generate complex spectra from easy calculations by performing only a small number of operations. This method consists of distorting a sinusoidal function with an amplitude function *I(t)* (called the index of distortion) with a nonlinear function γ. The function γ can easily be linked to the spectrum of the sound generated for an index value *I(t)* = 1. In this case, the coefficients of the Chebyshev decomposition of γ are given by the values of the modulus of the spectrum to be generated:

$$\gamma(\cos\omega_0 t) = \sum_{k=0}^{\infty}\alpha_k T_k(\cos\omega_0 t) = \sum_{k=0}^{\infty}\alpha_k \cos k\omega_0 t \qquad (1)$$

The index of distortion is said to be bounded (–1 ≤ *I(t)* ≤ 1), and the waveshaping function will be chosen so that the synthetic signal obtained for *I(t)* = 1 corresponds to the richest part of the real signal (i.e., a *fortissimo* sound). The goal is then to associate with the waveshaping index a measurable value such as the driving pressure to control the spectral evolution of the synthetic signal. One great disadvantage of the global synthesis technique is that the representation of signals is not complete, and it is therefore not possible to reconstruct any spectral evolution by simply changing the index. Nevertheless, as we shall see, the index can be estimated so that the reconstructed signal satisfies perceptual criteria.

One well-known perceptual criterion is the spectral centroid criterion (Beauchamp 1982). It relates to the brightness of a sound and is given by the first order moment of the spectrum. We first applied this criterion to the case of the flute, but discovered that it did not work because very few of the components (mainly the first through sixth) in a flute spectrum change with the dynamic level of the sound. This means that greater importance must be given to these components. We therefore adopted another perceptual criterion called the tristimulus criterion (Pollard and Jansson 1982). This criterion deals with the loudness of three separate parts of the spectrum: one where the evolution of the fundamental is considered; one where the second, third, and fourth components are considered; and one where the rest of the components are considered. The loudness value of each group can be computed using Stevens' formula (Stevens 1972):

$$N_i^n = 0.85\,N_{\max} + 0.15\sum_i^n N_i \qquad (2)$$

where N_i^n is the required equivalent loudness (for the group which contains components *i* through *n*), $N_{\max}$ is the loudest part of the group, and $\sum_i^n N_i$ is the loudness of all the partials in the group. The total loudness *N* of the sound is then given by the sum of the three loudness groups:

$$N = N_1 + N_2^4 + N_5^n \qquad (3)$$

With this method, the tristimulus can be given in an acoustic tristimulus diagram, where

$$x = \frac{N_5^n}{N}$$

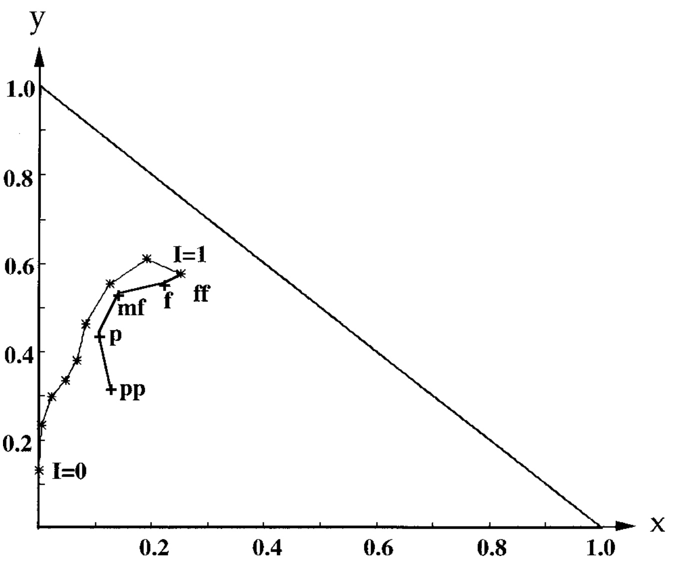


Figure 7. Tristimulus diagram of real(+) and synthetic(*) flute sounds.

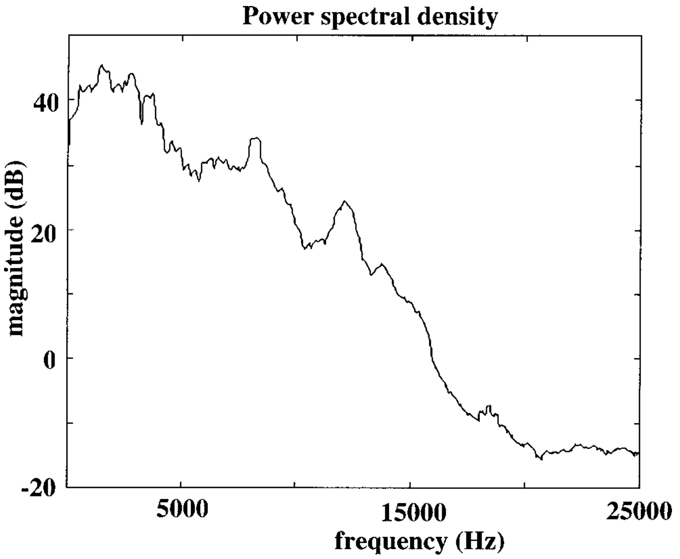


Figure 8. Power spectral density of the stochastic part of the source signal (the same pitch as in Figure 6).

$$y = \frac{N_2^4}{N},$$

and

$$z = \frac{N_1}{N}.$$

Since $x + y + z = 1$, it is sufficient to use two of the coordinates(x and y) to draw the tristimulus diagram as shown in Figure 7. Here, the tristimulus of the sounds generated by waveshaping synthesis with index values ranging from 0 to 1 are represented, along with five flute sounds with different dynamic levels (*pianissimo* to *fortissimo*). The nonlinear function was chosen so that the spectrum generated would coincide with the real fortissimo spectrum for $I = 1$. Since the tristimulus is a perceptual criterion, only its global behavior is important. This means that the real and the synthetic sounds will not show exactly the same behavior, but they will be located in the same area of the diagram and have the same global evolution.

By minimizing the difference between the real and the synthetic flute sounds, we observed that the index of the waveshaping function tends to vary from $I = 0.5$ to $I = 1$, depending on the logarithm of the driving pressure. Consequently, the spectral evolution of the source signal can be controlled by the pressure from the player's mouth, which was measured with the microphone replacing the cork.

Modeling the Stochastic Part

The stochastic part of the source signal was assumed to be stationary and ergodic. That is, we assumed the excitation noise can be described by its power spectral density and its probability density function. From the perceptual point of view, the "coloring" of the noise is mainly related to its power spectral density. Its probability density function $f_B(x)$ can also be relevant. It is related to the histogram of the values χ involved in the noisy process B. It can be easily estimated provided that the random process can be separated from the deterministic one, which is generally true in the case of source signals. In the case of the flute, the histogram is symmetric and follows an exponential law. The power spectral density of a flute sound is shown in Figure 8. By filtering broadband noise (the response of which is given by the extracted power spectral density), one can generate a satisfactory flute source noise.

This model, together with the model of the deterministic part of the source signal, gives a general model of the source signal based on signal modeling procedures. By combining the source model with the physical model simulating the behavior of the

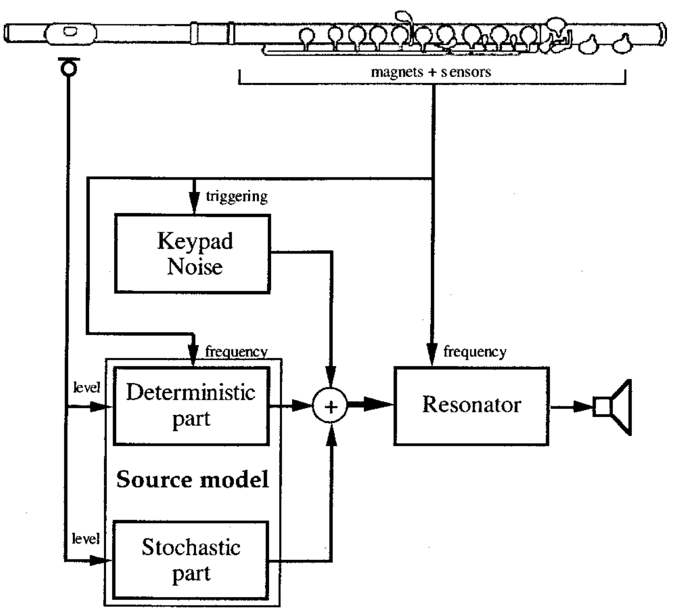


Figure 9. The hybrid model obtained by combining a signal model with a physical model.

waves while they are propagating through the medium, very general sound models can be constructed. In the next section, we shall see how this general model can be applied to the case of the flute.

Hybrid Model

We have called the complete model a *hybrid model*, because it is a combination of two classes of synthesis models: signal models and physical models. This is a very powerful model, as it benefits from advantages of both classes of sound models. Figure 9 shows the flute model, consisting of the source model containing the deterministic and stochastic contributions, and the physical model simulating the resonator of the instrument. To complete the picture, we have added a third part to this synthesis model. In this part, the goal was to generate fixed sounds such as the noise produced by the key pads. This was done using a sampler-like method which consists of reading a previously stored signal designed so that its passage through the resonator gives a realistic key pad impulse. This signal can be obtained, for example, from recordings of the real key pad noise. It would also be of musical interest to use sounds from percussive instruments at this stage.

The Musical Context

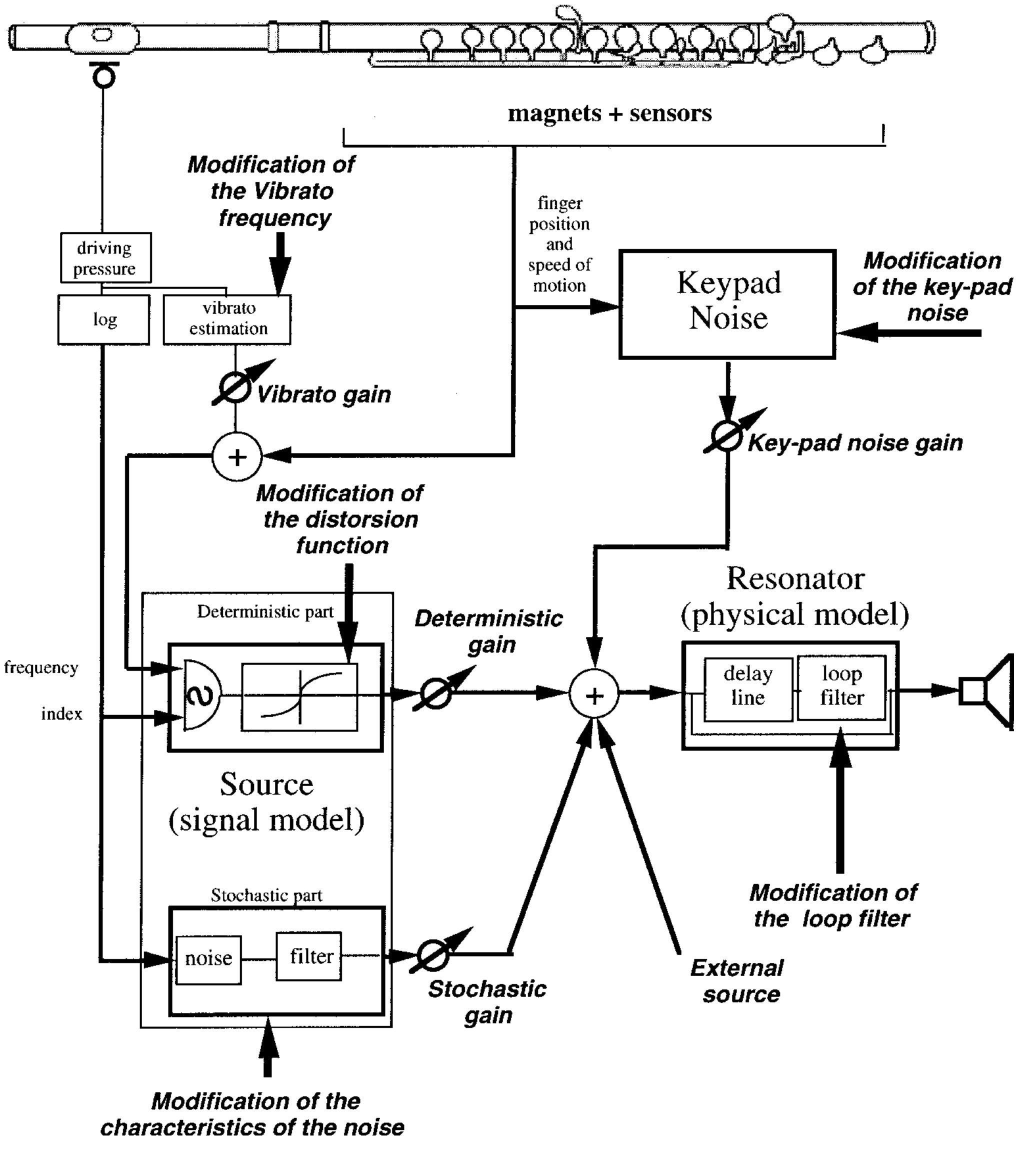
In a musical context, the "virtually real" flute can be played in various ways. It can still be played like a traditional flute, it can be used like a traditional instrument while controlling the synthesis model in real-time, and it can be muted by inserting damping material near the mouthpiece so that only the sounds from the synthesis model can be heard. Additionally, it can simply be used as a MIDI controller to drive other MIDI instruments.

When the instrument is used with the synthesis model, it can effect sound transformations. In the first part of this section, we describe the sound transformations which can be obtained by varying different parameters of the synthesis model. There are many possibilities, and in a performance situation it is therefore necessary to limit the number of parameters to be modified to prevent the instrument from becoming too complicated for the performer. In the second part of the section, we give an example of an application of the instrument where the player is able to modulate four different parameters of the model and where the instrument also controls a Yamaha Disklavier MIDI piano.

As mentioned earlier, the interface was designed so that the possibility of playing the flute using traditional techniques could be retained. This means that musicians can use their hands and mouths to vary the frequency and the pressure, as under normal performance conditions. The player may therefore have to use other parts of the body to regulate the parameters of the synthesis model in order to make sound transformations. The I-Cube System which is used to power the Hall effect sensors connected to the key pads of the instrument makes it possible, for instance, to use volume pedals to control the model's parameters. In Figure 10, we have indicated some of the parameters of the synthesis model which can give interesting sound effects.

The source signal of the flute model comprises three different contributions which must be calculated and mixed before being processed by the resonator. These contributions correspond to the deterministic part, the stochastic part, and the noise generated by the key pads. The deterministic part of the source signal consists of an oscillator

Figure 10. Control of the model's parameters. The possible modifications are shown in an italicized, boldface font.


magnets + sensors
Modification of the Vibrato frequency
finger position and speed of motion
driving pressure
log
vibrato estimation
Keypad Noise
Modification of the key-pad noise
Vibrato gain
Key-pad noise gain
Modification of the distorsion function
Deterministic part
Deterministic gain
Resonator (physical model)
delay line
loop filter
frequency
index
Source (signal model)
Stochastic part
noise
filter
Stochastic gain
External source
Modification of the loop filter
Modification of the characteristics of the noise

and a nonlinear function, while the stochastic part consists of filtered noise. The driving pressure controls the stochastic part as well as the amplitude of the oscillator component of the deterministic part. The sensors connected to the key pads detect the player's finger position, and thus controls the delay line of the resonator as well as the frequency of the oscillator. The speed at which the key pads are closed is also detected and used to control the key pad noise. By bandpass filtering the pressure envelope, the vibrato can be estimated and added to the frequency component of the oscillator, causing fluctuations in the resonance peaks of the source. This means that when the source is injected into the resonator, the resonance peaks of the source and those of the resonator will not be tuned all the time. The output amplitude of the system will therefore fluctuate and be stronger when the two systems are tuned than when they are not tuned. The amplitude fluctuations (tremolo) will therefore follow the frequency fluctuations (vibrato) as on a traditional flute.

With all the components that constitute the synthesis model in mind, we can now describe how to produce interesting timbral variations by controlling them independently. By changing the gain of the filter, the depth of the vibrato can be changed. Special shapes of vibrato can also be artificially generated. The waveshaping index can also be controlled. The waveshaping index is a highly sensitive parameter that was estimated to fit the spectral evolution of the flute sound. Nevertheless, a change in the correspondence between the internal pressure and the distortion index can be envisioned. The flute can be given a brassy effect, for example, by increasing the variation domain of the distortion index.

Changing the characteristics of the distortion function dramatically affects the timbre of the deterministic part of the source signal. A distortion function with a decomposition that contains only odd Chebychev polynomials can be used to generate a clarinet-like or pan flute–like source, for example.

The characteristics of the noise can be modified via the noise filter (power spectral density) and the statistics (probability density function). The relationship between the deterministic and stochastic parts of the source signal can also be changed by adjusting the noise gain. If the deterministic part is removed, then the resulting sound would be a noise filtered by the resonator.

The level of the key pad noise can be adjusted by adding a gain to the key pad noise table output. If both the deterministic and stochastic parts of the source are removed, the resulting sound will correspond to that obtained by closing the key pads. The key pad noise can also be altered by modifying the corresponding table and could be replaced by any percussive sound.

One can also use an external input with its own level control to drive the resonator via an external signal. This can be used to take the various noises made by the player's mouth into account, for example.

The loop filter of the resonator can also be adjusted. The loop filter characterizes the resonator and represents dissipation and dispersion phenomena present in the bore. Altering this filter will change the characteristics of the medium in which the waves are propagating. Cross-synthesis effects can be obtained using parameters corresponding to the source of one instrument and the resonator of another instrument. Using a loop filter corresponding to a string with a flute excitation, a very particular sound can be generated, corresponding to blowing "into" a string. Likewise, external sources to be filtered by the resonator can be added. This would make it possible, for example, to generate the noises made by the flautist while playing.

One can also vary the delay line's length and the oscillator's frequency. By changing the offset of these parameters, one can simulate instruments with unrealistic sizes, such as an extremely long flute.

All these manipulations show the advantages of sound modeling. With this particular sound model, one can model not only synthetic sounds, but also natural sounds.

