Analyse et modèle génératif de l'expressivité : application à la Parole et à l'Interprétation musicale

Grégory Beller

THESE DE DOCTORAT

UNIVERSITE PARIS VI – PIERRE ET MARIE CURIE

ECOLE DOCTORALE D'INFORMATIQUE, TELECOMUNICATIONS
ET ELECTRONIQUE (EDITE) DE PARIS

Spécialité **Informatique**

Préparée à l'IRCAM
par

Grégory Beller

En vue de l'obtention du titre de

DOCTEUR DE L'UNIVERSITE PIERRE ET MARIE CURIE

Analyse et Modèle Génératif de l'Expressivité

Application à la Parole et à l'Interprétation Musicale

Thèse soutenue le mercredi 24 Juin 2009

devant le jury composé de :

Gérard Bailly	rapporteur	GIPSA-lab
Christophe D'Alessandro	Examinateur	LIMSI-CNRS
Laurence Devillers	rapporteure	LIMSI-CNRS
Thierry Dutoit	Examinateur	TCTS
Axel Röbel	Examinateur	IRCAM
Xavier Rodet	Directeur	IRCAM
Jean-Luc Zarader	Examinateur	ISIR

Thèse préparée à l'IRCAM
(Institut de recherche et de coordination acoustique/musique)

Sous la direction de Xavier Rodet

IRCAM-CNRS UMR STMS 9912
Equipe Analyse Synthèse
1 place Igor Stravinsky
75004 Paris – France

Dédicace

Je dédie cette thèse à Christian Le Vraux, le premier artiste que j'ai rencontré, qui m'a permis de me lancer dans ce projet fou que d'utiliser les machines les plus savantes de notre siècle, pour en faire émerger le son le plus commun de l'humanité : le rire. Je lui dédie ce travail dont il n'aura jamais vu l'aboutissement.

Remerciements

Je tiens à remercier les personnes qui m'ont accompagné dans l'accomplissement de ce travail...

... Mon directeur de thèse, Xavier Rodet, pour la confiance, les bons conseils et la liberté d'entreprise qu'il m'a donnés. C'est grâce à son ouverture d'esprit que j'ai pu aborder des sujets assez divers et mener des expériences dont l'issue n'était pas toujours courue d'avance...

... Les membres du jury, qui ont accepté de prendre en compte ce travail, de l'apprécier et de l'évaluer dans sa totalité. En particulier, j'aimerais remercier les rapporteurs qui ont suscité des corrections et qui ont ainsi participé grandement à la clarté et au contenu de ce manuscrit...

... Mes collègues chercheurs de l'IRCAM, Axel Roebel, Geoffroy Peeters, Christophe Veaux, Gilles Degottex, Pierre Lanchantin, Thomas Hueber, Diemo Schwarz, Fernando Villavicencio, Aurélien Marty, Arshia Cont, Julien Bloit, Nicolas Rasamimanana, Olivier Lescurieux, Clara Suied, Norbert Schnell, Snorre Farner, qui ont tous apporté, un jour, une pierre à cet édifice ...

... Et plus largement, l'ensemble de mes collègues de l'IRCAM qui ont fait de mon cadre de travail, un environnement plaisant, motivant et performant. Je les remercie notamment pour m'avoir aider à y organiser un colloque international sur l'expressivité dans la musique et dans la parole ... à créer une mailing liste pour les personnes intéressés par les liens entre la parole et la musique [musicspeech@ircam.fr]... A conduire divers enregistrements d'acteurs et d'instrumentistes... et pour la bonne humeur contagieuse...

... Mes collègues chercheurs hors IRCAM, Klaus Scherer, Marc Schroeder, Nick Campbell, Jürgen Trouvain, Annirudth Patel, Jacqueline Vaissière, Véronique Aubergé, Laurence Devillers, Nicolas Audibert, Félix Burkhardt, Maëva Garnier, Cédric Gendrot, Martine Adda-Decker, Philippe Boula de Mareüil, Chloé Clavel, grâce à qui j'ai beaucoup appris, tant sur les émotions que sur la parole...

... Dans l'organisation du cycle de conférences internationales EMUS (Expressivity in MUsic and Speech), mes collaborateurs internes Nicolas Obin, Florence Quilliard et Andrew Gerzso, ainsi que mes collaborateurs externes Anne Lacheret, Aliyah Morgenstern, Antoine Auchlin, Christophe d'Alessandro, Sandra Madureira et Plinio Barbosa...

... Dans l'organisation de la journée littérature musique, encadrée par la collaboration CDMC-CNSMDP, François Nicolas et Michel Chaillou, qui m'ont permis de faire une expérience en temps réel avec d'éminents poètes.

... Chinkel, IRISA et France Telecom pour les discussions que nous avons pu avoir dans le cadre du projet RIAM VIVOS, et tout particulièrement, Gaëlle Vidal (IRISA) avec qui j'ai élaboré l'annotation des corpus expressifs, à plus de 200Km de distance (merci à l'Internet au passage), et Martin Delille (Chinkel), avec qui j'ai conduit certains enregistrements. Voxler, Cantoche, et LIMSI pour les discussions dans le cadre du projet ANR Affective Avatar. Cantoche et Chinkel à nouveau pour le projet ANR ReSpoken...

... Les acteurs qui m'ont parlé de leur métier : Jacques Gamblin, Abbi Patrix, Martin Girard, et tout particulièrement, ceux qui ont livré leur parole aux enregistrements : Jacques Combes, Philippe Roullier, Olivia, Daniele (demandez les noms à Martin Delille)...

... Les instrumentistes, qui m'ont parlé de leur métier : Bertrand Brayard, Florimont Dal Zotto, Benjamin Chabert, Michèle Buirette, ainsi que ceux qui se sont prêtés à l'exercice difficile des enregistrements, les violonistes...

... Les compositeurs qui me permettent de croire en l'utilité de ce travail pour la composition musicale : David Chaillou, David Coll, Lara Morciano, Evdokja Danaïlovska, Stephano Gervasoni, Georges Aperghis...