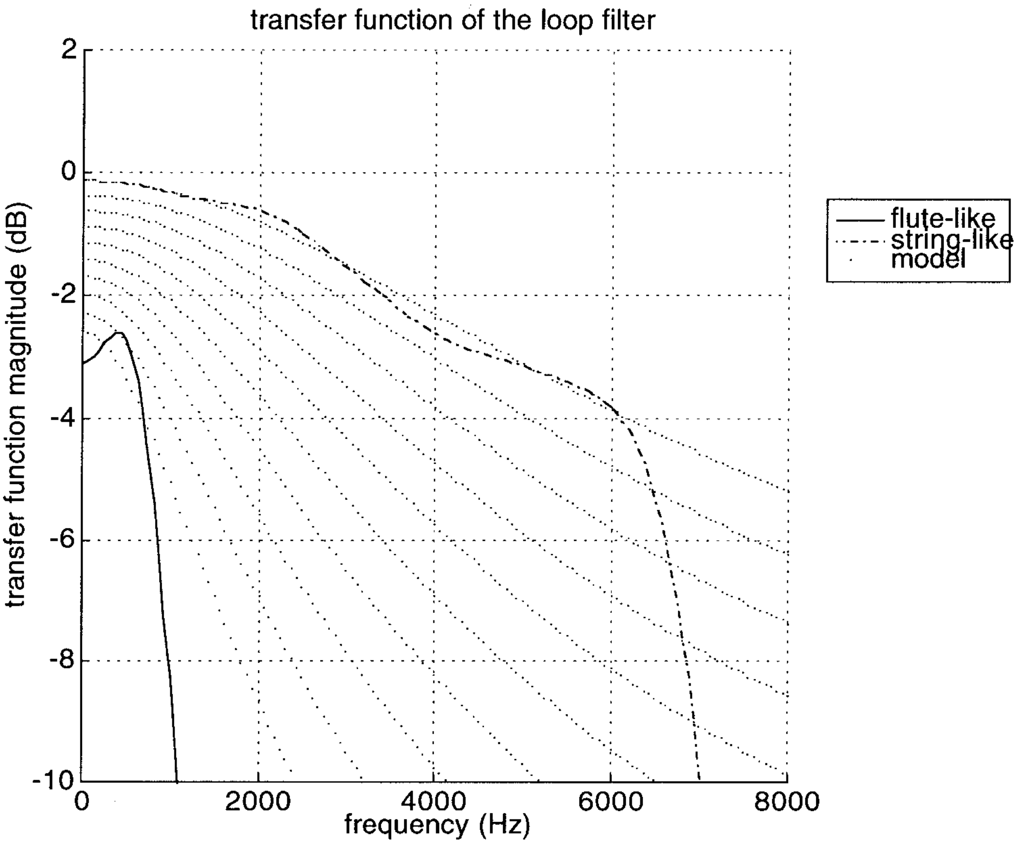
A special thought must be given to the diffusion problem. The instrument has several outputs: one corresponding to the acoustic signal, the others corresponding to the outputs of the synthesis model at different stages. These outputs can be post-processed (with reverberation, equalization, etc.) separately and then diffused or spatialized in the concert hall. This offers many possibilities for live performances, such as independent spatialization effects for each component of the sound. These effects can be controlled either by the flautist (using pedals or specific actions on the instrument), or programmed in advance.

To end this section, we give an example of an application of the digital flute where the player can control four different parameters with pedals, and where the flute also controls a Yamaha Disklavier MIDI piano.

In our setup, the first pedal is used to adjust the ratio between the deterministic and the stochastic parts of the source signal. This makes it possible to make a sound with little or no noise, or to remove the deterministic part in order to obtain a filtered noise.

The second pedal controls the nonlinear function. In fact, the nonlinear function has been split into an odd and an even function, which makes it possible to adjust the amount of odd and even harmonics. A continuous morphing can then be obtained between a pan flute–like sound and a clarinet-like sound. The real flute sound corresponds to a median position.

Figure 11. Transfer function of the loop filter in the case of a flute-like sound (solid) and a string-like sound (dashed). The dotted curves show various transfer functions of the loop filter obtained by jointly adjusting parameters G and a.



The third pedal controls the noise produced by the key pad and is triggered when a key is pressed on the instrument. The level of the impulse noise is controlled by the closing speed of the key pad detected by the flute interface. This effect makes it possible to play the flute without blowing into it.

The fourth pedal allows the user to change the characteristics of the resonator of the instrument while the flute is being played. This means that we must design a filter allowing for large changes in the transfer function via few control parameters. We made a filter allowing continuous "morphing" between a flute-like and a string-like sound. By adjusting the filter's transfer function with a control pedal, interesting effects can be obtained. To obtain the filter, we first had to estimate the two transfer functions. These transfer functions, as shown in Figure 11, were obtained from measurements on a flute and a guitar string. Roughly speaking, they can be said to be low-pass filters with different cut-off frequencies and slopes. This suggests the use of a classical one pole filter with a z-transfer function as follows:

$$H(z) = G\frac{1-a}{1-az^{-1}}. \qquad (4)$$

A filter of this kind yields reasonable approximations of the two above mentioned transfer functions. In this way we can obtain interpolations between these transfer functions by jointly adjusting parameters G and a, as shown in Figure 11.

In this application, the flute and Disklavier can interact in several ways. By adding a software layer to the MIDI interface, one can for instance use pressure variations from the flautist to control the speed at which the piano strings are struck, or control the tempo of a prerecorded sequence played by the piano.

Conclusion

In this article, we have described a new interface that was adapted to a traditional flute. The goal in designing this interface was to give flautists access to the world of digital sounds without obliging them to change their traditional playing techniques. A synthesis model adapted to this interface was designed which makes it possible to resynthesize and transform the original flute sound. Because this synthesis model is a mixture between a signal model and a physical model, we have called it a hybrid model. Thanks to the physical model, we are now able to give the instrument a new set of physical characteristics, and thus give a physical interpretation of the sound transformation. This means that we can, for example, simulate the sound of a gigantic flute, or replace the resonator of the flute by the resonator of another instrument. As far as the signal model is concerned, its advantage is that it can simulate a sound no matter how complicated the physical processes underlying the source are. This means that when the physical behavior of a sound source is not fully understood, or when the physical equations which describe the source are too complicated to be implemented in real time, signal models can be used. These models generally yield a satisfactory sound resynthesis, and they enable one to perform sound transformations. In addition, this flute is MIDI-compatible, which means that it can be used to control other MIDI instruments, and can be played in any tuning system by assigning it arbitrary frequency values for a given key state.

This instrument has been presented twice to composers and musicians, and we are currently working with a composer to make a piece where the flute controls a MIDI piano. Traditional flute players who are interested in contemporary music have also given very positive feedback to this instrument, especially because they can use traditional playing techniques.

Acknowledgments

This work has been done in the research group Modeling, Synthesis, and Control of Sound and Musical Signals at the Laboratoire de Mécanique et d'Acoustique in Marseille. The authors would like to thank Richard Kronland-Martinet and Philippe Guillemain for fruitful discussions and collaboration.

This work was supported in part by the Norwegian Research Council.

References

Arfib, D. 1979. "Digital Synthesis of Complex Spectra by Means of Multiplication of Non-linear Distorted Sine Waves." *Journal of the Audio Engineering Society* 27:757–768.

Beauchamp, J. W. 1982. "Synthesis by Spectral Amplitude and `Brightness' Matching of Analyzed Musical Instrument Tones." *Journal of the Audio Engineering Society* 30(6):396–406.

Chaigne, A. 1995. "Trends and Challenges in Physical Modeling of Musical Instruments." Paper presented at the 1995 International Congress on Acoustics, 26–30 June, Trondheim, Norway.

Coltman, J. W. 1968. "Sounding Mechanism of the Flute and Organ Pipe." *Journal of the Acoustical Society of America* 44(1):983–992.

Cook, P. R. 1992. "A Meta-Wind-Instrument Physical Model Controller and a Meta-Controller for Real-Time Performance Control." *Proceedings of the 1992 International Computer Music Conference*. San Francisco: International Computer Music Association, pp. 273–276.

De Laubier, S. 1998. "The Meta-Instrument." *Computer Music Journal* 22(1):25–29.

Fabre, B., A. Hirschberg, and P. J. Wijmands. 1996. "Vortex Shedding in Steady Oscillation of a Flue Organ Pipe." *Acta Acustica* 82(6):863–877.

Fletcher, N. H. 1975. "Acoustical Correlates of Flute Performance Technique." *Journal of the Acoustical Society of America* 57(1):233–237.

Fletcher, N. H., and T. D. Rossing. 1990. *The Physics of Musical Instruments.* Berlin: Springer-Verlag.

Guillemain, P., and R. Kronland-Martinet. 1996. "Characterisation of Acoustics Signals Through Continuous Linear Time-frequency Representations." *Proceedings of the IEEE* 84(4):561–585.

Karjalainen, M., et al. 1991. "Transmission-Line Modeling and Real-Time Synthesis of String and Wind Instruments." *Proceedings of the 1991 International Computer Music Conference*. San Francisco: International Computer Music Association, pp. 293–296.

Kergomard, J. 1981. "Champ interne et champ externe des instruments à vent. " Doctoral dissertation, Université Paris IV.

Kronland-Martinet, R. 1988. "The Wavelet Transform for Analysis, Synthesis, and Processing of Speech and Music Sounds." *Computer Music Journal* 12(4):11–20.

Kronland-Martinet, R., and A. Grossmann. 1991. "Application of Time-Frequency and Time-Scale Methods (Wavelet Transforms) to the Analysis, Synthesis, and Transformation of Natural Sounds. " In G. De Poli, A. Piccialli, and C. Roads, eds. *Representations of Musical Signals*. Cambridge, Massachusetts: MIT Press, pp. 45–85.

Kronland-Martinet, R., T. Voinier, and P. Guillemain. 1997. "Agir sur le Son Musical avec la Baguette-Radio." In H. Genevois and R. de Vivo, eds. *Les Nouveaux Gestes de la Musique*. Marseille: Editions Parenthèses, pp. 181–193.

Le Brun, M. 1979. "Digital Waveshaping Synthesis." *Journal of the Audio Engineering Society* 27:250–266.

Mathews, M. V. 1991a. "The Conductor Program and Mechanical Baton." In M. V. Mathews and J. R. Pierce, eds. *Current Directions in Computer Music Research*. Cambridge, Massachusetts: MIT Press, pp. 263–281.

Matthews, M. V. 1991b. "The Radio-Baton and Conductor Program, or: Pitch, the Most Important and Least Expressive Part of Music. " *Computer Music Journal* 15(4):37–46.

McNally, G. W. 1984. "Dynamic Range Control of Digital Audio Signals." *Journal of the Audio Engineering Society* 32(5):316.

Nederveen, C. J. 1998. *Acoustical Aspects of Woodwind Instruments.* Dekalb, Illinois: Northern Illinois University Press.

Pollard, H.F., and E. V. Jansson. 1982. "A Tristimulus Method for the Specification of Musical Timbre." *Acustica* 51:162–171.

Puckette, M. S., T. Apel, and D. Zicarelli. 1998. "Real-time Audio Analysis Tools for Pd and MSP." *Proceedings of the 1998 International Computer Music Conference*. San Francisco: International Computer Music Association, pp. 109–112.

Serra, X. 1989. "A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic plus Stochastic Decomposition." PhD Thesis, Stanford University.

Smith, J.O. 1992. "Physical Modeling Using Digital Waveguides." *Computer Music Journal* 16(4):74–91.

Stevens, S. S. 1972. "Perceived Level of Noise by Mark VII and Decibels (E). " *Journal of the Acoustical Society of America* 51:575.

Verge, M.P. 1995. "Aeroacoustics of Confined Jets with Applications to the Physical Modeling of Recorder-Like Instruments." PhD thesis, Eindhoven University.

Widrow, B., and S. D. Stearns. 1985. *Adaptive Signal Processing.* Englewood Cliffs, New Jersey: Prentice-Hall.

Ystad, S. 1998. "Sound Modeling Using a Combination of Physical and Signal Models." PhD Thesis, University of Aix-Marseille II.

Ystad, S. 2000. "Sound Modeling Applied to Flute Sounds. " *Journal of the Audio Engineering Society*, 48(9):810–825.

Sound Modeling Applied to Flute Sounds*

SØLVI YSTAD, *AES Member*

CNRS, Laboratoire de Mécanique et d'Acoustique, 13402 Marseille cedex 20, France

An original way of modeling musical sounds by a combination of physical and signal models is presented. To take into account the physical characteristics of the instrument, a physical model is designed. A technique for extracting parameters for synthesis purposes is then proposed. To generate sustained sounds, a source signal to be injected into the physical model has been constructed. This source signal, modeled using signal models, consists of independent, deterministic, and stochastic parts. A new way of fitting the parameters of the signal model by perceptual criteria such as the tristimulus criterion is given. Finally examples are given of how sounds from the model can be manipulated. These manipulations can be implemented in real time and piloted by an interface mirroring a real flute.

0 INTRODUCTION

Sound modeling is a part of a more general concept of analysis–synthesis, which consists of constructing a synthetic sound from a natural sound [1]. The first step in this process is to design analysis methods that give representations of the real sounds. If these representations are invertible we can construct a synthetic sound identical to the real sound. This corresponds to the horizontal path in the center of Fig. 1. However, by following this path the resulting sound is identical to the original. It would therefore be more interesting to be able to modify the representation through transformations on the original sound. This can be done by following the upper path of the diagram [2], [3].

This engineering report is mainly related to the lower part of the diagram, corresponding to sound modeling by constructed synthesis models. Synthesis models can be divided into two main groups: signal models and physical models. Signal models reproduce a perceived effect through mathematical relations, whereas physical modeling gives a physical description of the sound-generating system. Synthesis models can use parameters extracted from the analysis of real sounds. Physical models can in addition use theoretical data. By acting on the parameters of the synthesis model, one can make transformations on the sound.

In this work we constructed a sound model by combining physical and signal models into a hybrid model. Such a model can take advantage of the positive aspects of both methods. This engineering report proposes a new technique based on the analysis of transient sounds to construct the physical model. The signal model makes use of perceptual criteria such as the tristimulus criterion, which is well adapted to finding the parameters corresponding to nonlinear synthesis methods.

1 ANALYSIS OF SOUNDS

When modeling sounds, one has to use or construct analysis methods that give a good representation of the sound. This is important for the extraction of synthesis parameters. Since sounds are generally nonstationary, and since the time course of a sound is important from a perceptual point of view, time–frequency representations should be used. For this purpose we used linear representations such as the Gabor and the wavelet transforms, which decompose the signal into elementary functions. The Gabor transform can be written as

$$S(\omega, \tau) = \int_{-\infty}^{\infty} s(t)\overline{W}(t - \tau)\mathrm{e}^{-\mathrm{i}\omega(t-\tau)}\,\mathrm{d}t$$

where the elementary functions are fixed windows translated in time and in frequency. The wavelet transform can be written as

$$S(a, \tau) = \frac{1}{\sqrt{a}}\int s(t)\overline{g}\left(\frac{t - \tau}{a}\right)\mathrm{d}t$$

where the elementary functions are wavelet functions obtained by translation in time and dilation of a basic function $g(t)$. These analysis methods give time–frequency representations such as in Fig. 2, where the

* Manuscript received 1999 February 16; revised 1999 December 12 and 2000 April 28.

upper part corresponds to the modulus and the lower part to the phase of the representation of a synthetic musical sound [4].

To extract parameters from this representation, one should use special methods such as the spectral-line estimation method developed by Guillemain [5]. This method constructs special filters that adapt with time, making it possible to extract one spectral component and reject the others. One of these filters is illustrated in Fig. 3. If the signal is analytic, one can associate one amplitude and one frequency modulation law with each spectral component [6]. (Henceforce we will use the term envelope when talking about modulation laws.) As an example, Figs. 4 and 5 show the amplitude and frequency envelopes of the third component of a flute sound.

The spectral component varies considerably with time. This time course is very important from a perceptual point of view [7]. We will also describe another method we designed for extracting the envelopes, called the matched analysis method, which was designed to optimize parameter extraction for transient sounds.

2 SYNTHESIS BY SIGNAL MODELING

Signal models use a purely mathematical description of sounds. They are numerically easy to implement, and they guarantee a close relation between the synthesis parameters and the resulting sound. These methods are similar to shaping and edification of structures from materials, and the three principal groups can be classified as follows:

- Additive synthesis
- Subtractive synthesis
- Global (or nonlinear) synthesis.

As a first approach to synthesizing flute sounds we

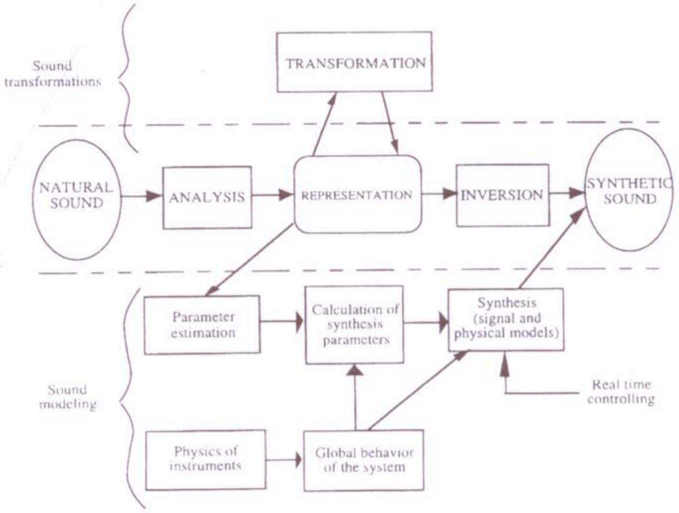


Fig. 1. General organization of analysis–synthesis and modeling concept.

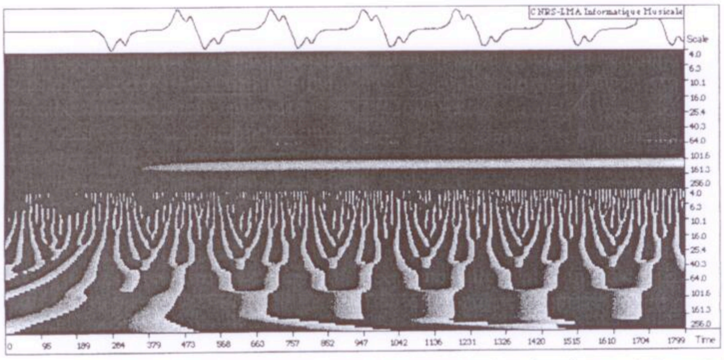


Fig. 2. Time–frequency representation of synthetic musical sound. Horizontal axis—time (in samples); vertical axis—frequency.

constructed an additive synthesis model by considering the behavior of the amplitude and frequency envelopes of the attack of a flute sound. This method shows how the parameters from the analysis can be used directly to construct a synthesis model.

Additive synthesis uses the amplitude and frequency envelopes directly extracted from the sound. It can be written as

$$s(t) = \sum_k A_k(t) \cos\left[\omega_k t + \int_0^t \nu_k(u)\,\mathrm{d}u\right]$$

where $A_k(t)$ and $\nu_k(t)$ represent the amplitude and frequency envelopes of the kth component. Additive synthesis consists of constructing a signal by adding up its amplitude- and frequency-modulated components. If we can find a relation between each amplitude envelope and a relation between each frequency envelope such that $A_k(t)$ can be written as $\alpha_k A(t)$ and $\nu_k(t)$ can be written as $kV_0(t)$, we can construct a group additive synthesis model [8] based on the formula

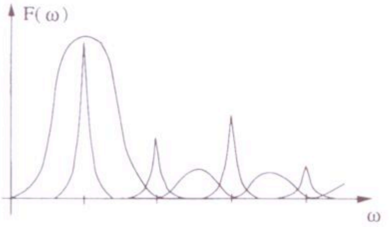


Fig. 3. Adaptive filter for extracting spectral lines.

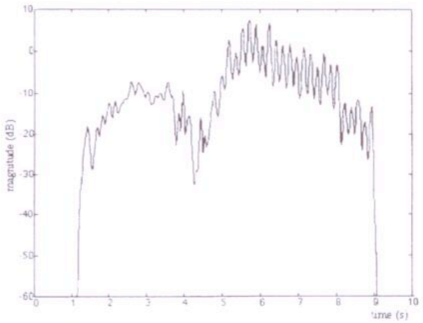


Fig. 4. Amplitude envelope of 3rd component of A1 (f_0 = 440 Hz).

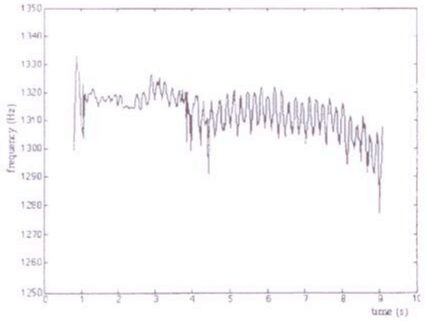


Fig. 5. Frequency envelope of 3rd component of A1 (f_0 = 440 Hz).

$$s(t) = A(t) \sum_k \alpha_k \cos\left[\omega_k t + k \int_0^t \nu(u)\,\mathrm{d}u\right] .$$

To model the attack of a flute sound by such a model, we studied the amplitude envelopes corresponding to the beginning (about 0.5 s) of flute sounds. An example of these envelopes is given in Fig. 6, where the first five components of a C4 flute sound are plotted together with the amplitude envelope of the total sound. The components are not generated at the same time, which causes a delay between them. In addition the slopes of the amplitude envelopes are steeper for higher harmonics.

To construct the group additive synthesis model, each amplitude envelope must be expressed as a function of the amplitude envelope of the whole signal. Fig. 7 shows the amplitude envelope of the first component versus the total amplitude envelope. Fitting a straight line to the curve (a first-order approximation) gives an empirical formula providing an expression for each amplitude envelope A_k as a function of the amplitude envelope A_T of the total pressure,

$$A_k = K_k + [0.75 + 0.3(k - 1)]A_T .$$

The constant K_k is related to the delays between the components (Fig. 6). The frequency fluctuations of the sound $\nu(t)$ can be found by bandpass filtering the driving pressure. We finally propose a group additive synthesis model corresponding to a flute sound given by the formula

$$s(n) = \sum_{k=1}^{K} A_k(n) \cos\left[k\omega_0 \frac{n}{\omega_s} + k\varphi(n)\right]$$

where

$$\varphi(n) = \varphi(n - 1) + \nu(n) .$$

The synthesized sound constructed by this model is rather close to a real flute sound, but the noise due to the air jet interactions with the flute body [9] is missing. Moreover, piloting such a model with an interface requires complicated rules taking into account the relations between the note played and the behavior of the components. To take into account physical aspects of the sound-generating system, a physical approach should be constructed.

3 SYNTHESIS BY PHYSICAL MODELING

Physical models simulate the behavior of existing or virtual sound sources. Such models can be constructed either from the equations describing the behavior of the waves propagating in the structure and their radiation in air, or from the behavior of the solution of the same equations. Here we chose to base our research on the latter alternative by using the waveguide synthesis models [10]. These models have the advantage of being easy to construct and of having a behavior close to that of a real sound generator. The waveguide model can model different kinds of systems such as strings and tubes. The tube case has been used here to describe how to construct such a model.

3.1 Propagation of Longitudinal Waves in Resonant Tubes

The wave equation of the acoustic pressure y inside the tube when viscothermal losses are taken into account can be written [11]

$$\frac{1}{c^2}\frac{\partial^2 y(x,r,t)}{\partial t^2} - \Delta y(x,r,t) - \frac{l_{\mathrm{hv}}}{c}\Delta\frac{\partial y(x,r,t)}{\partial t} = \Psi(x,r,t)$$

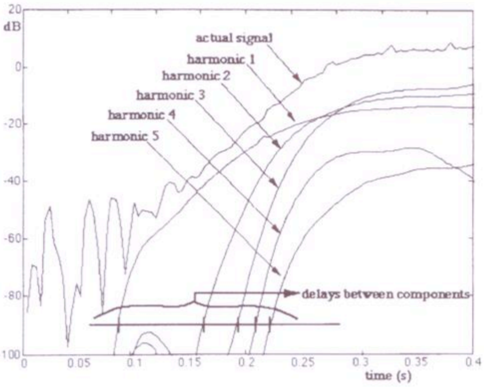


Fig. 6. First 0.4 s of amplitude envelopes of first five harmonics of a flute sound whose fundamental frequency is 261 Hz (note C1).

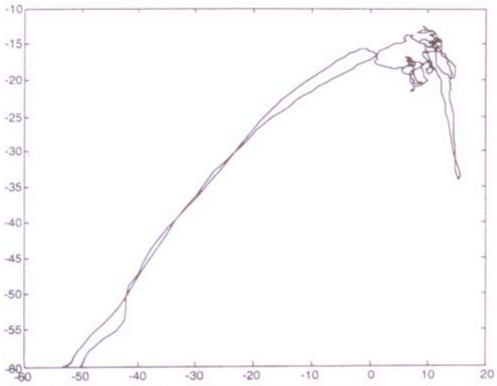


Fig. 7. Energy density of first component versus energy density of whole sound.

with boundary conditions $y(0, r, t) = y(L, r, t) = 0$. Here l_{hv} represents the characteristic length (l_{hv} is on the order 10^{-8} m in free air-filled space), c is the velocity of sound, and Ψ is a time-dependent source. If we suppose that only a plane wave propagates in the tube, we find that the response at point x (to the one-dimensional equation) of an impulse at point x_0 is

$$y(x,t) = \frac{2}{L}\sum_{n=1}^{\infty}\frac{\sin(n\pi(x_0/L))\sin(n\pi(x/L))}{(1/c^2)\omega_n}\sin(\omega_n t)e^{-\alpha_n t}.$$

The response is a sum of sinusoidal functions exponentially damped. The damping factors are given by

$$\alpha_n = l_{\mathrm{hv}}c\frac{n^2\pi^2}{2L^2}$$

and they depend on the square of the mode rank. The frequencies are given by the relation

$$\omega_n^2 = c^2\frac{n^2\pi^2}{L^2} - l_{\mathrm{hv}}^2c^2\frac{n^4\pi^4}{4L^4}$$

where the first term corresponds to the perfectly harmonic case and the second term to an inharmonicity term related to the dispersion in the medium.

Figs. 8 and 9 show the theoretical damping factor and

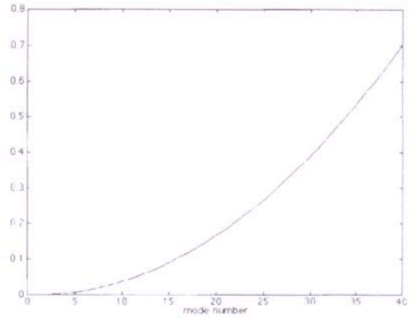


Fig. 8. Theoretical damping factors as a function of mode rank corresponding to a tube of length $L = 58$ cm.

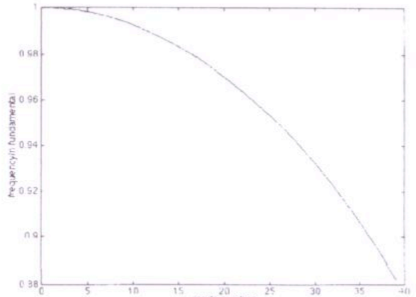


Fig. 9. Theoretical inharmonicity of modes in a tube.

the inharmonicity as a function of the mode rank. The damping factor increases with the mode rank, which means that higher order modes are more rapidly attenuated than lower order modes. Fig. 9 shows that the inharmonicity is nonnegligible even in the tube case. This behavior has been verified and validated by experimental measurements [12] (see also Section 3.3).

3.2 Construction of a Waveguide Model

To make a model that takes into account these phenomena, a propagative model was constructed which is a generalization of the waveguide model first proposed by Karplus and Strong [13]. If we consider a tube or a string and cut it into elementary cells, we can look at the behavior of the waves in each elementary cell during propagation in the medium. When the medium is excited, waves propagate in both directions. When the waves propagate from one cell to another, they take a certain time, which can be simulated by an elementary delay z^{-1}. In addition they are attenuated and dispersed depending on the frequency. This can be simulated by elementary attenuation and dispersion filters. In this way we can construct propagative lines as shown in Fig. 10.

Since the system is linear, the elementary delays can be grouped into a single delay line D, and the elementary filters can be grouped into a single filter, as shown in Fig. 11. The filters R take into account the boundary conditions of the finite medium. This model can be simplified further by grouping the delay lines into a single delay line and the filters into a single filter taking into account dispersion, attenuation, and boundary conditions (Fig. 12).

Propagation in strings and tubes has been simulated by this model. In analysis–synthesis the problem to solve is the construction of a loop filter taking into account physical phenomena found from the theoretical equations. To design such a filter, we compared the transfer function of the waveguide model with the theoretical response of the system. To do that, consider the power spectral density of the model's transfer function

$$S(\omega) = \frac{1}{1 + |F(\omega)|^2 - 2|F(\omega)| \cos(\omega d - \phi)}$$

and the power spectral density of the response of the theoretical equation, which is a sum of Lorentzian functions,

$$S_{\mathrm{Th}}(\omega) = \sum_n \frac{C_n^2}{\alpha_n^2 + (\omega - \omega_n)^2} .$$

Comparing the power spectral densities at the resonance peaks yielded an expression for the discrete values of the loop filter at the resonance peaks $\omega = \omega_n$,

$$|F(\omega_n)| = \frac{2 + d^2\alpha_n^2 \pm \sqrt{(2 + \alpha_n^2 d^2)^2 - 4}}{2}$$

as well as the corresponding phase,

$$\phi(\omega_n) = \omega_n d - 2n\pi .$$

It is important to know the behavior of the loop filter for different kinds of resonators so as to control sound transformations such as the continuous morphing of a resonant tube into a string. To get an idea of what the loop filter looks like for different instruments, the discrete loop filter values of the tube and the string are plotted in Figs. 13 and 14. The calculus corresponding to the string is given in [14].

The modes in the tube are much more rapidly attenuated than those in the string, and the response of a string contains many more significant modes than does the

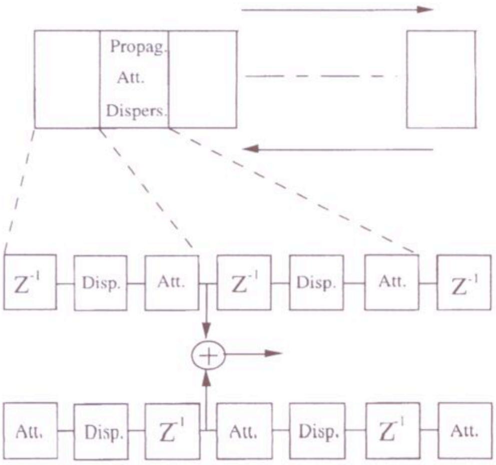


Fig. 10. Construction of propagative lines with elementary filters.

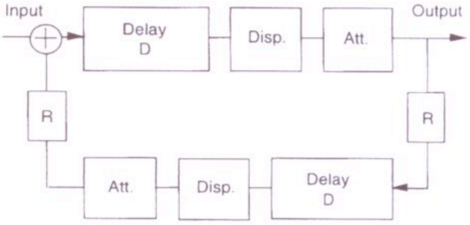


Fig. 11. Grouping elements for simplification of model.

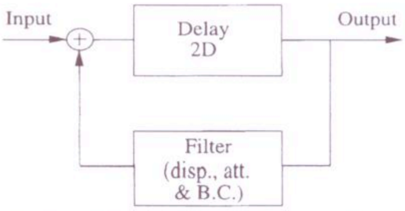


Fig. 12. Waveguide model with one filter and one delay line.

response of a tube. These findings correspond to what we expect, since the spectrum of the sound from a string is much richer than that of a tube.

To take into account dispersion and dissipation in a temporal scheme, the impulse response of the filter should be constructed. Although algorithms already exist which take into account dispersion and dissipation [15], they are based on approximations and are not precise enough for resynthesis where the reconstructed sound should closely resemble the original sound. We therefore designed a new method for constructing an impulse response based on time–frequency representations of a transient sound. Fig. 15 shows the time–frequency representation of a transient sound propagating in a piano string. This example is used to describe the method since the dispersion phenomena in a piano string are more visible than for the flute. Nevertheless the method acts in the same way on flute sounds. From this representation one can see that the energy of the signal is localized along curves called the ridges of the transform [16]. These ridges are related to the group delay, which is given by the relation

$$t_k = T_1 - kT_k = \frac{1}{f_1} - \frac{k}{f_k}$$

and which is due to the fact that the propagation speed depends on the frequency. The theory states that one can reconstruct the signal by summing up elementary grains along the group delay, as shown in Fig. 16 [17]. The impulse response is then given by

$$f(t) = \sum_{k=1}^{K} \{\alpha_k \cos[\omega_k(t - t_k)] + \beta_k \sin[\omega_k(t - t_k)]\} \, \mathrm{e}^{-(t-t_k)^2/\sigma^2} .$$

To find the coefficients α_k and β_k corresponding to the amplitudes of the elementary grains, we chose Gaussian

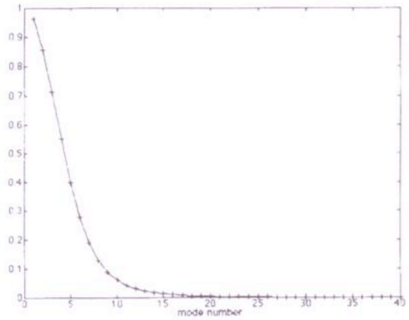


Fig. 13. Modulus of filter F in tube case.

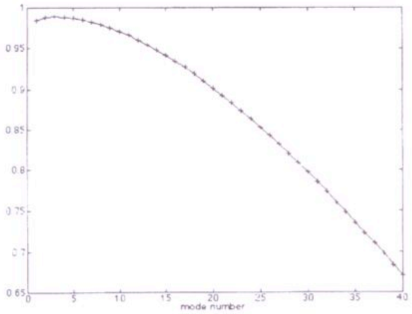


Fig. 14. Modulus of filter F in string case.

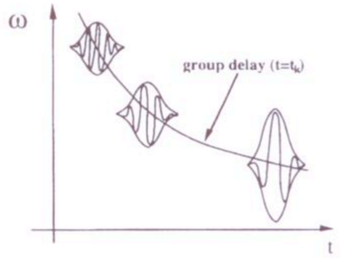


Fig. 16. Elementary grains along group delay.

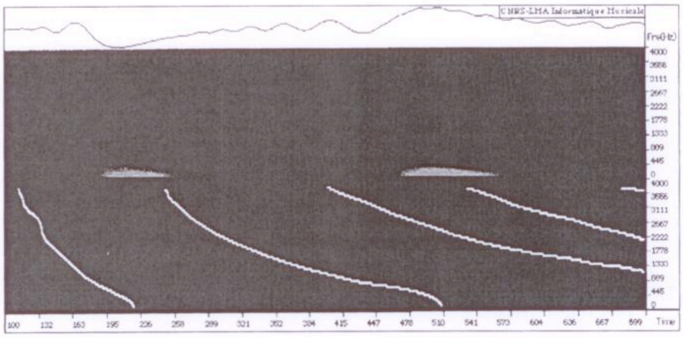


Fig. 15. Modulus and ridge of Gabor representation of first 66 ms of speed at one point of a piano string measured by laser vibrometry. White curves correspond to ridges related to group delay of wave. One can see that deformation increases with time.

functions as elementary grains. Since Gaussian functions are invariant by Fourier transform, the amplitudes we are searching for correspond to the amplitudes of the Gaussian functions in the frequency domain (Fig. 17). Thus by summing up the functions, a continuous curve can be constructed which coincides exactly with the discrete filter values found by comparing the power spectral densities.

In the frequency domain the constructed impulse response can now be written

$$F(\omega) = \frac{\sigma\sqrt{2\pi}}{2} \sum_{k=1}^{K} \{\alpha_k M(\omega) - \mathrm{i}\beta_k M(\omega)\}\, \mathrm{e}^{-\mathrm{i}\omega t_k}$$

where

$$M(\omega) = \mathrm{e}^{-(1/2)(\omega-\omega_k)^2\sigma^2} + \mathrm{e}^{-(1/2)(\omega+\omega_k)^2\sigma^2} .$$

This means that we can find coefficients α_k and β_k by solving a linear system [12].

This original method makes it possible to construct filters that have the exact values of dissipation and dispersion at the resonance peaks. Figs. 18 and 19 show the impulse response of the loop filter corresponding to the tube and string models, respectively. As expected, the dispersion is much larger in the string than in the tube, where the impulse response is almost symmetric. Also, in the string the higher order modes propagate faster than the lower order modes.

3.3 Feeding the Waveguide Model with Data from Real Sounds

This section addresses the problem of feeding the waveguide model in order to resynthesize sounds. Parameters must be extracted from natural sounds to be fed into the propagative model. To do that, we designed an analysis method called matched analysis, which uses special analysis functions adapted to the signal to be analyzed. As for the flute model, the bandwidth of the spectral components depends on the square of the frequency. This means that by choosing analysis functions that have the same properties as the analyzed signal, we can precisely extract the parameters from the signal. The matched analysis can be written

$$M(\tau, \omega) = \int s(t) W[(t - \tau)\alpha(\omega)]\, \mathrm{e}^{\mathrm{i}\omega(t-\tau)}\, \mathrm{d}t .$$

Applied to the flute, the analysis functions can be written $W[(t - \tau)K\omega^2]$, giving

$$M(\tau, \omega) = \int s(t) W[(t - \tau)K\omega^2]\, \mathrm{e}^{\mathrm{i}\omega(t-\tau)}\, \mathrm{d}t .$$

The filters corresponding to a matched time–frequency analysis are shown in Fig. 20. The bandwidth of the functions is exaggerated for illustrative purposes. The matched analysis technique applied to a transient flute sound gave very good results. Fig. 21 shows the amplitude envelopes of the first and sixth components of the flute sound produced by rapidly closing a fingerhole (without exciting at the embouchure).

The sixth component is much more rapidly attenuated than the first, which corresponds to the theory. The real damping factors can now be found by measuring the slope of the logarithm of the amplitude envelope. This makes it possible to compare the real and theoretical dissipation and dispersion, as shown in Figs. 22 and 23.

To fit the real and theoretical cases, we can adjust the parameters in the theoretical equation. In the tube the curves have been fitted by adjusting the value of the parameter l_{hv} in the theoretical equation. The theoretical curves plotted here correspond to a characteristic length $l_{\mathrm{hv}} = 10^{-3}$ m.

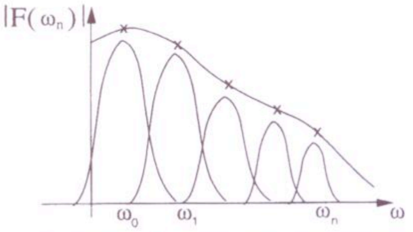


Fig. 17. Construction of a continuous filter response by summing up Gaussian functions.

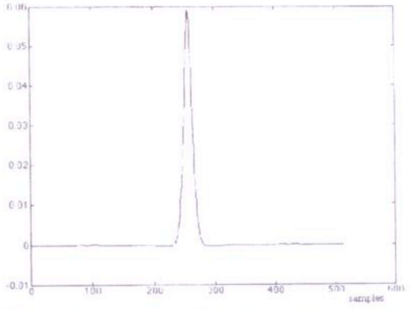


Fig. 18. Impulse response of filter F in waveguide model using theoretical values for damping coefficients and eigenfrequencies corresponding to tube case (f_s = 32 000 Hz).

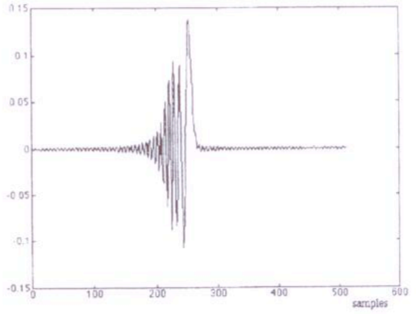


Fig. 19. Impulse response of filter F in waveguide model using theoretical values for damping coefficients and eigenfrequencies corresponding to string case (f_s = 32 000 Hz).

The resynthesis of transient sounds obtained by the model is very convincing. However, this model is insufficient to simulate sustained sounds such as those from wind instruments. In this case we chose to model the source separately before injecting it into the waveguide model. In the flute the source corresponds to an air jet fluctuating around a sharp edge called the labium and interacting with the standing waves in the resonator. The physical phenomena related to the source were described by authors such as Verge [18] and Fabre et al. [19]. Although their work provided a good physical understanding of the sound-generating system, their physical models are too complicated to be implemented in real time. Authors like Cook [20] propose the construction of a nonlinear waveguide to simulate the interaction of the source and the tube. Nevertheless, these algorithms do not allow the resynthesis of a given natural sound. Since we wanted to develop a digital instrument sounding like a given real one, to be manipulated and piloted in real time by simulating the perceptive effect of the source, we used a signal model for the source.