... Les artistes qui m'ont livré leurs avis sur ce travail, Norbert Godon, Florent Truchel, Béatrice de Fays, Pierre Alexandre...

... Eric Daubresse, pour la confiance qu'il m'a donné...

... Ma mère, mon père, et mes nombreuses familles... Beller, Patrix, Le Vraux, Palais, Andral, Guyho, Laghrani, Guillonnet, et tout particulièrement Marie-France Andral qui s'est prêtée au jeu de la chasse aux fautes d'orthographe...

... Isabel Woehler, pour sa patience hors du commun...

... Anne Sédès, Martin Laliberté, Horacio Vagione, ainsi que tous mes collègues et étudiants des universités Paris 8 et Paris Est - Marne la vallée...

... Tous mes amis, mes colocataires, tout ceux que j'oublie et tout ceux que je n'oublie pas...

Résumé

Cette thèse s'inscrit dans les recherches actuelles sur les émotions et les réactions émotionnelles, sur la modélisation et la transformation de la parole, ainsi que sur l'interprétation musicale. Il semble que la capacité d'exprimer, de simuler et d'identifier des émotions, des humeurs, des intentions ou des attitudes, est fondamentale dans la communication humaine. La facilité avec laquelle nous comprenons l'état d'un personnage, à partir de la seule observation du comportement des acteurs et des sons qu'ils émettent, montre que cette source d'information est essentielle et, parfois même, suffisante dans nos relations sociales. Si l'état émotionnel présente la particularité d'être idiosyncrasique, c'est-à-dire particulier à chaque individu, il n'en va pas de même de la réaction associée qui se manifeste par le geste (mouvement, posture, visage, ...), le son (voix, musique, ...), et qui, elle, est observable par autrui. Ce qui nous permet de penser qu'il est possible de transformer cette réaction dans le but de modifier la perception de l'émotion associée.

C'est pourquoi le paradigme d'analyse-transformation-synthèse des réactions émotionnelles est, peu à peu, introduit dans les domaines thérapeutique, commercial , scientifique et artistique. Cette thèse s'inscrit dans ces deux derniers domaines et propose plusieurs contributions.

D'un point de vue théorique, cette thèse propose une définition de l'expressivité (et de l'expression neutre), un nouveau mode de représentation de l'expressivité, ainsi qu'un ensemble de catégories expressives communes à la parole et à la musique. Elle situe l'expressivité parmi le recensement des niveaux d'information disponibles dans l'interprétation qui peut être vu comme un modèle de la performance artistique. Elle propose un modèle original de la parole et de ses constituants, ainsi qu'un nouveau modèle prosodique hiérarchique.

D'un point de vue expérimental, cette thèse fournit un protocole pour l'acquisition de données expressives interprétées. Colatéralement, elle rend disponible trois corpus pour l'observation de l'expressivité. Elle fournit une nouvelle mesure statistique du degré d'articulation ainsi que plusieurs résultats d'analyses concernant l'influence de l'expressivité sur la parole.

D'un point de vue technique, elle propose un algorithme de traitement du signal permettant la modification du degré d'articulation. Elle présente un système de gestion de corpus novateur qui est, d'ores et déjà, utilisé par d'autres applications du traitement automatique de la parole, nécessitant la manipulation de corpus. Elle montre l'établissement d'un réseau bayésien en tant que modèle génératif de paramètres de transformation dépendants du contexte.

D'un point de vue technologique, un système expérimental de transformation, de haute qualité, de l'expressivité d'une phrase neutre, en français, synthétique ou enregistrée, a été produit, ainsi qu'une interface web pour la réalisation d'un test perceptif en ligne.

Enfin et surtout, d'un point de vue prospectif, cette thèse propose différentes pistes de recherche pour l'avenir, tant sur les plans théorique, expérimental, technique que technologique. Parmi celles-ci, la confrontation des manifestations de l'expressivité dans les interprétations verbales et musicales semble être une voie prometteuse.

Mots-clés

Emotions, expressivité, performance artistique, interprétation musicale, parole, prosodie, transformation du signal de parole, modélisation générative, apprentissage, réseau bayésien.

Abstract

This thesis joins in the current searches(researches) on the feelings and the emotional reactions, on the modelling and the transformation of the speech, as well as on the musical performance. It seems that the capacity to express, to feign and to identify emotions, humors, intentions or attitudes, is fundamental in the human communication. The ease with which we understand the state of a character, from the only observation of the behavior of the actors and the sounds which they utter, shows that this source of information is essential and, sometimes, sufficient in our social relationships. If the emotional state presents the peculiarity to be idiosyncratic, that is private to every individual, it does not also go away of the associated reaction which shows itself by the gesture (movement, posture, face), the sound (voice, music), and which, it is observable by others.

That is why paradigm of analysis - transformation - synthesis of the emotional reactions grows on into the therapeutic, commercial, scientific and artistic domains. This thesis joins in these last two domains and proposes several contributions. From a theoretical point of view, this thesis proposes a definition of the expressivity, a definition of the neutral expression, a new representation mode of the expressivity, as well as a set of expressive categories common to the speech and to the music. It places the expressivity among the census of the available levels of information in the performance which can be seen as a model of the artistic performance. It proposes an original model of the speech and its constituents, as well as a new hierarchical prosodic model.

From an experimental point of view, this thesis supplies a protocol for the acquisition of performed expressive data. Collaterally, it makes available three corpora for the observation of the expressivity. It supplies a new statistical measure of the degree of articulation as well as several analysis results concerning the influence of the expressivity on the speech.

From a technical point of view, it proposes a speech processing algorithm allowing the modification of the degree of articulation. It presents an innovative database management system which is used, already, used by some other automatic speech processing applications, requiring the manipulation of corpus. It shows the establishment of a bayesian network as generative model of context dependent transformation parameters.

From a technological point of view, an experimental system of high quality transformation of the expressivity of a French neutral utterance, either synthetic or recorded, has been produced, as well as an on-line interface for perceptive tests.