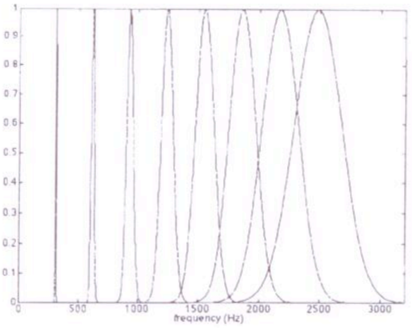
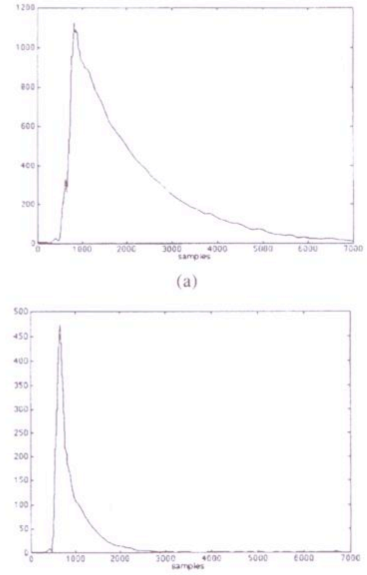


Fig. 20. Filters corresponding to matched time–frequency analysis. Relation between α and ω is based on mode behavior in a tube: $\alpha = K\omega^2$, where K was chosen to be larger than in reality for illustrative purposes.



(a)

(b)

Fig. 21. Amplitude modulation laws. (a) First component of transient sound. (b) Sixth component of transient sound. (f_s = 48 000 Hz.)

4 SYNTHESIS BY HYBRID MODELING

Modeling the source of a flute sound requires extracting its contribution from the rest of the signal. We started by assuming that the source and the resonator could be separated. Although this is not correct from a physical point of view, we shall see that it is a good assumption when a perceptual effect is to be reconstructed. The corresponding source signal was extracted by a deconvolution method.

4.1 Source Identification by Deconvolution

As seen in the previous section, a physical model corresponding to the resonator of the instrument can be

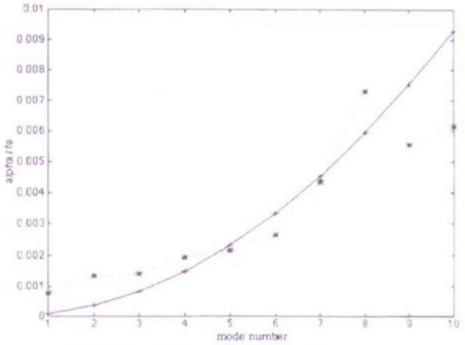
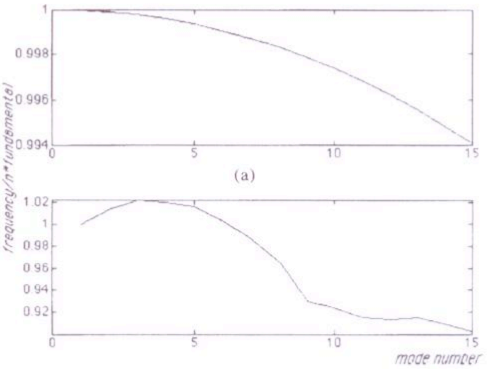


Fig. 22. Damping factors in real (*) and theoretical (+) cases.



(a)

(b)

Fig. 23. $f_n/n f_0$ for a tube of length L = 0.58 m. (a) Theoretical case. (b) Real case.

constructed by a filter and a delay, as shown in Fig. 24. The transfer function of the resonant system is given by

$$H(\omega) = \frac{1}{1 - F(\omega)\, e^{-i\omega d}}$$

and corresponds to an all-pole filter. This means that its inverse exists, and that the deconvolution between the real flute sound and the resonator is a justifiable mathematical operation. The source signal $x(t)$ is then given by

$$x(t) = (y * h^{-1})(t) .$$

Fig. 25 shows the transfer function of the resonator corresponding to a flute and its inverse. The deconvolution between this resonator and the real flute sound gives the source, represented by its spectrum in Fig. 26.

The source consists of two main contributions, a stochastic part corresponding to the noise in the signal and a deterministic part that is the sum of spectral components. To model the source signal we therefore split it into a deterministic and a stochastic part, as proposed in Serra [21], and modeled them independently.

4.2 Splitting Deterministic and Stochastic Components of the Source Signal

To split the two contributions, we used for the first time in this field a so-called LMS (least mean square) algorithm [22], which consists of using a special adaptive filter to remove from an input signal all the components correlated with a reference signal, as shown in Fig. 27. If the input signal is the source signal and if the reference signal is an estimate of the deterministic signal, the output will be the stochastic signal. The adaptive filter W_k can be updated by the formula

$$\boldsymbol{W}_{k+1} = \boldsymbol{W}_k - \mu \bar{\nabla}_k$$

where the gradient estimate $\bar{\nabla}_k$ is given by

$$\bar{\nabla}_k = \frac{\partial e_k^2}{\partial w_k} = 2e_k \frac{\partial e_k}{\partial w_k} = -2e_k s_k$$

The deterministic estimate can be obtained using the matched analysis.

Now that the source is split into a deterministic and a stochastic part, we will model these parts independently, starting with the deterministic part.

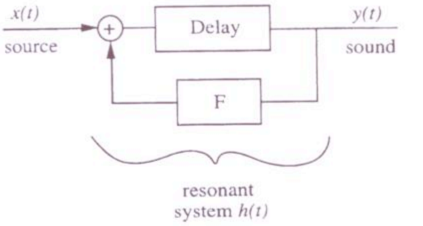
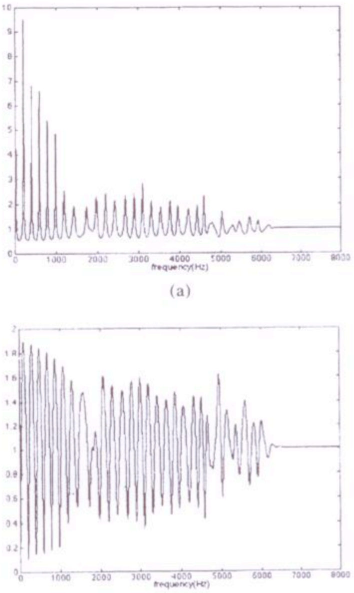


Fig. 24. Filtering of a source by a resonant system.



(a)

(b)

Fig. 25. (a) Transfer function $H(\omega)$ and (b) inverse transfer function $H^{-1}(\omega)$ of resonance model corresponding to a flute.

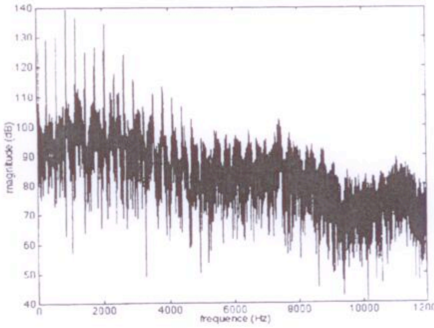


Fig. 26. Spectrum of deconvoluted signal, flute sound D1.

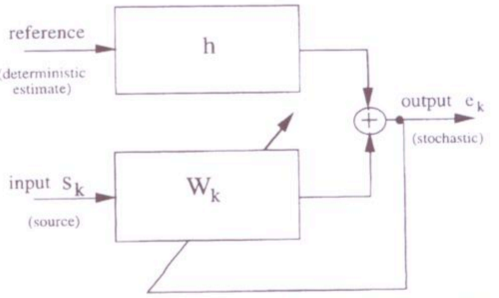


Fig. 27. Illustration of the principle of the LMS algorithm used to separate deterministic signal and noise.

4.3 Modeling the Deterministic Part of the Source Signal

For different dynamic levels the source spectra do not evolve in a linear way, since the increase in the dynamic level does not correspond to a global amplification of the spectrum. This can be seen by comparing the source spectra of a pianissimo and a fortissimo sound (Figs. 28 and 29). The first four components of the spectra change considerably, and the third component becomes very strong in the fortissimo case. This nonlinear behavior of the source is common to many musical instruments. It is of great importance from a perceptual point of view since it means that the timbre of the flute sound changes when the dynamic level of the sound changes.

To model this nonlinear behavior we used a waveshaping synthesis model developed by Arfib [23] and Le Brun [24]. It consists of constructing a signal by using a nonlinear function γ, the argument of which is a monochromatic signal (a simple cosine) with amplitude $I(t)$, called the index of distortion. This index has a large influence on the spectrum of the signal. The signal obtained by the waveshaping method is given by

$$s(t) = \gamma[I(t)\cos(\omega_0 t)]$$

with

$$\gamma(x) = \sum_{k=1}^{K} \alpha_k T_k(x) .$$

Waveshaping synthesis makes it possible to generate a desired spectrum for a given index. It is convenient to decompose the nonlinear function into Chebyshev polynomials, since this gives a simple relation between the function and the generated signal. Actually, when the index of distortion equals 1, the coefficients of the decomposition of γ in the Chebyshev basis are given by the values of the modulus of the spectrum to be generated,

$$\gamma(\cos\omega_0 t) = \sum_{k=0}^{\infty} \alpha_k \cos(k\omega_0 t) .$$

Fig. 30 shows the waveshaping function for a flute sound.

At this point the challenge is to find how to vary the waveshaping index in order to get an evolution of the synthetic spectrum, which corresponds to the evolution of the real spectrum for different dynamic levels. Since waveshaping synthesis does not make it possible to model spectral evolution, we propose a new method, which makes use of perceptive criteria, to find the variation range for the index.

4.3.1 Spectral Centroid Criterion

The first perceptive criterion we tested, probably the most well known, was the spectral centroid criterion given by

$$C = \frac{\int_0^{\infty} \omega |h(\omega)| \, d\omega}{\int_0^{\infty} |h(\omega)| \, d\omega}$$

which was proposed by Beauchamp [25] and which is directly related to the brightness of the sound. Fig. 31 shows the spectral centroid corresponding to different dynamic levels of the flute sound C2 as a function of the logarithm of the energy. The centroid varied from 2.8 to 3.7, a rather small range of variation. Fig. 32 shows the spectral centroid (corresponding to a synthetic sound) as a function of the waveshaping index. Here the spectral centroid varied from 0 to 3.7 as the waveshaping index varied from 0 to 1. To make the synthetic centroid vary the same way as the real centroid, the waveshaping

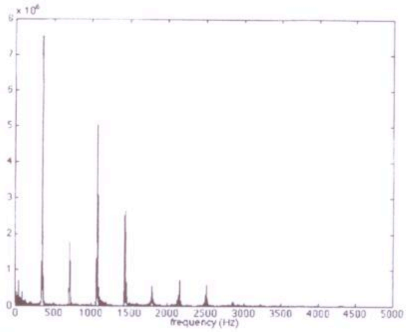


Fig. 28. Spectrum of a pianissimo played flute sound G1 (f = 392 Hz).

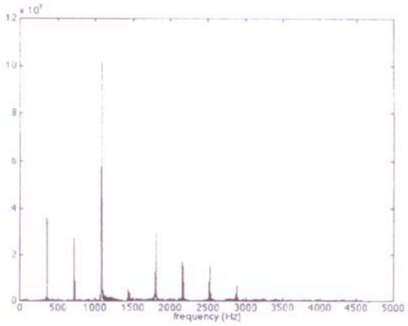


Fig. 29. Spectrum of a fortissimo played flute sound G1 (f = 392 Hz).

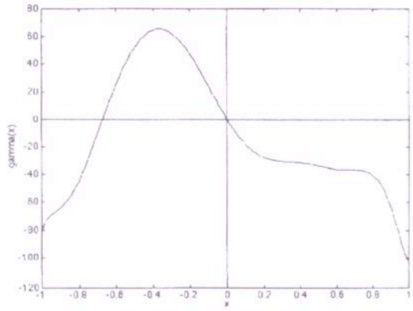


Fig. 30. Waveshaping function of a flute sound.

index should vary from 0.97 to 1. This is a very small range of variation, but it does not matter if this variation range gives the desired spectral evolution. Unfortunately for flute sounds, such spectral evolution cannot be obtained, which means that the spectral centroid criterion does not work on these sounds. In fact, the spectral centroid criterion is suitable for sounds whose spectral components increase globally and not for sounds whose spectrum changes dramatically.

4.3.2 Tristimulus Criterion

Another perceptual criterion, more adapted to the flute case, is the tristimulus criterion proposed by Pollard and Jansson [26]. This criterion consists of cutting the total loudness of the sound into three contributions: one takes into account the fundamental component, another the second to fourth components, and the third the rest of the components. The total loudness can then be written as a sum of these three contributions:

$$N = N_1 + N_2^4 + N_5^n$$

Each contribution can be calculated using Stevens' equation [27],

$$N_i^n = 0.85 N_{\max} + 0.15 \sum_i^n N_i \,.$$

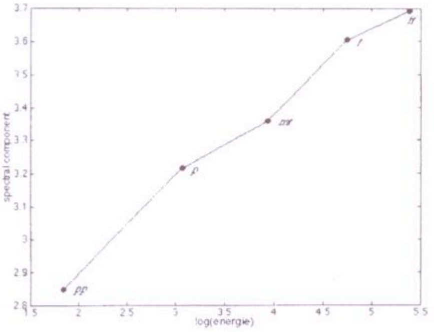


Fig. 31. Spectral centroid of C2 (fundamental frequency 523 Hz) as a function of the logarithm of energy.

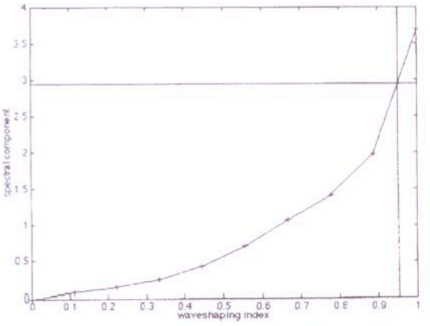


Fig. 32. Spectral centroid of waveshaping function for values of waveshaping index from 0 to 1.

The tristimulus is then given by the three normalized contributions so that their sum is 1:

$$x = \frac{N_5^n}{N}\,, \qquad y = \frac{N_2^4}{N}\,, \qquad z = \frac{N_1}{N}\,.$$

This means that the tristimulus can be plotted in a two-dimensional diagram where the x axis (corresponding to the normalized high-frequency contribution) is the abscissa and the y axis (corresponding to the mid-frequenty partials) is the ordinate, and where the fundamental contribution z is implicit (Fig. 33). Thus when the values of x and y are small, the fundamental frequencies of the sound are strong; when the value of x is small and y is strong, the midfrequency components are strong, and so on. The tristimulus corresponding to a flute sound is shown in Fig. 34.

Here the lower curve corresponds to the tristimulus of the real sound plotted for different dynamic levels of the flute sound, whereas the upper curve corresponds to the synthetic tristimulus when the index of distortion varies from 0 to 1. It is important to be aware of the fact that the tristimulus is a perceptual criterion, and

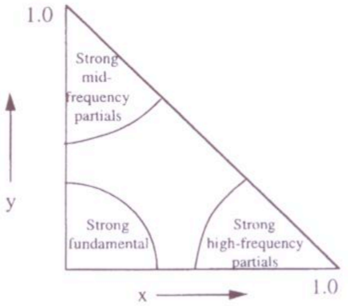


Fig. 33. Acoustic tristimulus diagram.

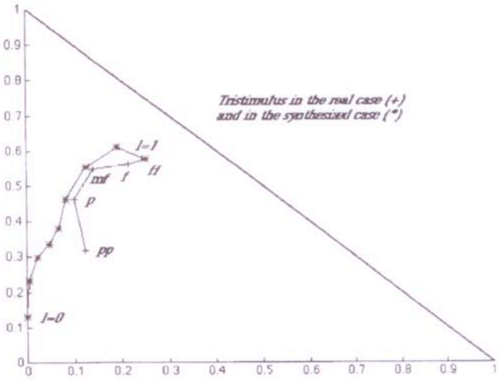


Fig. 34. Tristimulus diagram of sounds generated by a waveshaping function with index values from 0 to 1 and of five flute sounds (fundamental frequency 523 Hz) with different dynamic levels (pp to ff).

only its global behavior is important. Therefore we do not expect the real and the synthetic sounds to show exactly the same behavior. What is important is that the trajectories be located in the same area of the tristimulus diagram, and that they have the same global evolution, which is the case for flute sounds. By fitting the two curves, we then find a correspondence between the index and the driving pressure, as shown in Fig. 35. The index varied quite linearly with the logarithm of the driving pressure. As an example, for a driving pressure corresponding to a pianissimo sound the index of distortion should be 0.5.

4.4 Modeling the Stochastic Part of the Source Signal

Now that we have a model for the deterministic part of the source signal, the stochastic part should be modeled. Thanks to the extraction of the resonator contribution, the noisy part of the source signal becomes quite simple to model with the help of a smooth response filter. We start by assuming that the process is stationary and ergodic. The time–frequency representation of the stochastic part of the flute source (Fig. 36) shows that this is a rather good assumption since the transform is almost time invariant.

This means that we can characterize the noise by its probability density function and its power spectral density. Measurements showed that in the flute the probability density function follows an exponential law,

$$f_e(x) = \frac{\lambda}{2e^{-\lambda|x|}} .$$

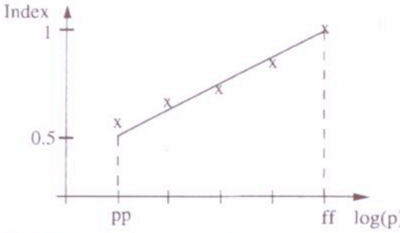


Fig. 35. Relation between waveshaping index and driving pressure.

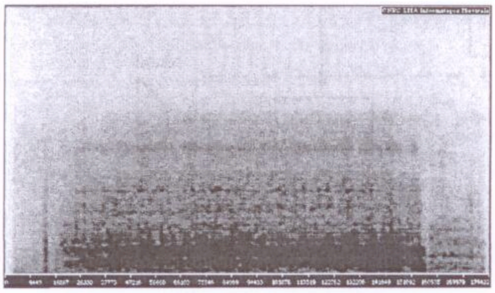


Fig. 36. Spectrogram of a flute noise.

The function is shown in Fig. 37. The power spectral density corresponds to a low-pass filtered noise with three bumps (Fig. 38). The third bump probably corresponds to the so-called edge tone of the flute sound.

Now that we have a model for the stochastic part of the source signal, we can combine it with the deterministic model to make a signal model for the source. Then by combining this signal model with the physical model we can construct a general hybrid model that can be applied to several instruments, and in particular to the flute.

5 CONTROL OF THE HYBRID FLUTE MODEL

The hybrid model corresponding to the flute is presented in Fig. 39. The physical model with its loop filter and its delay line simulates the resonator of the instrument. Its input signal is the signal model simulating the source of the instrument. The model can be piloted by both the driving pressure and the played frequency, which is given by the finger position when a specific flute interface is used. The source injected into the resonator is obtained by a noise-generation system, which represents its stochastic contribution, and by a nonlinear system, which represents its deterministic part. The input of the deterministic part of the source is a sine generator. Its amplitude input corresponds to the

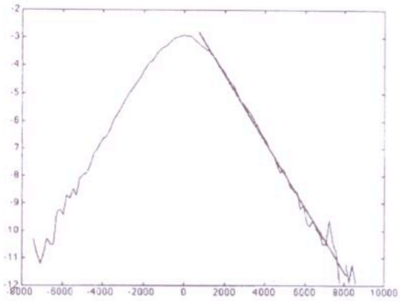


Fig. 37. Natural logarithm of estimated probability density function corresponding to the stochastic part of the source signal of a flute. Superposed straight line allows the estimation of λ.

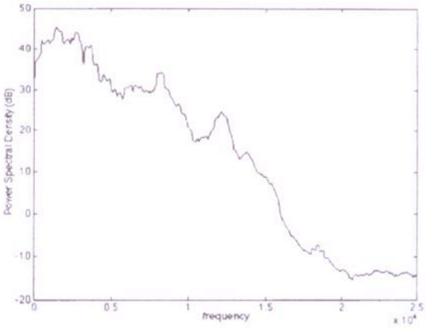


Fig. 38. Power spectral density of stochastic part of the source of a flute noise.

index of distortion given by the logarithm of the driving pressure, and its frequency input is given by the finger position together with a vibrato extracted by bandpass filtering the driving pressure. The noise from the key pads that is obtained by rapidly closing a fingerhole is added to the input of the resonator since it is an important part of the flute sound.

This general model gives good resynthesized sounds and has been implemented in real time. In addition, a suitable interface should be made to pilot the model. We considered several interfaces. One possibility was to use the radio baton designed by Matthews, which makes it possible to associate the synthesis parameters with the position of two batons in space [28]. By moving the batons one can then modify the synthesis parameters in order to transform the sound. This is a good tool for sculpting the sound. However, if we want to play as a flutist plays the instrument, we need to control parameters such as the driving pressure. Because such a parameter is almost impossible to modify in a natural way by using a baton, another interface mirroring a real flute was designed. This interface is shown in Fig. 40. It is a real flute with sensors connected to the keypads [29].

The sensors make it possible to detect the finger position and thus the note played. In addition a microphone is placed at the embouchure level to measure the driving pressure. Fig. 41 shows how the information from the sensors can be transformed and sent to the real-time processor. In this case the information is transformed into MIDI codes through a MIDI coding processor. MIDI (musical instrument digital interface) is the specification for a set of digital codes that transmit music control and timing information in real time. MIDI is also a specification for the hardware interface through which the codes are transmitted [30].

The information can be manipulated by the MAX program [31], which can link the instrument to other MIDI systems. MAX can also be used to alter the sound by modifying the synthesis parameters.

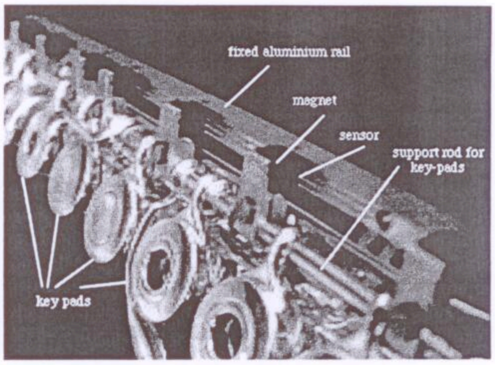


Fig. 40. Close view of a flute with magnets and sensors.

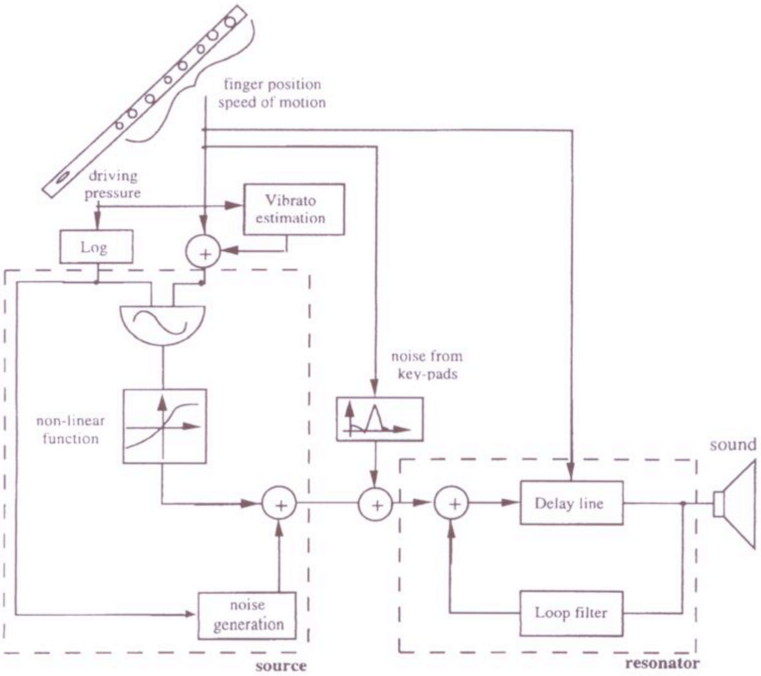


Fig. 39. Hybrid flute model.

As was already mentioned, sound modeling consists not only of resynthesizing sounds but also of transforming the sound without the mechanical constraints of the instrument. This is the most interesting part of the digital instrument. Thus to transform sounds by this instrument, one should modify the parameters of the model. Fig. 42 shows some modifications. We can, for example, modify the vibrato of the sound to remove, alter, or amplify it. By modifying the index of distortion we change the timbre of the sound. Modifications on the waveshaping function make it possible to play clarinet sounds with the flute interface. We can modify the noise to change its characteristics or modify the deterministic or stochastic levels. We can also modify the characteristics of the noise from the keypads, or amplify it, or we can add an external source to the input of the resonator to sing into it. By modifying the loop filter, we can also change the characteristics of the resonator and make a "flutar" corresponding to blowing into a string with a flute embouchure. Even more curious sounds can be made by combining the different transformations we proposed.

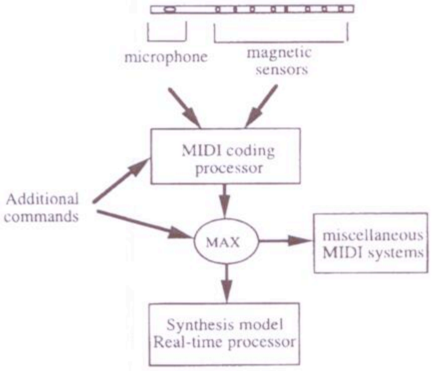


Fig. 41. Synoptic of connection of a digital interface of flute type.

6 CONCLUSION

This study shows how to design a sound model by using a combination of signal and physical models. The physical models take into account the most relevant physical characteristics of the sound-generating system, whereas the signal models take into account perceptual effects. Also, a way of designing physical models that will simulate the instrument resonator has been described. This model takes into account both dispersion and dissipation. Such effects are important from a perceptual point of view. Signal models were used to model the source of the instrument, which was extracted from the sound by a deconvolution method. We further pro-

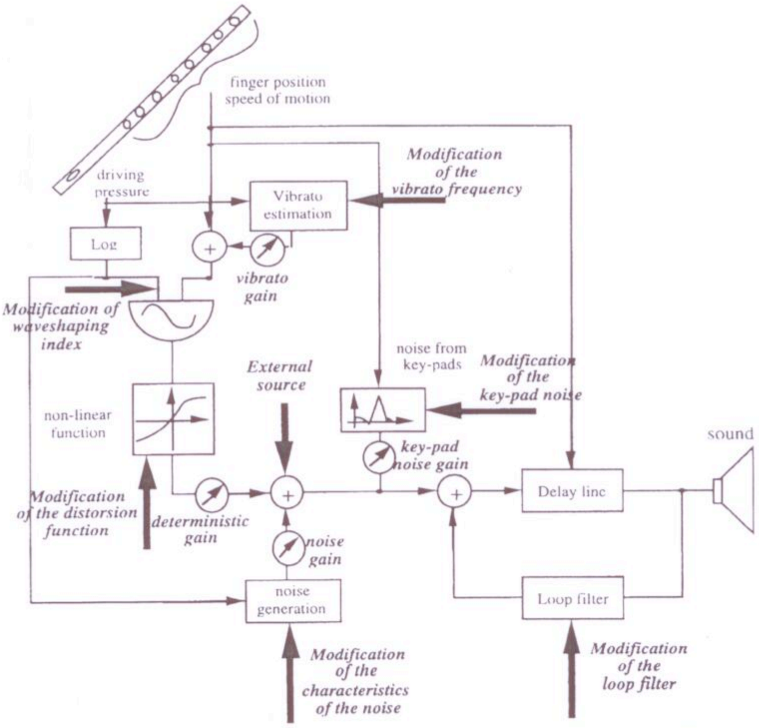


Fig. 42. Control of model parameters. Bold italic labels show possibilities for modification.

posed to split the source signal into a deterministic and a stochastic part by the LMS algorithm and to model these contributions independently. The deterministic part was modeled by waveshaping synthesis to take into account the nonlinearities of the source signal. Perceptual criteria were then applied to find the synthesis parameters. The stochastic part was easy to implement by separating source and resonator. It can be modeled by linear filtering of white noise. In fact the stochastic part of the flute sound is colored because the noise propagates in the resonator. As mentioned in the beginning, sound modeling consists of both resynthesis and transformation of natural sounds. We have therefore shown how sounds can be manipulated by the proposed model and how these manipulations can be done in real time and piloted by an interface we designed.

In the future it would also be of interest to model the coupling between the source and the resonator more precisely. This would correspond more closely to the behavior of a real instrument. By coupling we mean reinjecting the output of the resonator into the input of the source. Although this may seem rather trivial, an important problem is related to this coupling: since the nonlinear source element in this case will be inside the loop, the system can become unstable and even chaotic. The present study gives several clues on how to address this problem.

7 ACKNOWLEDGMENT

This work was supported in part by the Norwegian Research Council. The author would like to thank Philippe Guillemain, Richard Kronland-Martinet, and Thierry Voinier for fruitful discussions and collaboration. She also thanks the reviewers for their constructive comments.

8 REFERENCES

[1] R. Kronland-Martinet, P. Guillemain, and S. Ystad, "Modelling of the Natural Sounds Using Time–Frequency and Wavelet Representations," *Organised Sound* (Cambridge University Press), vol. 2, no. 3, pp. 179–191 (1997).

[2] R. Kronland-Martinet and A. Grossmann, "Application of Time-Frequency and Time-Scale Methods to the Analysis, Synthesis and Transformation of Natural Sounds," in *Representations of Musical Signals*, C. Roads, G. De Poli, and A. Picciali, Eds. (MIT Press, Cambridge, MA, 1990).

[3] D. Arfib and N. Delprat, "Musical Transformations Using the Modification of Time–Frequency Images," *Comput. Music J.*, vol. 17, no. 2, pp. 66–72 (1993).

[4] R. Kronland-Martinet, J. Morlet, and A. Grossman, "Analysis of Sound Patterns through Wavelet Transforms," *Int. J. Pattern Recognit. Artifi. Intelligence*, vol. 11, no. 2, pp. 97–126 (1987).

[5] P. Guillemain, "Analyse et modelisation de signaux sonores par des représentations temps–fréquence linéaires," PhD thesis, Université Aix-Marseille II, France (1994 June).

[6] B. Picinbono, *Signaux Aléatoires*, vols. 1, 2, 3 (Dunot, Paris, 1993).

[7] J. C. Risset and D. Wessel, "Exploration of Timbre by Analysis and Synthesis," in *The Psychology of Music*, 2nd ed. (Academic Press, 1999), pp. 113–169.

[8] A. Horner and L. Ayers, "Modeling Acoustic Wind Instruments with Contiguous Group Synthesis," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 46, pp. 868–879 (1998 Oct.).

[9] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments* (Springer, New York, 1990).

[10] J. O. Smith, "Physical Modeling Using Digital Waveguides," *Comput. Music J.*, vol. 16, no. 4, pp. 74–91 (1992).

[11] J. Kergomard, "Champ interne et champ externe des instruments à vent," Thèse d'Etat, Université Paris IV, France (1981).

[12] S. Ystad, "Sound Modeling Using a Combination of Physical and Signal Models," PhD thesis, Université Aix-Marseille II, France (1998 Mar.).

[13] K. Karplus and A. Strong, "Digital Synthesis of Plucked String and Drum Timbres," *Comput. Music J.*, vol. 2, no. 7, pp. 43–55 (1983).

[14] P. Guillemain, R. Kronland-Martinet, and S. Ystad, "Physical Modelling Based on the Analysis of Natural Sounds," in *Proc. ISMA* (Int. Symp. on Musical Acoustics), vol. 19, pt. 5 (University of Edinburgh, Scotland, 1997 Aug.), pp. 445–450.

[15] B. Yegnanarayana, "Design of Recursive Group-Delay Filters by Autoregressive Modeling," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, pp. 632–637 (1982).

[16] P. Guillemain and R. Kronland-Martinet, "Characterization of Acoustics Signals through Continuous Linear Time–Frequency Representations," *Proc. IEEE (Special Issue on Wavelets)*, vol. 84, pp. 561–585 (1996).

[17] N. Delprat, B. Escudié, P. Guillemain, R. Kronland-Martinet, P. Tchamitchian, and B. Torrésani, "Asymptotic Wavelet and Gabor Analysis: Extraction of Instantaneous Frequencies," *IEEE Trans. Inform. Theory*, vol. 38, pp. 644–664 (1992 Mar.).

[18] M. P. Verge, "Aeroacoustics of Confined Jets with Applications to the Physical Modeling of Recorder-Like Instruments," PhD thesis, Eindhoven University, Eindhoven, The Netherlands (1995).

[19] B. Fabre, A. Hirschberg, and P. J. Wijnands, "Vortex Shedding in Steady Oscillation of a Flue Organ Pipe," *Acta Acoustica*, vol. 82, pp. 863–877 (1996).

[20] P. R. Cook, "A Meta-Wind-Instrument Physical Model Controller and a Meta Controller for Real-Time Performance Control," in *Proc. 1992 Int. Computer Music Conf.* (San Francisco, CA, 1992), pp. 273–276.

[21] X. Serra, "A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic plus Stochastic Decomposition," PhD thesis, Stanford University, Stanford, CA (1989 Oct.).

[22] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. (Prentice-Hall, Englewood Cliffs, NJ, 1985).

[23] D. Arfib, "Digital Synthesis of Complex Spectra by Means of Multiplication of Nonlinear Distorted Sine Waves," *J. Audio Eng. Soc.*, vol. 27, pp. 757–768 (1979 Oct.).

[24] M. Le Brun, "Digital Waveshaping Synthesis," *J. Audio Eng. Soc.*, vol. 27, pp. 250–266 (1979 Apr.).

[25] J. W. Beauchamp, "Synthesis by Spectral Amplitude and 'Brightness' Matching of Analyzed Musical Instrument Tones," *J. Audio Eng. Soc.*, vol. 30, pp. 396–406 (1982 June).

[26] H. F. Pollard and E. V. Jansson, "A Tristimulus Method for the Specification of Musical Timbre," *Acoustica*, vol. 51 (1982).

[27] S. S. Stevens, "Perceived Level of Noise by Mark VII and Decibels (E)," *J. Acoust. Soc. Am.*, vol. 51, p. 575 (1972).

[28] M. V. Mathews and C. Abbot, "The Sequential Drum," *Comput. Music J.*, vol. 4, no. 4 (1980).

[29] S. Ystad and T. Voinier, "Design of a Flute Interface to Control Synthesis Models," in *Proc. 1999 Int. Computer Music Conf.* (Beijing, China, 1999), pp. 228–232.

[30] B. Moog, "MIDI: Musical Instrument Digital Interface," *J. Audio Eng. Soc. (Features)*, vol. 34, pp. 394–404 (1986 May).

[31] M. Puckette and D. Zicarelli, "MAX—An Interactive Graphic Programming Environment," Opcode Systems (Palo Alto, CA, 1985).

THE AUTHOR

Sølvi Ystad was born in Bø, Telemark, Norway, in 1968. She graduated from Norges Tekniske Høgskole (NTH) in Trondheim as an electronics engineer in 1992 and in 1998 obtained a Ph.D. degree from the University of Aix-Marseille II, France, and from Norges Teknisk Naturvitenskapelige Universitet (NTNU), Norway, for her work on "Sound Modeling Using a Combination of Physical and Signal Models." Dr. Ystad is currently working in the research group Modeling, Synthesis and Control of Sound and Musical Signals at the Laboratoire de Mécanique et d'Acoustique, National Center for Scientific Research in Marseille, France, where her research activity mainly concerns modeling, analysis, and synthesis of audio sounds with particular emphasis on musical applications. Dr. Ystad is a flute player and has also studied composition in electroacoustic and contemporary music.

Modelling of natural sounds by time–frequency and wavelet representations*

R. KRONLAND-MARTINET, PH. GUILLEMAIN and S. YSTAD

CNRS, Laboratoire de Mécanique et d'Acoustique, 31 Chemin Joseph Aiguier, 13402 Marseille cedex 20, France
E-mail: kronland@lma.cnrs-mrs.fr guillem@lma.cnrs-mrs.fr ystad@lma.cnrs-mrs.fr

Sound modelling is an important part of the analysis–synthesis process since it combines sound processing and algorithmic synthesis within the same formalism. Its aim is to make sound simulators by synthesis methods based on signal models or physical models, the parameters of which are directly extracted from the analysis of natural sounds. In this article the successive steps for making such systems are described. These are numerical synthesis and sound generation methods, analysis of natural sounds, particularly time–frequency and time–scale (wavelet) representations, extraction of pertinent parameters, and the determination of the correspondence between these parameters and those corresponding to the synthesis models. Additive synthesis, nonlinear synthesis, and waveguide synthesis are discussed.

1. THE SOUND MODELLING CONCEPT

Analysis–synthesis is a set of procedures to reconstruct a given natural sound and collect information about it. Different methods can be applied, and the success of each method depends on its adaptive possibilities and the sound effect to be produced. Figure 1 shows the most commonly used procedures. The three parts of the figure correspond to different processes. The central level corresponds to a direct analysis–synthesis process and consists of reconstructing a sound signal by inversion of the analysis procedure. This is a useful process which uses analysis to get information about a sound, and synthesis (inversion) to verify that no information is lost. The analysis will make it possible to classify and characterise audio signals (Kronland-Martinet, Morlet and Grossman 1987), but the result of the process will simply be a reproduction of the natural sound. From a musical point of view, a representation of sounds by analysis is useful when one intimates sound modifications. This sound transformation process corresponds to the upper path in figure 1 and consists of altering the coefficients of the representation between the analysis and the synthesis procedures. According to the analysis method used, different aspects of the sound can be altered. The energy distribution and/or the frequency of the partials can, for example, be manipulated through spectral analysis. Time–frequency analysis allows the separation of the time and frequency characteristics associated with the sound and is of great interest (Kronland-Martinet 1988). However, this approach conflicts with a very important mathematical principle which states that one cannot arbitrarily modify a time–frequency representation of a signal. This constraint is due to the existence of the so-called 'reproducing kernel' which takes into account the redundancy of such representations (Kronland-Martinet *et al.* 1987). It corresponds to the uncertainty principle which states that one cannot be as precise as one wishes in the localisation of both the time and the frequency domains. This principle imposes the concept that the time–frequency domain corresponding to the uncertainties (time–frequency atoms) be considered instead of isolated values of the representation. It is then natural to make the transformations act on these domains in order to conserve the necessary correlations between close representation values. This is done by using a mathematical operation known as a projection, which can transform any image into a time–frequency representation. The constraints limit the time–frequency transformation processes and make it difficult to determine the correspondence between the altered values and the obtained sounds. Nevertheless, very interesting sounds can be obtained by carefully using such altering procedures (Arfib and Delprat 1993). In this article we will pay special attention to the lower part of figure 1 which corresponds to sound modelling. In this part, the representations obtained from the analysis provide parameters corresponding to the synthesis models. The concept of the algorithmic sampler (Arfib, Guillemain and Kronland-Martinet 1992) consists of simulating natural sounds through a synthesis process that is well adapted to algorithmic and realtime manipulations. The resynthesis and the transformation of natural sounds are then part of the same concept.