Finally and especially, from a forward-looking point of view, this thesis proposes various research tracks for the future, both on the theoretical, experimental, technical, and technological aspects.

Among these, the confrontation of the demonstrations of the expressivity in the speech and in the musical performance seems to be a promising way.

Keywords

Emotions, expressivity, artistic performance, musical performance, speech, prosody, speech signal transformation, generative model, machine learning, bayesian network.

_ ***C'est une erreur flagrante que d'assimiler la science à la raison pure et à la logique, comme l'art à l'intuition et à l'émotion.***

A. Köstler, *Le cri d'Archimède : art de la découverte*, 1965

Table des matières

Chapitre 1

Introduction

Sommaire

> ***Tout le talent de l'acteur consiste à faire éprouver aux spectateurs des émotions qu'il ne ressent pas lui-même.***
>
> S. Guitry, dans *Sacha Guitry : Roi du théâtre* de René Benjamin, 1933

Pour transformer l'expression de la parole, il est tout d'abord nécessaire de définir la notion d'expressivité, d'un point de vue théorique. Puis, d'un point de vue pratique, il faut modifier le signal de parole, de manière à ce que la perception de l'expression en soit changée. Cette apparente simplicité cache de nombreux verrous théoriques et techniques (notamment liés à la nature idiosyncrasique des émotions, et à la variabilité, à la richesse et à la complexité du phénomène de la parole), que cette thèse tente de mettre en évidence, tout en essayant d'y apporter quelques éléments de réponse. D'un point de vue pratique, un programme expérimental a été crée, permettant de conférer à n'importe quelle phrase, en français, enregistrée ou synthétisée, une expression désirée avec un certain degré d'intensité.

Cette thèse s'inscrit dans les recherches actuelles sur les émotions et les réactions émotionnelles associées, sur la modélisation de la parole, sur la transformation du signal de parole, ainsi que sur l'interprétation musicale. Après une mise en situation dans le contexte actuel, cette introduction permet de présenter la démarche scientifique utilisée, la structure de ce manuscrit, les problématiques abordées (qui révèlent, plus en détail, la structure de ce manuscrit), ainsi que les contributions apportées par cette thèse.

1.1 Contexte actuel

Il semble que la capacité d'exprimer, de simuler et d'identifier des émotions, des intentions ou des attitudes, est fondamentale dans la communication humaine. La facilité avec laquelle nous comprenons le jeu des comédiens illustre bien notre capacité fondamentale à décoder les émotions, les attitudes, ou les humeurs des autres, à partir de la seule observation de leurs comportements et des sons qu'ils émettent. Car, en effet, si l'état émotionnel présente la particularité d'être idiosyncrasique, c'est-à-dire particulier à chaque individu [Picard 1997], il n'en va pas de même de la réaction associée qui, elle, est observable par plusieurs personnes. Ce qui nous permet de penser qu'il est possible de transformer cette réaction dans le but de modifier la perception de l'émotion associée.

C'est pourquoi, l'analyse-transformation-synthèse des émotions est, peu à peu, introduite dans de nouvelles applications artistiques, scientifiques, thérapeutiques, commerciales, juridiques [Eriksson 2007]... Le projet européen HUMAINE[1], devenu, aujourd'hui, une association, présente sur son portail, une application technologique différente par jour qui implique, de manière directe ou indirecte, les

[1] HUMAINE : http://emotion-research.net/

émotions. L'ISRE[2] propose, quant à elle, des applications en sociologie et en psychologie. L'un des champs d'application important réside évidemment dans l'introduction des émotions dans les nouveaux types d'interaction Homme-machine [Bailly 1992, Dutoit 2003]. En effet, les synthétiseurs de parole produisent, aujourd'hui, une parole intelligible et parfois naturelle. Un nouveau domaine de recherche a alors récemment émergé, dans le but de conférer à cette parole, certains attributs : l'identité ou certains traits d'un locuteur spécifique (sexe, âge, origine, état de santé...), le style employé (professionnel, relationnel,...), et, enfin, l'émotion exprimée.

1.2 Enjeu de la thèse

L'enjeu de la thèse est l'établissement d'un système de transformation de l'expressivité pour différents domaines artistiques. Des metteurs en scène, des compositeurs, des studios de doublage et des producteurs de jeux vidéo et d'avatars sont, en effet, intéressés par les multiples possibilités que pourrait fournir un système capable de transformer et de simuler des réactions émotionnelles, dans la voix, comme le montre le projet ANR-VIVOS[3].

Le système Expresso, qui est décrit dans ce manuscrit, doit permettre à un utilisateur de contrôler la voix d'un acteur virtuel. Cela est possible grâce à la synthèse de parole expressive à partir du texte. Comme il existe déjà de nombreux synthétiseurs TTS (Text To Speech) de bonne qualité et qu'il est intéressant de pouvoir aussi modifier l'expression d'une phrase enregistrée, le problème a été découpé en deux : D'un côté, la synthèse de parole neutre et intelligible, et de l'autre, la modification de l'expression de la parole. Ce découpage permet notamment de concentrer l'effort sur la modification de l'expression d'une phrase possédant déjà une structuration linguistique. Ainsi, le système Expresso se destine à la modification artificielle de la manière de parler, c'est à dire de la prosodie. L'utilisateur présente une phrase au système et choisit une expression ainsi qu'un degré d'expression. Après plusieurs étapes, le système lui retourne la même phrase, modifiée de sorte que l'expression choisie puisse être perçue.

L'élaboration du système Expresso a nécessité, au préalable, la définition d'un cadre théorique de l'expressivité, l'établissement d'un modèle de la parole et le choix d'un protocole expérimental adéquat. Ensuite, des données ont été acquises afin de fournir des exemples. L'analyse de ces données a permis l'établissement d'un modèle génératif de l'expressivité dans la parole. Ce modèle est utilisé pour contrôler des algorithmes de traitement du signal qui modifient la manière de parler. Enfin, l'ensemble du système Expresso a été évalué par un test perceptif.