The paper is organised as follows. We describe the most commonly used synthesis methods, analysis methods such as time–frequency and wavelet transforms, and the algorithms used for separating and

* This work has been partly supported by the Norwegian Research Council.

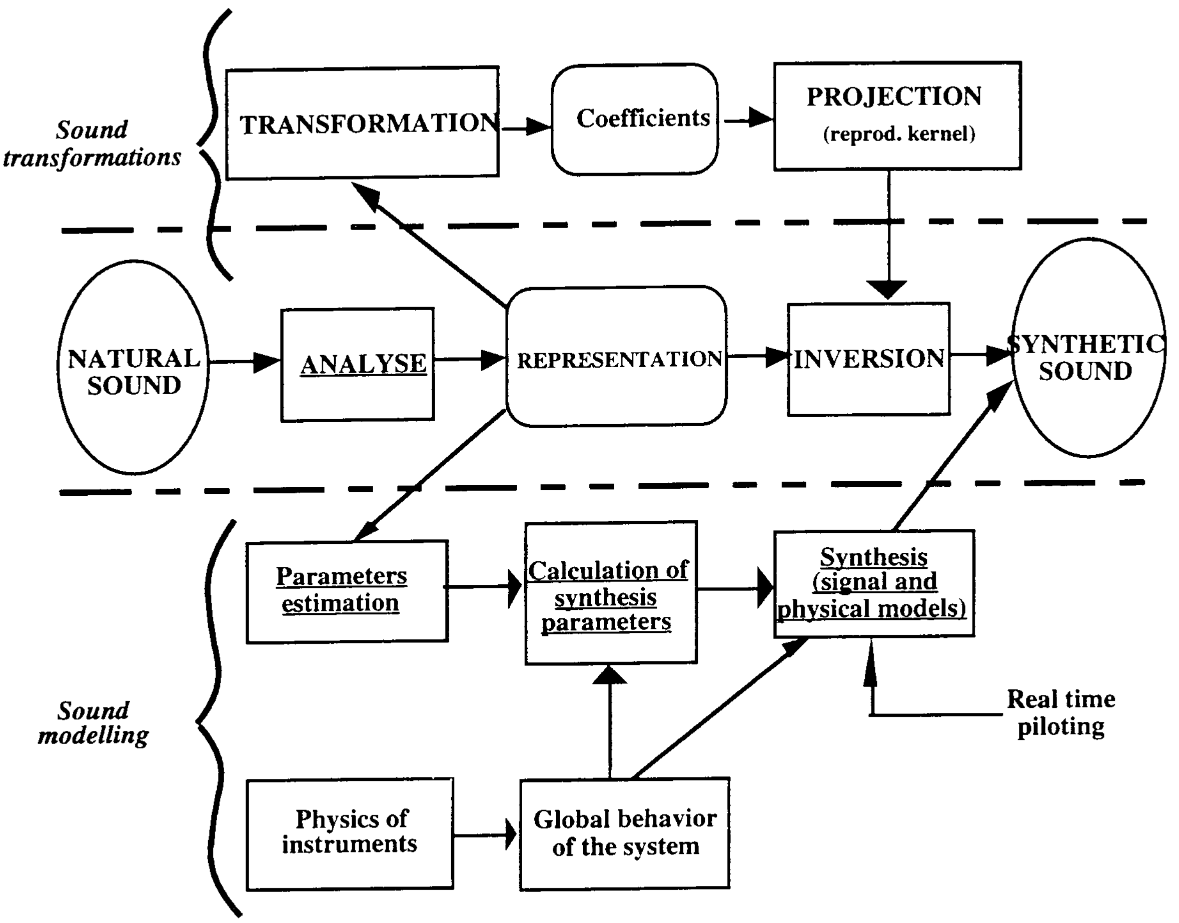


Figure 1. General organisation of the analysis–synthesis and modelling concept. Each underlined item corresponds to a section in the text.

characterising spectral components. We conclude by showing how the analysis of real sounds can be used to estimate the synthesis parameters of the signal models and of the physical models. Most of these techniques have been developed in our laboratory in Marseille, France.

2. SOUND SYNTHESIS MODELS

Digital synthesis uses methods of signal generation that can be divided into two classes:

- signal models aimed at reconstructing a perceptive effect without being concerned with the specific source that made the sound,
- physical models aimed at simulating the behaviour of existing or virtual sound sources.

2.1. Signal model synthesis

Signal models use a purely mathematical description of sounds. They are numerically easy to implement, and they guarantee a close relation between the synthesis parameters and the resulting sound. These methods are similar to shaping and building structures from materials, and the three principal groups can be classified as follows:

- additive synthesis,
- subtractive synthesis,
- global (or nonlinear) synthesis.

2.1.1. Additive synthesis

A complex sound can be constructed as a superposition of elementary sounds, generally sinusoidal signals modulated in amplitude and frequency (Risset 1965). For periodic or quasi-periodic sounds, these components have average frequencies that are multiples of one fundamental frequency and are called harmonics. The periodic structure leads to electronic organ sounds if one does not consider the microvariations that can be found through the amplitude and frequency modulation laws of the components of any real sound. These dynamic laws must therefore be very precise when one reproduces a real sound. The advantage of these synthesis methods is the potential for intimate and dynamic modifications of the sound.

Granular synthesis can be considered as a special kind of additive synthesis, since it also consists in summing elementary signals (grains) localised in both the time and the frequency domains (Roads 1978).

2.1.2. Subtractive synthesis

A sound can be constructed by removing undesired components from an initial, complex sound such as noise. This synthesis technique is closely linked to the theory of digital filtering (Rabiner and Gold 1975) and can be related to some physical sound generation systems such as speech (Flanagan, Coker, Rabiner, Schafer and Umeda 1970, Atal and Hanauer 1971). The advantage of this approach (excluding the physical aspects of physical modelling synthesis, discussed later) is the possibility of uncoupling the excitation source and the resonance system. The sound transformations related to these methods often use this property to make hybrid sounds or crossed synthesis of two different sounds by combining the excitation source of a sound and the resonant system of another (Makhoul 1975, Kronland-Martinet 1989). A well-known example of cross-synthesis is the sound of a talking 'cello obtained by associating an excitation of a 'cello string and a resonance system corresponding to the time-varying formants of the vocal tract.

2.1.3. Global synthesis

Simple and 'inert' signals can be dynamically modelled using global synthesis models. This method is nonlinear since the operations on the signals are not simple additions and amplifications. The most well-known example of global synthesis is audio frequency modulation (FM) updated by John Chowning (Chowning 1973) which revolutionised commercial synthesizers. The advantages of this method are that it calls for very few parameters, and that a small number of operations can generate complex spectra. These simplify numerical implementation and control. However, it is difficult to control the shaping of a sound by this method, since the timbre is related to the synthesis parameters in a nonlinear way and continuous modification of these parameters may give discontinuities in the sound. Other related methods have proved to be efficient for signal synthesis, such as waveshaping techniques (Arfib 1979, Le Brun 1979).

2.2. Synthesis by physical modelling

This is a more recent technique, which we will describe more precisely than signal model synthesis. Unlike signal models which use a purely mathematical description of sounds, physical models describe the sound generation system with respect to its physical behaviour. Such models can be constructed either from the equations describing the behaviour of the waves propagating in the structure and their radiation in air, or from the behaviour of the solution of these equations. The first approach is costly in terms of calculations and is generally used only in connection with research work (Chaigne 1995), unless one uses a simplified version consisting of modelling the structure by an association of simple elements (Cadoz, Luciani and Florens 1984) (masses, springs, dampers,...). Synthesis by simulation of the solution of the propagation equation has led to waveguide synthesis models (Smith 1992), which have the advantages of being easy to construct and of having a behaviour close to that of a real instrument. Such synthesis methods are, consequently, well adapted to the modelling of acoustical instruments.

We describe below the principles of these methods in order to reveal their parameters, together with their correlation to physical mechanisms. These parameters are related to the structure of the instrument as well as to the instrumental performance. If we consider a vibrating string, the Shannon theorem states that one can, without loss of information, split the movement into a succession of instantaneous cliches separated by an interval of time T called the sampling period. If c is the propagation speed of the waves in the string, this is equivalent to cutting the string into intervals of length $x = cT$ and considering the propagation as a passage from one elementary cell to another. This operation corresponds to a spatial 'discretisation' of the structure: one can then consider the wave propagation as the result of a succession of transformations or filterings of the initial excitation.

In the ideal case in which we neglect losses and nonlinearities, there is only a displacement of the waves (in two directions), and the result can thus be simulated by a succession of delay lines corresponding to the sampling period T, symbolised in digital signal processing by the variable z^{-1}. In the more realistic case in which the waves undergo an attenuation depending on the frequency, a filter P should be added between each delay. If in addition the medium is dispersive, a 'dephasor' or an all-pass filter D should be added (figure 2).

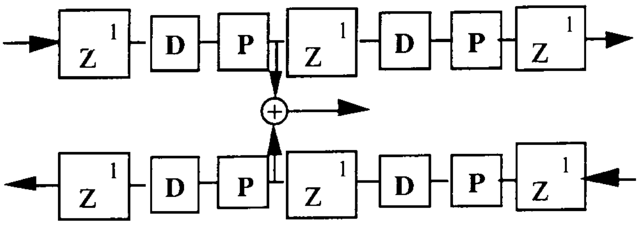


Figure 2. Discrete simulation of the wave propagation in a dissipative and dispersive medium.

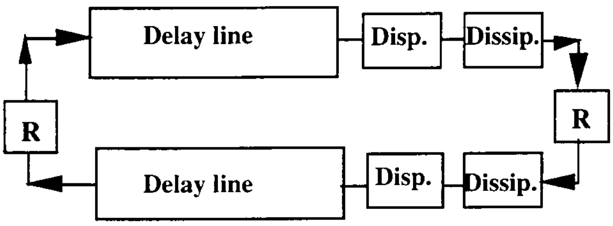


Figure 3. Propagation model in a bounded dissipative and dispersive medium.

The theory of digital filters allows elements of the same type to be gathered. Thus, the propagation medium can be represented by a succession of 3 elements, i.e. a delay line, an attenuating filter for simulating the dissipation, and an all-pass filter for simulating the dispersion. Real instruments have strings of finite length and the waves propagated through it are reflected at the ends. The reflections correspond to a return of the initial waves, with modifications depending on the boundary conditions. Thus one can simulate the wave behaviour corresponding to the solution of the equations. For that purpose, one uses a looped system which, in addition to the delay line, attenuating filter and the all-pass filter, also makes use of a filter corresponding to the reflections R (figure 3).

Synthesis models related to a particular digital filter are known as waveguide models. They can be used to simulate many different systems, such as a tube representing the resonant system in wind instruments (Cook 1992).

3. ANALYSIS OF REAL SOUNDS

The analysis of natural sounds calls for several methods giving a description or a representation of pertinent physical and perceptive characteristics of the sound (Risset and Wessel 1982). Even though the spectral content of a sound is often of great importance, the time course of its energy is at least as important. This can be shown by artificially modifying the attack of a percussive sound in order to make it 'woolly', or by playing the sound backwards. The time and frequency evolution of each partial component is also significant. The vibrato is a perceptively robust effect that is essential, for example for the synthesis of the singing voice. Another essential aspect that should be taken into account when creating a sound corresponding to a plucked vibrating string is the different decay times of the partials. These examples illustrate the need for analysis methods giving access to time and frequency variations of sounds. To solve this general analysis problem of signals, a collection of methods called joint representations has been designed.

The analysis methods of signals can be divided into two principal classes: parametric methods and nonparametric methods. The parametric methods require *a priori* knowledge of the signal, and consist of adjusting the parameters of a model. The nonparametric models do not need any knowledge of the signal to be analysed, but they often require a large number of coefficients.

3.1. Parametric methods

These techniques are generally optimal for the representation of signals adapted to the chosen parametric model. The most common method used for processing sounds is linear prediction (LPC). This technique is adapted to signals from sound production systems of the source–resonance type. The resonant filter should be modelled by a digital all-pass filter the coeffficients of which are related to the frequency and to the width of the formants. The applications of analysis–synthesis for speech signals are numerous, because of a good correspondence between the physics of the vocal tract and the linear filtering. The input signal of LPC systems is generally a broadband noise or a periodic signal adapted to a subtractive synthesis technique.

3.2. Nonparametric methods

Nonparametric techniques for analysis of sound signals generally correspond to representations with physically and/or perceptively meaningful parameters. The best known is the spectral representation obtained through the Fourier transform. In this case the signal is associated with a representation giving the energy distribution as a function of frequency. As mentioned earlier, this representation is not sufficient for characterising the timbre and the dynamic aspects of a sound. In what follows we describe the joint time–frequency representations considering both dynamic and frequency aspects. The time–frequency transformations distribute the total energy of the signal in a plane similar to a musical score in which one of the axes corresponds to the time and the other to the frequency. Such representations are to sound what musical scores are to melodies. There are two ways of obtaining this kind of representation depending on whether the analysis acts on the energy of the signal or on the signal itself. In the first case the methods are said to be nonlinear, giving, for instance, representations from the so-called 'Cohen's class'. The best known example of transformations within this class is the Wigner–Ville distribution (Flandrin 1993). In the other situation the representations are said to be linear, leading to the Fourier transform with a sliding window, the Gabor transform, or the wavelet transform. The linear methods have, at least as far as sound signals are concerned, a great advantage over the nonlinear methods. Linear methods make the resynthesis of signals possible and they ensure that no spurious data cause confusion during the interpretation of the analysis. These spurious data can occur in nonlinear analysis as a result of cross-terms appearing in the development of the square of a sum. This is why we shall focus on the linear time–frequency methods.

The linear representations are obtained by decomposing the signal into a continuous sum of elementary functions having the same properties of localisation both in time and in frequency. These elementary functions correspond to the impulse response of bandpass filters. The central frequency of the analysis band is related to a frequency parameter for time–frequency transformations and is related to a scaling parameter for wavelet transforms. The choice of the elementary functions gives the shape of the filter.

3.2.1. Gabor transform

In the case of the Gabor transform, the elementary functions, also called time–frequency atoms, are all generated from a 'mother' function (window) translated in time and in frequency. The 'mother' function is chosen to be well localised in time and frequency and to have finite energy (for instance a Gaussian function) (figure 4).

Each value of the transform in the time–frequency plane is obtained by comparing the signal to a time–frequency atom. This comparison is mathematically expressed by a scalar product. Each horizontal line of the Gabor transform then corresponds to a filtering of the signal by a bandpass filter centred at a given frequency with a shape that is constant as a function of frequency. The vertical lines correspond to the Fourier transform of a part of the signal, isolated by a window centred on a given time. The transform obtained this way is generally complex, since the atoms themselves are complex, giving two complementary images (Kronland-Martinet *et al.* 1987). The first one is the modulus of the transform and corresponds to a classical spectrogram, the square of the modulus being interpreted as the energy distribution in the time–frequency plane. The second image corresponding to the phase of the transform is generally less well known and less used, but it nevertheless contains a lot of information. This information concerns mainly the 'oscillating part' of the signal (figure 5). Actually, the time derivative of the phase has the dimension of a frequency and leads to the frequency modulation law of the signal components (Guillemain and Kronland-Martinet 1996).

3.2.2. Wavelet transform

The wavelet transform follows a principle close to that of the Gabor transform. Again the horizontal lines of the wavelet transform correspond to a filtering of the signal by a filter, the shape of which is independent of the scale, but whose bandwidth is

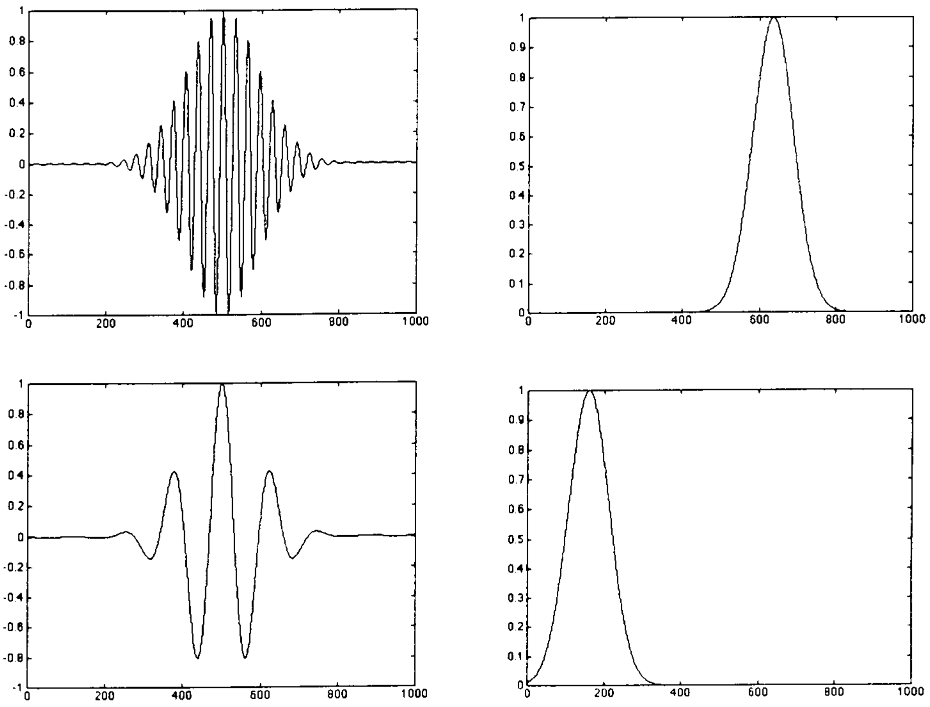


Figure 4. Two Gabor functions in the time domain (left), and their Fourier transform (right). In the Gabor representation, all the filters are obtained by shifting a 'mother' function in frequency, yielding a constant absolute bandwidth analysis.