Ces différentes étapes situent donc les contributions majeures de cette thèse dans le domaine du traitement automatique de la parole (TAP), et plus particulièrement dans le volet de recherche dédié à l'analyse-transformation-synthèse de l'expressi-

[2]ISRE : International Society for Research on Emotion : http://isre.org/index.php
[3]VIVOS : http://www.vivos.fr

vité de la parole. Toutefois, plusieurs résultats dépassent ce seul cadre et peuvent contribuer indirectement à d'autres domaines, comme le montre une partie plus prospective dédiée à l'étude de l'expressivité dans l'interprétation musicale.

1.3 Proposition centrale

1.3.1 Le paradoxe du comédien

Au théâtre, comme au cinéma, les acteurs s'efforcent d'exprimer, le mieux possible, les émotions et les états intérieurs que l'œuvre prête à leurs personnages. Pour cela, ils ne se contentent pas de dire leurs répliques. Ils jouent aussi avec un éventail de gestes vocaux, de mimiques faciales et de postures. Ils savent que le moindre signe, aussi fugace qu'une respiration ou qu'un battement de cil, est utilisé par les spectateurs, comme un indice pour percer le caractère d'un personnage ou l'état dans lequel il se trouve. Pourtant, ils ne sont pas forcément habiter par l'émotion qu'ils sont en train de communiquer. Cela est aussi vrai lors d'une interprétation musicale, durant laquelle les instrumentistes expriment certaines émotions tout en étant sujet à d'autres, comme le trac, par exemple. Enfin, cette différence entre l'émotion ressentie et l'émotion communiquée se retrouve dans le large cadre de la performance artistique, qui réunit l'interprétation verbale, musicale, gestuelle (la danse) et picturale, et encore plus largement, dans notre vie de tous les jours.

Le Paradoxe sur le comédien est un ouvrage de réflexions sur le théâtre rédigées par Denis Diderot entre 1773 et 1777. Diderot s'oppose à l'opinion courante qui suppose que, pour être convaincant, le comédien doit ressentir les passions qu'il exprime. Diderot soutient au contraire qu'il ne s'agit que de manifestations des émotions : il faut donc faire preuve de sang froid afin d'étudier ces passions, pour ensuite les reproduire. En effet, Diderot expose deux sortes de jeux d'acteurs : "Jouer d'âme", qui consiste à ressentir les émotions que l'on joue et "Jouer d'intelligence", qui repose sur le paraître. Les émotions jouées se lisent sur le visage de l'acteur, mais ce dernier ne les ressent absolument pas.

Le Paradoxe du comédien nuance la validité de théories sur les émotions, uniquement fondées sur l'observation de données émotionnelles. Nous verrons qu'en effet, ce paradoxe reste valable à la fois dans les méthodes directes et indirectes d'acquisition de données émotionnelles. De plus, il révèle la possibilité de communiquer certaines émotions, sans pour autant les ressentir. A l'instar du comédien, le spectateur peut percevoir et comprendre l'état émotionnel d'un personnage, sans pour autant être le siège d'une émotion induite. Ainsi, le paradoxe du comédien montre qu'il est nécessaire de dissocier l'émotion ressentie de l'émotion communiquée. De plus, il permet de penser qu'une machine est capable de simuler le jeu d'un acteur en faisant percevoir des émotions, sans pour autant être le siège d'une quelconque expérience émotionnelle.

1.3.2 L'expressivité

Le Paradoxe du comédien révèle l'importance de dissocier l'émotion ressentie de l'émotion communiquée. C'est pourquoi nous introduisons une définition de l'expressivité[4].

1.3.2.1 Une définition de l'expressivité

> L'expressivité est un niveau d'information dans la communication. Ce niveau regroupe les manifestations externes, contrôlées ou non, qui sont attribuables à des états internes inaccessibles à la perception d'autrui.

L'expressivité fait donc partie de la communication, en tant que démonstrateur d'un état interne qui est, par essence, inaccessible à autrui. Cet état interne peut être existant ou non, et sa manifestation peut donc être simulée ou non. Dans tous les cas, l'expressivité s'appuie sur l'ensemble des indices multimodaux par lesquels nous sommes capables d'inférer l'état interne d'autrui. Parmi ces états internes sont compris les phénomènes affectifs en général, c'est à dire les émotions (utilitaires, esthétiques), les attitudes (interpersonnelles, préférentielles), les sentiments et les humeurs (personnelles, référentielles) recensées par Scherer [Scherer 2005] :

- Emotions utilitaires : Emotions classiquement recensées facilitant notre adaptation ou notre réaction à un événement déterminant pour notre survie ou notre qualité de vie : "colère", "peur", "joie", "dégoût", "tristesse", "surprise", "honte", "culpabilité".
- Les préférences : Jugements relatifs à l'évaluation ou à la comparaison d'objets : "j'aime", "je n'aime pas", "je préfère"...
- Les attitudes : Jugements, prédispositions ou croyances, relativement stables, relatives à des objets, des événements, des personnes, des groupes ou des catégories de personnes : "Je déteste", "je hais", "j'apprécie", je désire"....
- Les humeurs : États affectifs diffus, dont la cause est difficilement identifiable, et qui affecte le comportement d'un individu : "joyeux", "mélancolique", "apathique", "déprimé", "allègre/enjoué".
- Les dispositions affectives : Traits de personnalité ou humeurs récurrentes pouvant favoriser l'apparition plus fréquente d'une émotion spécifique, même en réponse à de légers stimuli[5] : "nerveux", "anxieux", "irritable", "téméraire", "morose", "hostile", "envieux", "jaloux".
- Positions relationnelles : Style affectif émergeant spontanément ou employé stratégiquement lors d'un échange : "poli", "distant", "froid", "chaud", "ave-

[4]Sur la définition de l'expressivité. L'expressivité est définie dans le dictionnaire comme étant relative à l'expression. Nous avons donc choisi de définir ce terme plutôt que de redéfinir le terme d'expression qui est déjà largement polysémique. De plus, le terme expressivité est proposé comme un anglicisme du terme expressivity, largement employé aujourd'hui dans la communauté anglophone.

[5]Scherer souligne l'importance de différentier les humeurs passagères des dispositions affectives. Par exemple, être déprimé momentanément ne pose pas de problèmes comportementaux, tandis qu'être dépressif nécessite un suivi clinique.

nant", "méprisant".
- Emotions esthétiques : Kant définit les émotions esthétiques par des plaisirs désintéressés ("interesseloses Wohlgefallen"), mettant en exergue l'absence complète de considérations utilitaires : "ému", "effrayé", "plein de désir", "admiratif", "bienheureux", "extasié", "fasciné", "harmonieux", "ravissement", "solennité". Ces termes sont notamment utilisées pour décrire l'expérience musicale (voir partie 6).

1.3.2.2 L'expression "neutre"

S'il n'a pas encore été démontré qu'il existe un état émotionnel ou psychologique interne neutre, dans lequel ni émotion, ni humeur, ni sentiment, ni attitude n'existe, l'existence d'un niveau d'expressivité zéro, dans lequel le protagoniste ne donne aucune information sur son état interne, semble largement acceptée par de nombreux chercheurs qui décrivent cette absence d'information par l'expression : *neutre*. Une phrase prononcée avec l'expression neutre ne donne donc aucun renseignement sur l'état interne de son locuteur. C'est pourquoi l'expression neutre est souvent, voire toujours, utilisée comme état de référence dans l'étude et la comparaison des expressions [Tao 2006, Z.Inanoglu 2009].

1.3.3 Transformation de l'expressivité dans la parole

1.3.3.1 Paradigme central de la thèse

> Un stimulus expressif peut être vu comme un stimulus neutre modifié par l'expressivité.

Le stimulus neutre présente possède l'avantage de présenter tous les autres niveaux d'information disponibles dans la parole :
- le message sémantique, porté par les aspects linguistiques
- l'identité du locuteur, portée par les caractéristiques de sa voix
- le style de parole emprunté
- les aspects pragmatiques relatifs à la parole en contexte
- (l'expressivité)

Transformer l'expression d'un tel stimulus, sous-entend modifier seulement ce niveau d'information, sans altérer les autres, c'est à dire, en conservant le message sémantique, l'identité du locuteur, le style de parole emprunté et les aspects pragmatiques convoyés par la parole. Dès lors, ces niveaux d'information doivent être connus d'un modèle dont le but n'est de transformer que l'expression. De plus, cette expression peut-être modéliser comme une modulation de l'expression neutre. Le système de transformation de l'expressivité repose donc sur la mise en place d'un modèle capable d'apprendre et de générer des modulations relatives aux expressions.

1.3.3.2 Présentation d'Expresso

Dans le but de conférer une expression désirée, avec un degré variable de l'expression, à une phrase neutre donnée, qui peut être enregistrée ou synthétisée, le système Expresso transforme le signal de parole, sur la base d'exemples fournis par des acteurs. De sorte de ne pas modifier les autres niveaux d'information de la parole, Expresso compare les versions neutres et expressives des acteurs, pour ne modéliser que la variation due à l'expression. Un modèle génératif permet, en effet, de prédire la façon dont un des acteurs enregistrés aurait prononcé la même phrase que la phrase neutre à transformer, dans une version neutre et dans une version expressive. La différence entre ces deux versions permet de définir la variation qu'apporte l'expression. Cette variation est alors appliquée à la phrase neutre source, par des algorithmes de traitement du signal, dans le but de lui ajouter l'expression désirée.

1.3.3.3 Préparation du modèle

Tout au long de ce manuscrit va s'établir le modèle qui servira pour le système de transformation de l'expressivité. Ce modèle repose sur deux types de variables dont la notation est ici explicitée. D'un côté, les variables linguistiques symboliques, de nature discrète et/ou catégorielle (c'est à dire qui possède un alphabet déterminé et fini) sont désignées par un S comme "Symbolique". Par exemple, la variable $Smodalite^{phrase}$ désigne la modalité d'une phrase, tandis que la variable $Sphoneme^{phone}$ désigne la catégorie phonétique d'un phone. De l'autre côté, les variables acoustiques, de nature continue, (c'est à dire qu'elles peuvent prendre une infinité de valeur, même si leur support est borné) sont désignées par un A comme "Acoustique". Par exemple, la moyenne de la fréquence fondamentale sur une syllabe s'écrit $Af0^{syllabe}_{moyenne}$. Cette notation sera utilisée dans toute la suite du manuscrit. Les variables encadrées serviront particulièrement à la définition du modèle final.

1.4 Structure du manuscrit

Cette introduction permet de poser les éléments fondateurs de la thèse. La proposition centrale exposée est à la base de notre démarche expérimentale et technique, puisque le système Expresso de transformation de l'expressivité repose fondamentalement sur cette proposition. Le premier chapitre présente les choix qui ont permis de constituer des corpus expressifs. Un état de l'art des méthodes de représentation des émotions et des techniques d'acquisition de données émotionnelles montre à quel point ces dernières sont d'importance dans l'étude des émotions. Cependant, il montre aussi qu'il est difficile d'évaluer le degré de spontanéité d'une donnée émotionnelle. C'est pourquoi nous avons enregistré des acteurs simulant des expressions pour constituer des corpus expressifs. Le second chapitre propose un modèle de la parole qui est ensuite appliqué aux corpus, dans le but de fournir des résultats d'analyse. Ces résultats permettent notamment de définir des variables de la pa-

role qui sont varient selon les expressions. Le troisième chapitre présente et évalue le système Expresso. Expresso permet la transformation de l'expressivité d'une phrase neutre par transformation du signal. Un modèle statistique reposant sur les variables décrites précédemment, et entraîné sur les corpus expressifs, permet d'inférer la modulation relative à l'expressivité. Cette modulation est ensuite appliquée sur la phrase à transformer dans le but d'en modifier l'expression. Enfin, un test perceptif permet une évaluation préliminaire du système qui débouche sur une discussion. La conclusion générale permet de résumer les contributions de la thèse. Puis, un chapitre prospectif traite de l'expressivité dans l'interprétation musicale. Une analogie entre la parole et l'interprétation musicale est faite dans le but d'en comparer les expressivités. Un corpus d'interprétations musicales expressives a été enregistré et des pistes de recherche pour le futur sont données. Enfin, des annexes présentent certains travaux relatifs à la thèse, comme la constitution d'une plateforme pour la gestion des corpus ou la synthèse semi-paramétrique de rires.