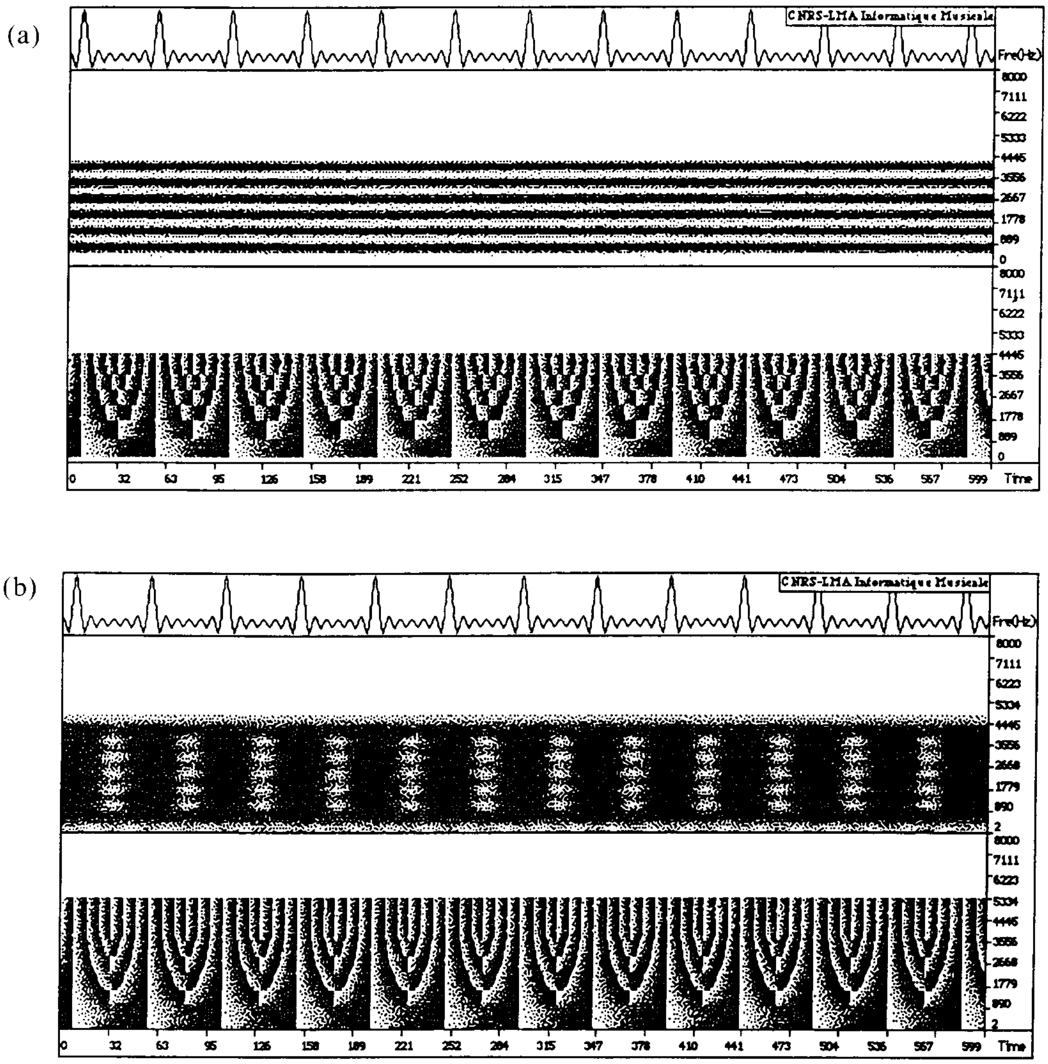


Figure 5. Gabor transform of the sum of six harmonic components analysed with two windows; the horizontal axis is time, the vertical axis is frequency. The upper picture is the modulus, the lower is the phase, represented by modulo-2π; their values are coded with a greyscale. In (a) the window is well localised in frequency, allowing the resolution of each component. In (b) the window is well localised with respect to time, leading to a bad separation of the components in the frequency domain, but showing impulses in time because the signal can also be considered as a filtered Dirac comb. In both figures, the phase behaves similarly, showing the periodicity of each component. This property has been used to estimate the frequencies of the components accurately.

inversely proportional to the scale. The analysis functions are all obtained from a 'mother' wavelet by translation and change of scale (dilation) (figure 6).

The 'mother' wavelet is a function with finite energy and zero mean value. These 'weak' conditions offer great freedom in the choice of this wavelet. One can, for example, imagine the decomposition of a speech signal in order to detect the word 'bonjour' pronounced at different pitches and with different durations. By using a 'mother' wavelet made of two wavelets separated, for example, by an octave, one can detect octave chords in a musical sequence (Kronland-Martinet 1988). This corresponds to a matched filtering at different scales. One important aspect of the wavelet transform is the localisation. By acting on the dilation parameter, the analysing function is automatically adapted to the size of the observed phenomena (figure 7). A high-frequency phenomenon should be analysed with a function that is well localised in time, whereas a low-frequency phenomenon requires a function well localised in frequency. This leads to an appropriate tool for the characterisation of transient signals (Guillemain *et al.* 1996). The particular geometry of the time–scale representation, where the dilation is represented according to a logarithmic scale (in fractions of octaves) permits the transform to be interpreted like a musical score associated with the analysed sound.

3.3. Parameter extraction

The parameter extraction method makes use of the qualitative information given by the time–frequency

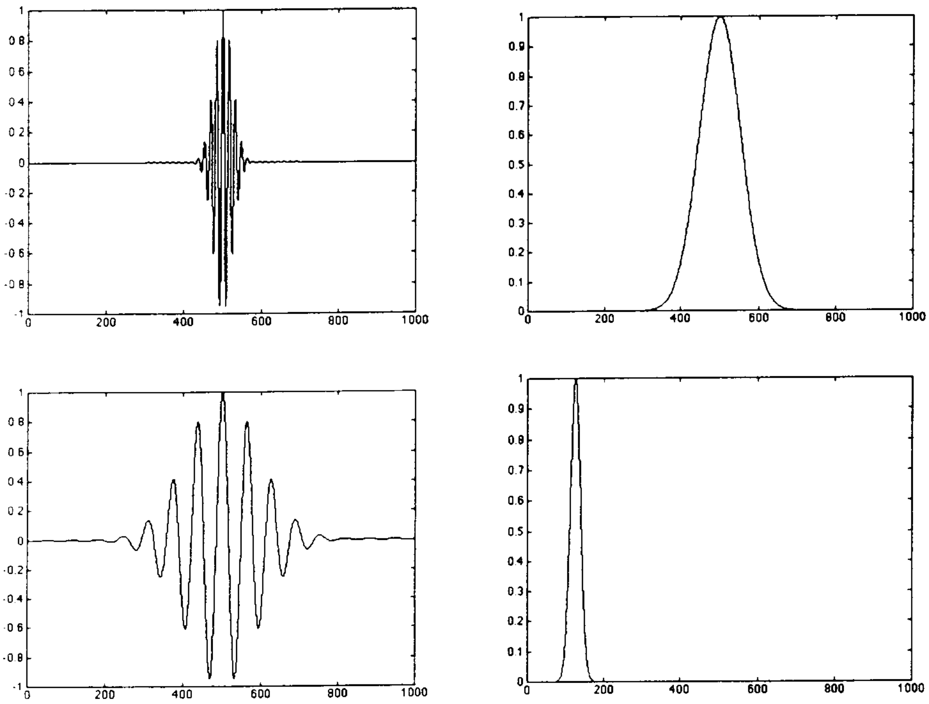


Figure 6. Two wavelets in the time domain (left), and their Fourier transform (right). In the wavelet representation, all the filters are obtained through dilation of a 'mother' function in time, yielding a constant relative ($\Delta\omega/\omega$) bandwidth analysis.

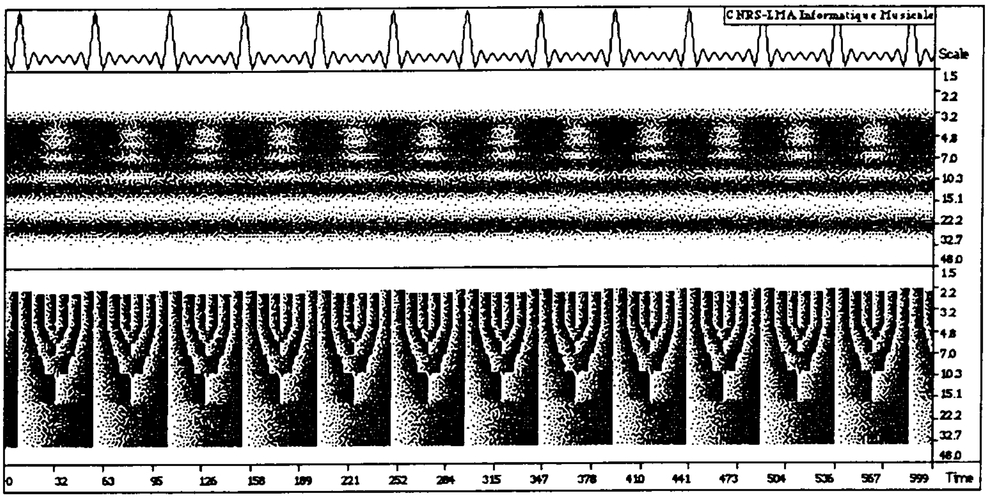


Figure 7. Wavelet transform of the same sum of six harmonic components. In contrast with figure 5 obtained through the Gabor transform, the wavelet transform privileges the frequency accuracy at low frequency (large scales) and the time accuracy at high frequency (small scales).

quantitative information from the signal. Even though the representations are not parametric, the character of the extracted information is generally determined by the supposed characteristics of the signal and by future applications. A useful representation for isolated musical instrument sounds is the additive model. It describes the sound as a sum of elementary components modulated in amplitude and in frequency, which is relevant from a physical and a perceptive point of view (figure 8).

Thus, to estimate parameters for an additive resynthesis of the sound, amplitude and frequency modulation laws associated with each partial should be extracted from the transform. Of course, this process must be efficient even for extracting components that are very close to each other and have rapidly changing amplitude modulation laws. Unfortunately, all the constraints for constructing the representation make this final operation complicated. The justification is of the same nature as the one given in the introduction in connection with sound transformation through modifying the representations. Absolute accuracy both in time and in frequency is impossible because of a mathematical relation between the transform at a point of the time–fre-

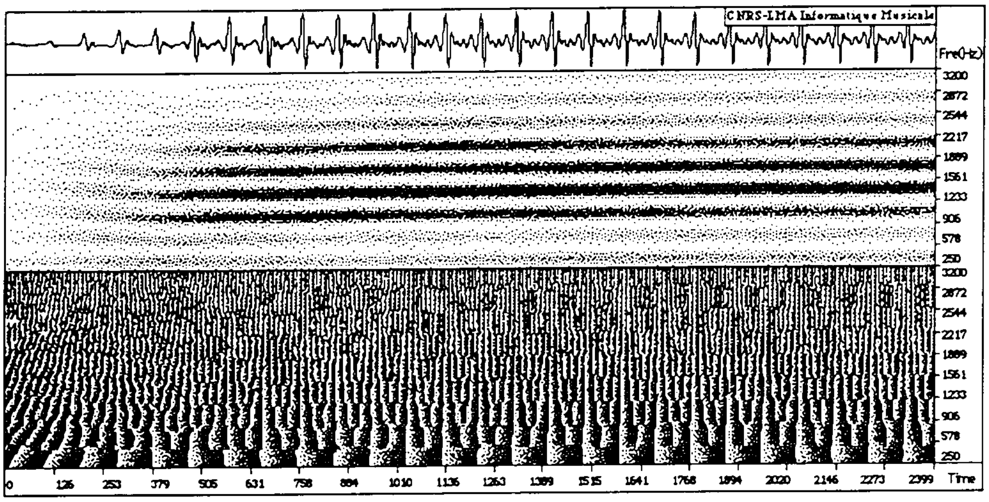


Figure 8. Gabor representation of the first 75 ms of a trumpet sound. Many harmonics with different time dependencies are visible in the modulus picture. The phase picture shows different regions, around each harmonic, where the phase wraps regularly at the time period of each harmonic, as in the previous figure.

Human hearing follows a rather similar 'uncertainty' principle: to identify the pitch of a pure sound, it must last for a certain time. The consequences of these limitations on the additive model parameter estimation are easy to understand. A high-frequency resolution necessitates analysis functions that are well localised in the frequency domain and therefore badly localised in the time domain. The extraction of the amplitude modulation law of a component from the modulus of the transform on a trajectory in the time–frequency plane smoothes the actual modulation law. This smoothing effect acts in a time interval with the same length as the analysis function. Conversely, the choice of well-localised analysis functions in the time domain generally yields oscillations in the estimated amplitude modulation laws, because of the presence of several components in the same analysis band. It is possible, however, to avoid this problem by astutely using the phase of the transform to precisely estimate the frequency of each component and by taking advantage of the linearity in order to separate them, without a hypothesis on the frequency selectivity of the analysis (figure 9).

The procedure uses linear combinations of analysis functions for different frequencies to construct a bank of filters with a quasi-perfect reconstruction. Each filter specifically estimates a component while conserving a good localisation in the time domain. Different kinds of filters can be designed, which permit an exact estimation of amplitude modulation laws locally polynomial on the time support of the filters (Guillemain *et al.* 1996) (figure 10).

The strict limitations of the wavelet transform or of the Gabor transform can be avoided by optimising the selectivity of the filter as a function of the density of the frequency components. The use of continuous transforms on the frequency axis is of great importance, since the central frequencies of the filters can be precisely calibrated at the frequencies of the components. Another important aspect of the musical sound is the frequency modulation of the components, in particular during the attack of the sound. Here the judicious use of the time derivative of the transform phase offers the possibility of developing iterative algorithms tracking the modulation laws, thus precluding the computation of the whole transform. These algorithms use frequency-modulated analysis functions, the modulations of which are automatically matched to the ones of the signal (Guillemain *et al.* 1996).

4. FEEDING THE SYNTHESIS MODELS

The extraction techniques using the time–frequency transforms directly provide a group of parameters which permit the resynthesis of a sound with the additive model. In addition, they can be used for identification of other synthesis models. The direct parameter identification techniques for the nonlinear models are difficult. Generally they do not give an exact reproduction of a given sound. The estimation criteria can be statistical (minimisation of nonlinear functions) (Horner 1996) or psychoacoustic (centroid of spectrum) (Beauchamp 1975). The direct estimation of physical or subtractive model parameters requires techniques like linear prediction, used, for instance, in speech synthesis (Markel and Gray 1976). Another solution consists in using parameters from the additive synthesis model to estimate another set of parameters corresponding to another synthesis model. In what follows we shall see how this operation can be done for the most common models.

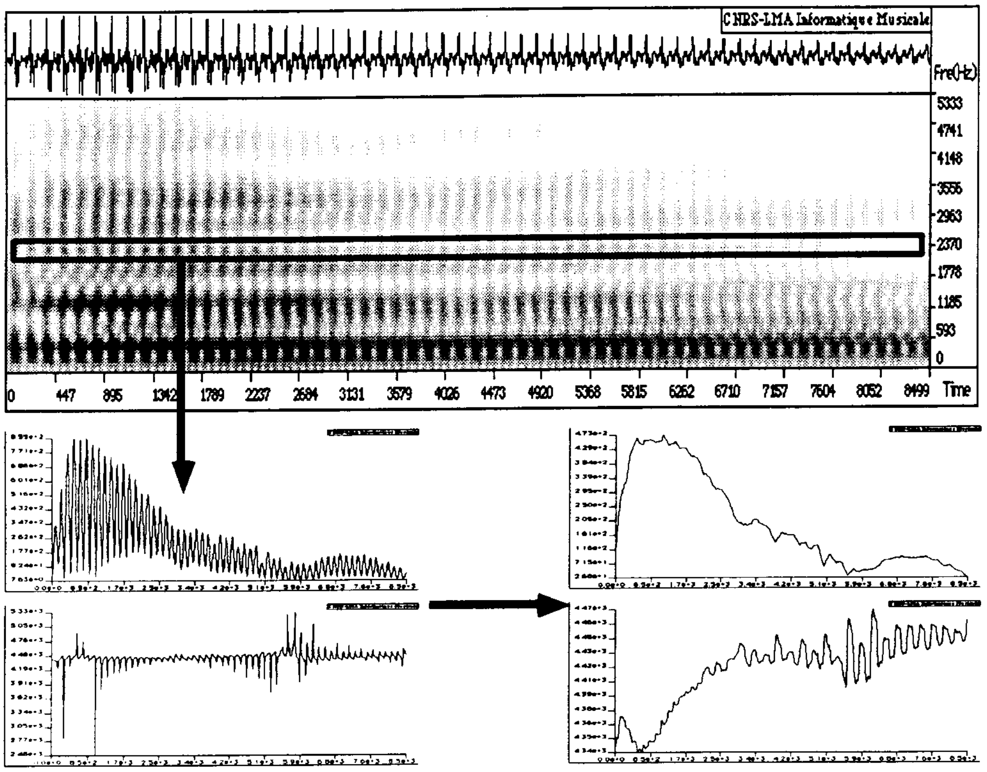


Figure 9. Estimation of the amplitude modulation law of a partial of a saxophone sound. The curves on the left show the estimated amplitude and frequency modulation laws using a straightforward Gabor transform. Several harmonics are present on the frequency support of the analysing function, yielding strong oscillations. The curves on the right show the estimated modulation laws using the filter bank displayed in figure 10. Although the time support remains the same, the oscillations are automatically cancelled by the algorithm.

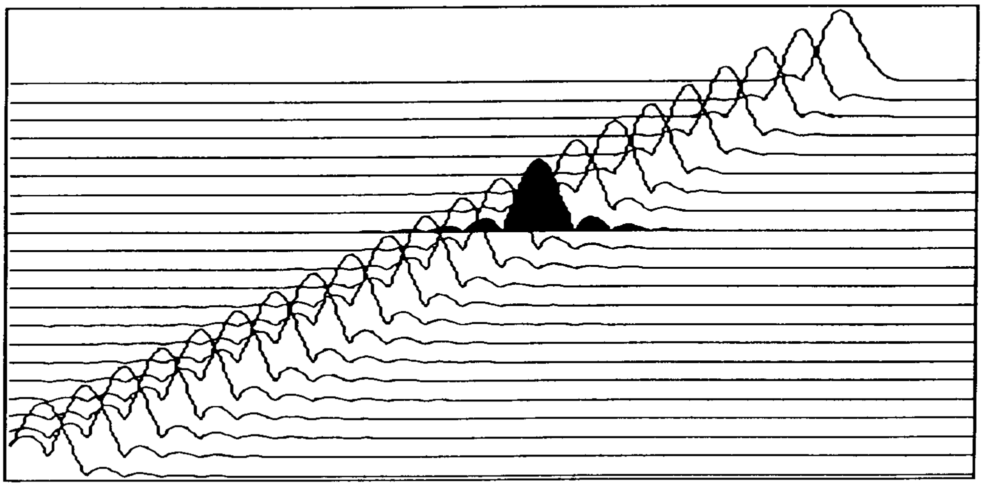


Figure 10. Filter bank in the frequency domain, allowing the estimation of spectral lines; one of the filters is darkened. The Fourier transform of each filter equals unity for the frequency it estimates, and zero for all the others. Its first derivative equals zero for all the frequencies. One can prove that this kind of filter allows exact estimation of locally linear amplitude modulation laws.

4.1. Additive synthesis

The parameter estimation for the additive model is the simplest one, since the parameters are determined in the analysis. The modelling of the envelopes can greatly reduce the data when one uses only perceptive criteria. The first reduction consists of associating each amplitude and frequency modulation law with a piecewise linear function (Horner and Beauchamp 1996) (figure 11). This makes it possible to automatically generate, for example, a Music V score associated with the sound.

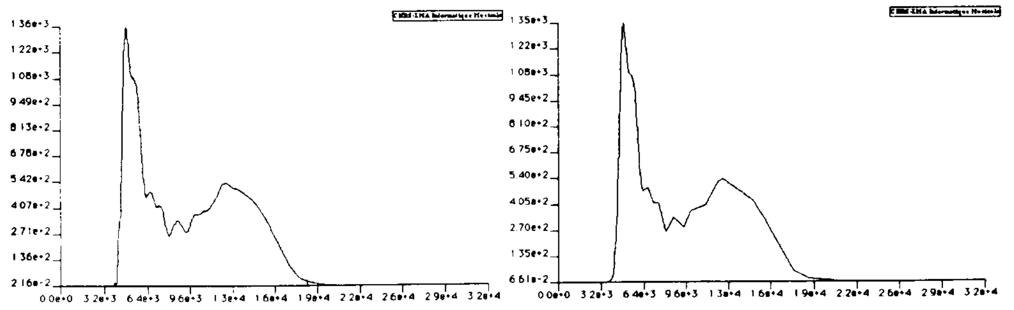


Figure 11. Original and modelled envelopes of a saxophone sound. The modelled curve is defined with thirty-five breakpoints and linear interpolation between them, while the original is defined on 32,000 samples.

Another possible reduction consists in grouping the components from the additive synthesis (group additive synthesis) (Kleczkowski 1989, Oates and Eagleston 1997). This can be done by statistical methods, such as principal component analysis, or by following an additive condition defined as the perceptual similarity between the amplitude modulations of the components (Kronland-Martinet and Guillemain 1993). This method offers a significant reduction in the number of synthesis parameters, since several components with a complex waveshape have the same amplitude modulation laws (figures 12 and 13).

4.2. Subtractive synthesis

An evolutive spectral envelope can be built by creating intermediate components obtained from the modulation laws of the additive modelling. Their amplitude modulation laws are obtained by interpolation of the envelopes of two adjacent components in the frequency domain (figure 14). These envelopes can then be used in order to ‘sculpt’ another sound (crossed synthesis). As we have already mentioned, physical modelling is sometimes close to subtractive synthesis. This aspect will be developed later.

4.3. Waveshaping and frequency modulation synthesis

From the parameters corresponding to the group additive synthesis (complex waves and their associated amplitude laws), one can deduce nonlinear synthesis parameters (Kronland-Martinet and Guillemain 1993). The technique consists of approaching each complex wave shape by an elementary nonlinear module. In the case of waveshaping, the knowledge of the complex wave allows the calculation of an exact distortion function. In the case of FM, the spectral components should be grouped, not only according to a perceptive criterion, but also according to a condition of spectral proximity. This condition is meaningful because real similarities between envelopes of neighbouring components are often observed. To generate the waveform corresponding to a group of components by an elementary FM oscillator, the perceptive approach is best suited. In that case, one can consider the energy and the spectral extent of the waveforms which are directly related to the modulation index. Other methods based on the minimisation of nonlinear functions by the simulated annealing or genetic algorithms have also been explored (Horner 1996). Attempts at direct estimation of the FM parameters by extraction of frequency modulation laws from the phase of the

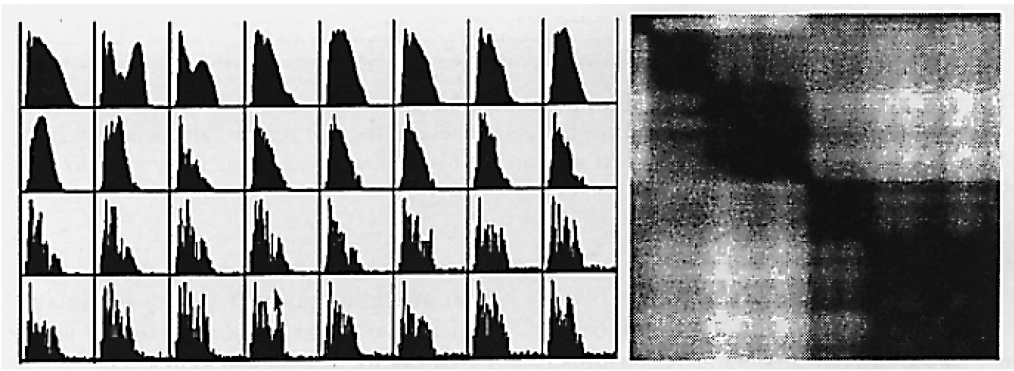


Figure 12. A whole set of envelopes of a violin sound, and the matrix showing the correlation between them. The dark regions around the diagonal correspond to curves that look similar and that correspond to components that are close in the frequency domain.

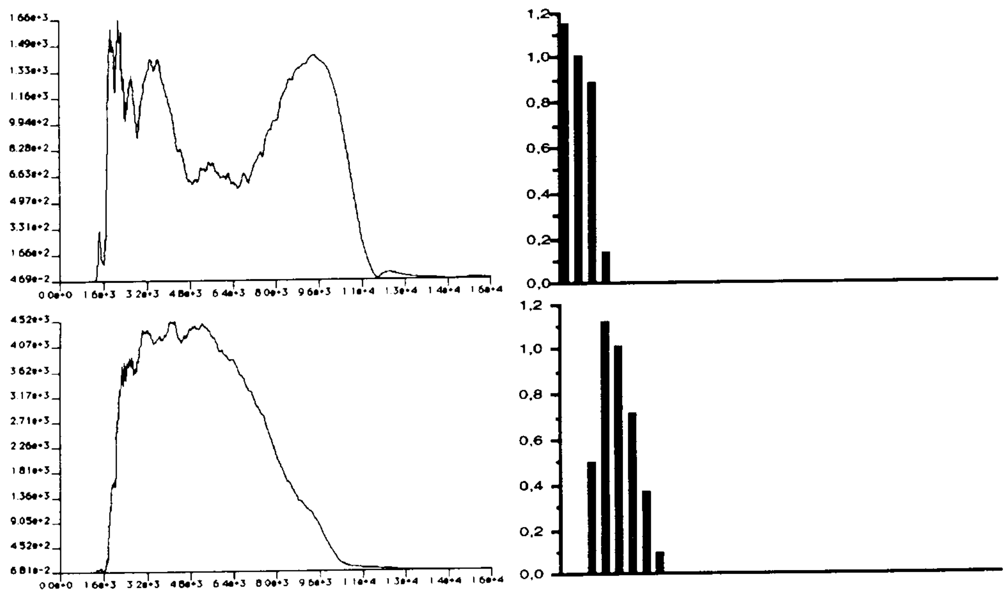


Figure 13. Two main envelopes of the group additive synthesis model, with the spectrum of their associated waveform. Psychoacoustic criteria can be used to generate a perceptively similar spectrum with nonlinear techniques.

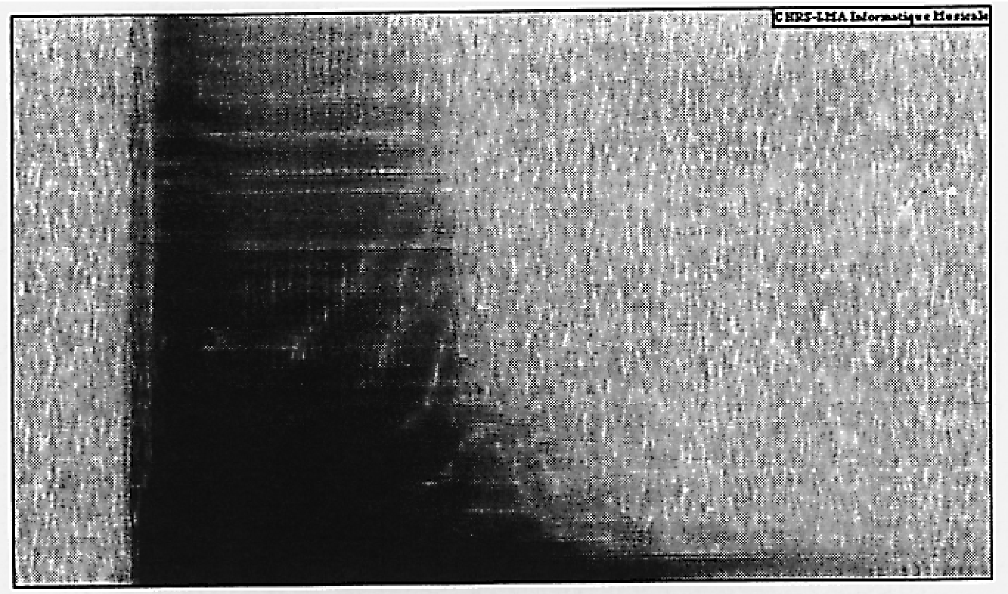


Figure 14. Spectral envelope of a saxophone sound built from the additive synthesis parameters. This envelope can be used to 'sculpt' the modulus of the Gabor transform of another sound in order to perform a crossed synthesis.

analytic signal related to the real sound have led to interesting results (Justice 1979, Delprat, Guillemain and Kronland-Martinet 1990).

4.4. Waveguide

The waveguide synthesis parameters are of a different kind. They characterise both the medium where the waves propagate and the way this medium is excited. From a physical point of view, it is difficult to separate these two aspects: the air jet of a wind instrument causes vortex sheddings interacting with the acoustic pressure in the tube (Verge 1995); the piano hammer modifies the characteristics of a string while it is in contact with it (Weinreich 1977). These source–resonator interactions are generally nonlinear and often difficult to model physically. However, a simple linear waveguide model often gives satisfactory sound results. In a general way, the study of linear wave propagation equations in a bounded medium shows that the response to a transient excitation can be written as a sum of exponentially damped sine functions. The inharmonicity is related to the dispersive characteristics of the propagation medium, the decay times are related to the dissipative characteristics of the medium, and the amplitudes are related to the spectrum of the excitation. In the same way, the impulse

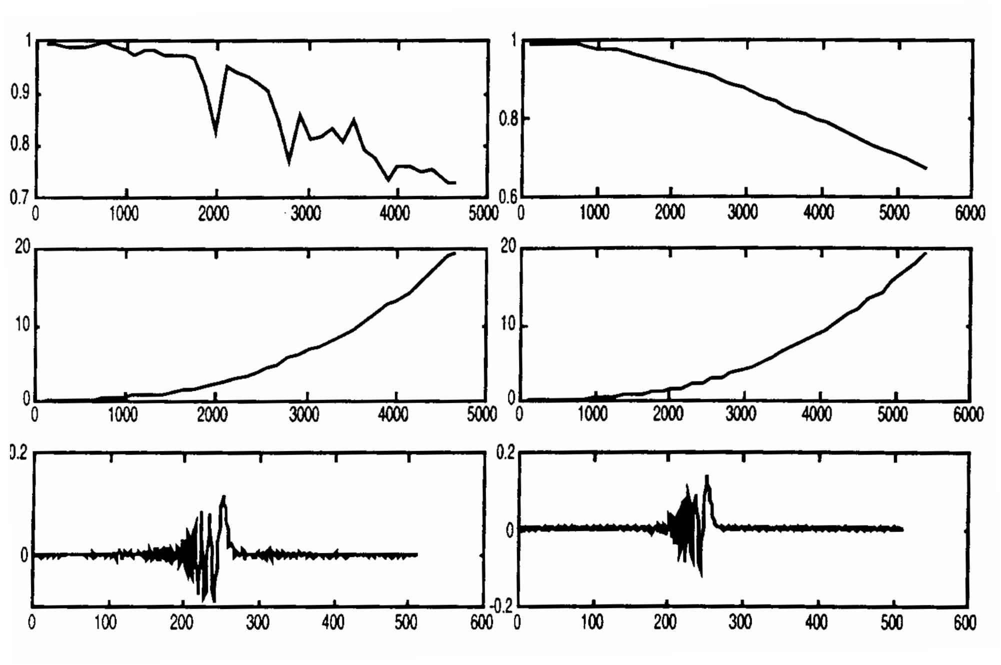


Figure 15. Parameter estimation for the waveguide model can be performed either from the solution of the partial differential equations of the string movement, or from the estimated damping factors and frequencies of the partials. Pictures on the left show the data from the estimation. Pictures on the right show the data from the movement equation of a stiff string. The figure shows, from top to bottom: modulus (related to losses during the propagation), phase derivative (related to the dispersion law of the propagation medium) of the Fourier transform of the filter inside the loop, and impulse response of the loop filter. The good agreement between theory and experimentation in this case can be used to fit mechanical parameters of the string from the experimentation.

approximated by a sum of exponentially damped sinusoids whose frequencies, amplitudes and damping rates are related in a simple way to the filter coefficients (Ystad, Guillemain and Kronland-Martinet 1996). Thanks to the additive synthesis parameters one can, for the percussive sound class, determine the parameters of the waveguide model, and also recover the physical parameters characterising the instrument (Guillemain, Kronland-Martinet and Ystad 1997) (figure 15). For sustained sounds, the estimation problem of the exciting source is crucial and necessitates the use of deconvolution techniques. This approach is entirely nonparametric, but it is also possible to use parametric techniques. Indeed, the discrete time formulation of the synthesis algorithm corresponds to a modelling of the so-called ARMA type (AutoRegressive Moving Average).

5. CONCLUSION

The modelling of sounds brings together the algorithmic synthesis process and the shaping of natural sounds. Such modelling may serve to develop an appropriate ‘algorithmic sampler’ to make all the intimate modifications offered by a mathematical description of the sounds. Time–frequency and time–scale representations of signals are helpful to extract relevant parameters describing the sounds. Both signal synthesis models and physical synthesis models can be fed in order to resynthesise and shape a given musical sound. Even though most musical sounds can be modelled from additive synthesis data, stochastic or very noisy sounds still remain difficult to model. Work is being conducted to fill this gap in order to offer in the near future a genuine sound simulator to musicians.

REFERENCES

Allen, J. B., and Rabiner, L. R. 1977. A unified approach to short-time Fourier analysis and synthesis. *Proc. of the IEEE* **65**: 1558–64.

Arfib, D. 1979. Digital synthesis of complex spectra by means of multiplication of non-linear distorted sine waves. *Journal of the Audio Engineering Society* **27**: 757–68.

Arfib, D., and Delprat, N. 1993. Musical transformations using the modifications of time–frequency images. *Computer Music Journal* **17**(2): 66–72.

Arfib, D., Guillemain, P., and Kronland-Martinet, R. 1992. The algorithmic sampler: an analysis problem? *Journal of the Acoustical Society of America* **92**: 2451.

Atal, B. S., and Hanauer, S. L. 1971. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America* **50**: 637–55.

Beauchamp, J. W. 1975. Analysis and synthesis of cornet tones using non-linear interharmonic relationships. *Journal of the Audio Engineering Society* **23**: 778–95.

Cadoz, C., Luciani, A., and Florens, J. L. 1984. Responsive input devices and sound synthesis by simulation of instrumental mechanisms. *Computer Music Journal* **8**(3): 60–73.

Chaigne, A. 1995. Trends and challenges in physical modeling of musical instruments. *Proc. of the ICMA*, Vol. III, pp. 397–400. Trondheim, Norway, 26–30 June.

Cheung, N. M., and Horner, A. 1996. Group synthesis with genetic algorithms. *Journal of the Audio Engineering Society* **44**: 130–47.

Chowning, J. 1973. The synthesis of complex audio spectra by means of frequency modulation. *Journal of the Audio Engineering Society* **21**: 526–34.

Cook, P. R. 1992. A meta-wind-instrument physical model controller, and a meta-controller for real-time performance control. *Proc. of the 1992 Int. Computer Music Conf.*, pp. 273–6. San Francisco: International Computer Music Association.

Delprat, N., Guillemain, P., and Kronland-Martinet, R. 1990. Parameter estimation for non-linear resynthesis methods with the help of a time–frequency analysis of natural sounds. *Proc. of the 1990 Int. Computer Music Conf.,* pp. 88–90. Glasgow: International Computer Music Association.

Depalle, P., and Rodet, X. 1992. A new additive synthesis method using inverse Fourier transform and spectral envelopes. *Proc. of the 1992 Int. Computer Music Conf.*, pp. 161–4. San Francisco: International Computer Music Association.

De Poli, G., Picciali, A., and Roads, C. (eds.) 1991. *The Representation of Musical Signals*. Cambridge, MA: MIT Press.

Dolson, M. 1986. The phase vocoder: a tutorial. *Computer Music Journal* **10**(4): 14–27.

Flanagan, J. L., Coker, C. H., Rabiner, P. R., Schafer, R. W., and Umeda, N. 1970. Synthetic voices for computer. *IEEE Spectrum* **7**: 22–45.

Flandrin, P. 1993. *Temps-frequence*. Hermes. Traite des nouvelles technologies, serie traitement du signal.

Gabor, D. 1947. Acoustical quanta and the nature of hearing. *Nature* **159**(4044): 591–4.

Grossman, A., and Morlet, J. 1984. Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM Journal of Mathematical Analysis* **15**: 723–36.

Guillemain, Ph., and Kronland-Martinet, R. 1996. Characterisation of acoustics signals through continuous linear time–frequency representations. *Proc. of the IEEE*, Special Issue on Wavelets, **84**(4): 561–85.

Guillemain, Ph., Kronland-Martinet, R., and Ystad, S. 1997. Physical modelling based on the analysis of real sounds. *Proc. of the Institute of Acoustics*, Vol. 19, pp. 445–50. Edinburgh: ICMA97.

Horner, A. 1996. Double-modulator FM matching of instruments tones. *Computer Music Journal* **20**(2): 57–71.

Horner, A., and Beauchamp, J. 1996. Piecewise-linear approximation of additive synthesis envelopes: a comparison of various methods. *Computer Music Journal* **20**(2): 72–95.

Justice, J. 1979. Analytic signal processing in music computation. *IEEE Trans. on Speech, Acoustics and Signal Processing* **ASSP-27**: 670–84.

Kleczkowski, P. 1989. Group additive synthesis. *Computer Music Journal* **13**(1): 12–20.

Kronland-Martinet, R. 1988. The use of the wavelet transform for the analysis, synthesis and processing of speech and music sounds. *Computer Music Journal* **12**(4): 11–20 (with sound examples on disk).

Kronland-Martinet, R. 1989. Digital subtractive synthesis of signals based on the analysis of natural sounds. In *Etat de la Recherche Musicale (au 1er janvier 1989).* Ed. A.R.C.A.M., Aix en Provence.

Kronland-Martinet, R., and Guillemain, P. 1993. Towards non-linear resynthesis of instrumental sounds. *Proc. of the 1993 Int. Computer Music Conf.*, pp. 86–93. San Francisco: International Computer Music Association.

Kronland-Martinet, R., Morlet, J., and Grossman, A. 1987. Analysis of sound patterns through wavelet transforms. *International Journal of Pattern Recognition and Artificial Intelligence* **11**(2): 97–126.

Laroche, J. 1993. The use of the matrix pencil method for the spectrum analysis of musical signals. *Journal of the Acoustical Society of America* **94**: 1958–65.

Le Brun, M. 1979. Digital waveshaping synthesis. *Journal of the Audio Engineering Society* **27**: 250–66.

Makhoul, J. 1975. Linear prediction, a tutorial review. *Proc. of the IEEE* **63**: 561–80.

Markel, J. D., and Gray, A. H. 1976. Linear prediction of speech. *Communication and Cybernetics* **12**. Berlin, Heidelberg, New York: Springer-Verlag.

McAulay, R., and Quatieri, T. 1986. Speech analysis–synthesis based on a sinusoidal representation. *IEEE Trans. on Speech, Acoustics and Signal Processing* **ASSP-34**: 744–54.

Moorer, J. A. 1978. The use of the phase vocoder in computer music applications. *Journal of the Audio Engineering Society* **26**: 42–5.

Oates, S., and Eagleston, B. 1997. Analytic methods for group additive synthesis. *Computer Music Journal* **21**(2): 21–39.

Rabiner, L. R., and Gold, B. 1975. *Theory and Application of Digital Signal Processing*. Englewood Cliffs, NJ: Prentice Hall.

Risset, J. C. 1965. Computer study of trumpet tones. *Journal of the Acoustical Society of America* **33**: 912.

Risset, J. C., and Wessel, D. L. 1982. Exploration of timbre by analysis and synthesis. In D. Deutsch (ed.) *The Psychology of Music*, pp. 26–58. New York: Academic Press.

Roads, C. 1978. Automated granular synthesis of sound. *Computer Music Journal* **2**(2): 61–2.

Ruskai, M. B., Beylkin, G., Coifman, R., Daubechies, I., Mallat, S., Meyer, Y., and Raphael, L. (eds.) 1992. *Wavelets and their Applications*. Boston: Jones and Bartlett.

Smith, J. 1992. Physical modeling using digital waveguides. *Computer Music Journal* **16**(4): 74–91.

Verge, M. P. 1995. *Aeroacoustics of Confined Jets with Applications to the Physical Modeling of Recorder-like Instruments*. PhD Thesis, Eindhoven University.

Weinreich, G. 1977. Coupled piano strings. *Journal of the Acoustical Society of America* **62**: 1474–84.

Ystad, S., Guillemain, Ph., and Kronland-Martinet, R. 1996. Estimation of parameters corresponding to a propagative synthesis model through the analysis of real sounds. *Proc. of the 1996 Int. Computer Music Conf.,* pp. 19–24. Hong Kong: International Computer Music Association.