1.5 Contributions

D'un point de vue théorique, cette thèse propose une définition de l'expressivité (et de l'expression neutre), un nouveau mode de représentation catégorico-dimensionnel de l'expressivité, ainsi qu'un ensemble de catégories expressives communes à la parole et à l'interprétation musicale. Elle situe l'expressivité parmi le recensement des niveaux d'information disponibles dans l'interprétation qui peut être vu comme un modèle de la performance artistique. Elle propose un modèle original de la parole et de ses constituants, ainsi qu'un nouveau modèle prosodique hiérarchique. D'un point de vue expérimental, elle fournit un protocole pour l'acquisition de données expressives interprétées. Colatéralement, elle rend disponible trois corpus pour l'observation de l'expressivité. Elle propose un système de gestion de corpus novateur qui est, d'ores et déjà, utilisé pour d'autres applications du traitement automatique de la parole, nécessitant l'usage de corpus. Elle fournit une nouvelle mesure statistique du degré d'articulation ainsi que plusieurs résultats d'analyse concernant l'influence de l'expressivité sur la parole. D'un point de vue technique, elle propose un nouvel algorithme de traitement du signal permettant la modification du degré d'articulation. Elle montre la mise en place d'un réseau bayésien en tant que modèle génératif de paramètres de transformation. D'un point de vue technologique, un système expérimental, de haute qualité, de transformation de l'expressivité d'une phrase neutre, synthétique ou enregistrée, en français a été produit. De même, un test perceptif reposant sur une interface web a été crée pour son évaluation. Enfin, cette thèse offre différentes pistes de recherche pour l'avenir, tant sur le plan théorique, qu'expérimental, technique et technologique.

Chapitre 2

Constitution d'un corpus expressif

Sommaire

Les mots manquent aux émotions.

V. Hugo, *Le Dernier Jour d'un condamné*, 1829

2.1 Résumé du chapitre

Ce premier chapitre permet de présenter les choix qui nous ont guidé dans la constitution de corpus expressifs. Tout d'abord, un mode de représentation de l'expressivité a du être défini afin d'être le socle commun au chercheur, au comédien et à l'auditeur. Un aperçu des différents modes de représentation des émotions a permis le choix d'une définition des expressions sous la forme de vecteurs lexicaux. Puis, les différentes techniques d'acquisition de données émotionnelles existantes sont recensées. La difficulté d'évaluer le degré de spontanéité d'une donnée émotionnelle, ainsi que l'enjeu de simuler le jeu d'un acteur, ont gouverné notre choix vers l'emploi de méthodes d'acquisition directes. Des acteurs-comédiens ont donc simulé un ensemble choisi d'expressions. Plusieurs corpus ont alors été enregistrés avec différents protocoles. Ces corpus de phrases constituant des paires d'expressions neutre-expressive ont notamment été définis de manière à isoler l'influence de l'expressivité sur la parole, de l'influence des autres niveaux d'information.

2.2 Représentation des émotions

Lorsque la communauté scientifique s'approprie un sujet d'étude tel que les émotions ou les réactions émotionnelles (hypothétiquement associées), elle se doit de définir une manière commune de les représenter. Or, la nature secrète, idiosyncrasique et mystérieuse de nos états émotionnels pose des problèmes dans la constitution d'une définition. Puisque l'expérience émotionnelle dépend de nombreux facteurs contextuels et subjectifs, un nombre conséquent de termes composent le champ sémantique des émotions. Aussi, certains chercheurs tentent de représenter les émotions par d'autres biais, suivant leurs caractéristiques communes, leurs fonctions ou d'autres critères communs. La projection de catégories discrètes dans des espaces continus, dont les axes sont qualificatifs de certains de ces critères, permet la représentation dimensionnelle des émotions.

2.2.1 Représentations catégorielles

La littérature scientifique abonde de terminologies tantôt dérivées du langage commun, tantôt issues de traductions de langues étrangères, parfois mal déterminées et possédant des doublons. Malgré la récurrence de certains termes et des efforts de standardisation, comme le montre la construction récente du language à balise "EmotionML"[1], il n'existe pas encore un dictionnaire de l'émotion dans la communauté scientifique. Seules les catégories définies par Ekman [Ekman 1999a] semblent consensuelles : peur, joie, colère, dégoût, surprise et tristesse. Le consensus terminologique sur ces émotions primaires, repose sur la constatation universelle d'Ekman qui provient de l'observation d'invariants de certaines réactions émotionnelles (faciales) [Ekman 1999b] dans différentes cultures.

Mais la définition d'une taxonomie est souvent dépendante de son utilisation. Par exemple, le terme "colère" est souvent employé, mais cette catégorie est parfois divisée en "colère froide" et en "colère chaude", selon les besoins. De plus, au cours d'une tâche d'étiquetage de réactions émotionnelles [Devillers 2003a], certains annotateurs préfèrent utiliser plusieurs termes pour décrire une manifestation émotionnelle, constituant par là, des vecteurs lexicaux de l'émotion. Cette idée de composition des émotions est d'ailleurs présente depuis LeBrun (passions simples et passions composées). La *palette-theory* [Scherer 1984] présente les émotions secondaires comme des compositions d'émotions primaires. Quelques exemples de termes emprunts au vaste champ lexical des émotions (souvent métaphoriques) sont exposés dans la partie 2.2.4.

2.2.2 Représentations dimensionnelles

Les notions de composition, d'agrégation et de proximité des catégories émotionnelles suscitent la possibilité de représenter celles-ci dans des espaces communs. La définition des axes de ces espaces communs, ou bien de métriques com-

[1] EmotionML : http://www.w3.org/2005/Incubator/emotion/XGR-emotionml-20081120/

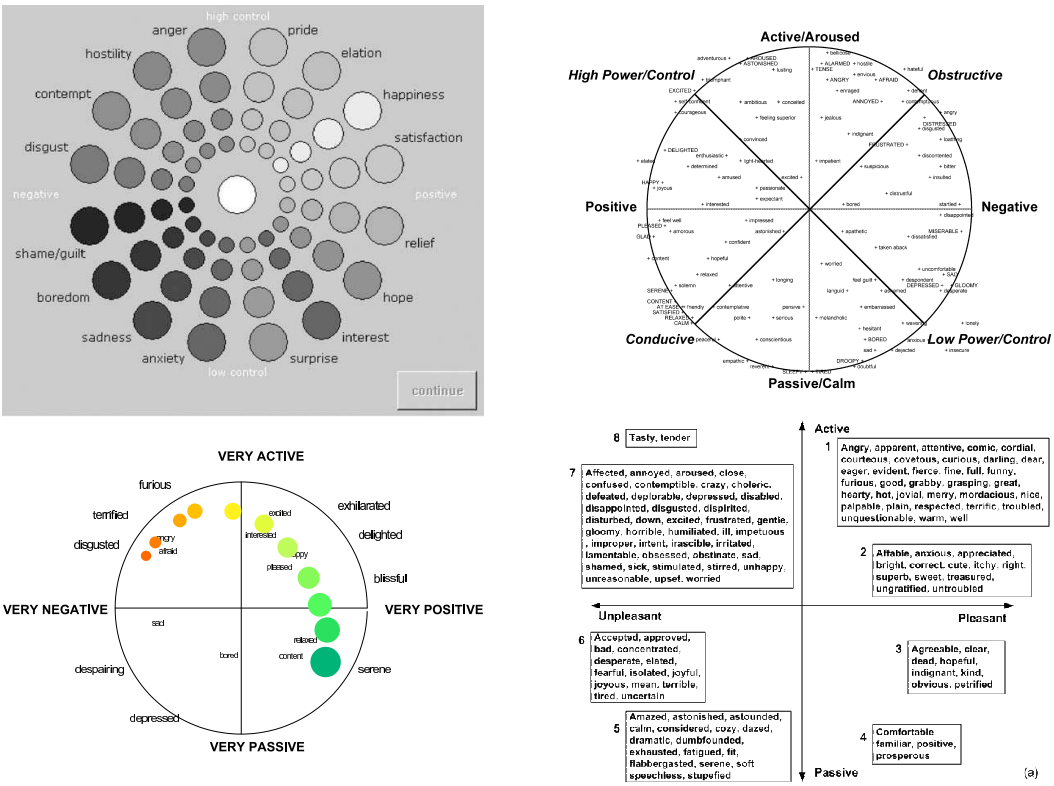

FIG. 2.1: Représentations dimensionnelles des émotions : En haut à gauche : Roue des émotions de Genève [Scherer 2005] ; En haut à droite : Variante de la roue des émotions de Genève [Scherer 2005] ; En bas à gauche : L'outil FeelTrace [Schroeder 2003] ; En bas à droite : Espace lexical affectif [Fitrianie 2006] ;

munes entre les catégories linguistiques discrètes, est au cœur des représentations dimensionnelles des émotions. Les représentations dimensionnelles reposent donc sur la définition d'attributs communs aux émotions. Celles-ci présentent les réactions émotionnelles, ou les émotions, dans des espaces continus, dont les axes peuvent différer [Schroeder 2004]. Parmi les axes les plus courants, on retrouve la valence (évaluation positive ou négative de l'émotion), le degré ou l'intensité ou la puissance, l'activation ("arousal") qui mesure la propension à l'action, l'introversion et l'extraversion de la réaction émotionnelle, et d'autres axes encore, relatifs à l'évaluation cognitive (nouveauté, intérêt, contrôle, effort...[Frijda 1986]). L'usage de ces représentations est favorisé par la recherche de corrélations entre ces axes perceptifs et des axes objectifs caractérisant les données émotionnelles (taille du sourire, hauteur moyenne de la voix, fréquence cardiaque...). Quelques exemples de représentations dimensionnelles sont réunies dans la figure 2.1

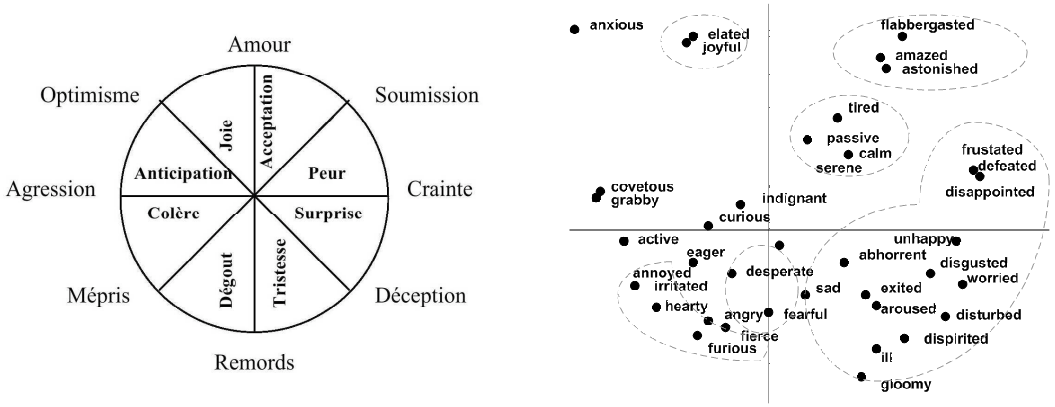

FIG. 2.2: Représentations géométriques des émotions : À gauche : Roue des émotions de Plutchik [Plutchik 1980] ; À droite : MDS de catégories émotionnelles [Fitrianie 2006].

2.2.3 Représentations géométriques

A l'instar de projeter des émotions dans des espaces dimensionnels, il existent des représentations métaphoriques, souvent catégorico-dimensionnelles, appelées "représentation géométriques des émotions". Ces représentations placent des termes appartenant aux représentations catégorielles dans des espaces continus adimensionnels, grâce à l'utilisation de métriques déduites ou arbitraires. Quelques exemples sont donnés dans la figure 2.2. Ces représentations peuvent être obtenues automatiquement par Multi Dimensional Scaling (MDS) [Fitrianie 2006] ou par l'utilisation de cartes auto-organisatrices de Kohonen, mais elle peuvent aussi être définies empiriquement, comme c'est le cas de la roue de Plutchik [Plutchik 1980] (voir figure 2.2) ou celui des émotions musicales d'Hevner [Hevner 1936] (voir figure 6.2).

2.2.4 Représentations contextuelles

La partie 2.3 montre que les informations relatives aux contextes d'apparition des émotions peuvent renseigner sur leurs degrés de spontanéité. Mais ils peuvent aussi nous renseigner sur d'autres aspects des émotions. De plus, la variété du contexte d'apparition d'une réaction émotionnelle complique sa qualification, qui ne peut, parfois, pas se résumer à un seul mot, compte tenu de l'exemple des vecteurs lexicaux [Devillers 2003a]. C'est pourquoi une terminologie commune doit prendre en compte tout ou partie pertinente de ce contexte. Le contexte d'apparition d'une émotion peut être partiellement décrit par sa fonction dans une situation, ainsi que par différents paramètres influant sur cette situation. Scherer [Scherer 2005] crée plusieurs sous-catégories de phénomènes affectifs dépassant le seul cadre des émotions. En effet, les émotions primaires d'Ekman, qu'il appelle "modales" ou "utilitaires" ne sont, dans cette classification, qu'une sous-catégorie des phénomènes affectifs. Les différentes sous-catégories de phénomènes affectifs recensées par Scherer, avec leurs définitions, ainsi que des termes appartenant à leurs champ lexicaux

ont été reportés dans la partie 1. Ces différentes sous-catégories se différencient fortement selon leurs contextes d'apparition, et selon différents critères communs relatifs à l'interaction entre le phénomène affectif et son contexte :

- Causalité : Critère exprimant la facilité avec laquelle le phénomène affectif peut être attaché à une cause, un stimulus.
- Evaluation intrinsèque : Critère mesurant le rapport entre le stimulus et les préférences de l'individu.
- Evaluation extrinsèque : Critère relatif au rapport entre le stimulus et les besoins/désirs/attitudes de l'individu.
- Synchronisation : Critère mesurant l'efficacité de la synchronisation des différents processus[2] composant la réponse affective.
- Vitesse d'adaptation : Critère mesurant la rapidité de changement du comportement.
- Impact comportemental : Critère relatif à la profondeur du changement du comportement.
- Intensité : Intensité de la réponse à un stimulus.
- Durée : Durée de la réponse à un stimulus.

Ces critères décrivent la manière dont les stimuli influent sur le comportement d'un individu. Ils permettent donc de définir tacitement le contexte d'apparition d'une réaction émotionnelle. Grâce à ces échelles d'évaluation et par des mesures expérimentales, Scherer distingue et compare les phénomènes affectifs.

2.2.5 Représentations sous la forme de vecteurs lexicaux

Ces différentes manières de représenter les émotions sont parfois complémentaires. Une émotion peut être décrite par un terme consacré de la langue, par des attributs cognitifs, par son placement géométrique vis à vis des autres émotions, ou encore par son contexte d'apparition. De manière à réunir les avantages des représentations catégorielles et ceux des représentations dimensionnelles, nous proposons une représentation des expressions sous la forme de vecteurs lexicaux. L'émotion est définie par un terme auquel est adjoint deux valeurs, relatives à deux dimensions provenant des représentations dimensionnelles : L'intensité et le degré d'extraversion. Ainsi, nous parlerons dans ce manuscrit de "tristesse extravertie d'intensité 5" (sur une échelle allant de 1 à 6, du moins au plus expressif). Les catégories retenues sont exposées dans la partie 2.6.5). Ce mode de représentation permet de distinguer certaines classes communément réunies, alors qu'elles se manifestent acoustiquement de manière différentes. Par exemple, la colère contenue est exprimée différemment de la colère explosive, tout comme la peur tétanisante, vis à vis de la peur alarmante.

[2]Les différents processus composant la réponse affective sont décrit par le "component process model" de Scherer : Composante cognitive (appréciation), composante neurophysiologique (symptômes corporels), composante motivationnelle (relative à l'action), composante motrice de l'expression (faciale, vocale...), composante du sentiment subjectif (expérience émotionnelle) [Scherer 1987]

Deux cas existent pour la création de ces représentations. Dans le premier cas, c'est souvent la donnée émotionnelle qui est représentée au départ. Puis, lorsque le nombre de données expérimentales est statistiquement significatif, ces données se transforment en catégories, censée représenter l'état interne, c'est à dire l'émotion susceptible d'être responsable de ces données. Dans le deuxième cas, l'étiquetage de ces données, par plusieurs annotateurs, peut conduire à des échelles perceptives ou bien à des champs lexicaux communs, qui sont ensuite employés pour définir les données. Dans tous les cas, ces représentations s'appuient sur l'observation de données émotionnelles, tout comme la plupart des théories sur les émotions.

2.3 Acquisition de données émotionnelles

Les différentes représentations des émotions reposent souvent sur l'observation de données émotionnelles. Parmi ces données, on distingue les *données internes* des *données externes*. Les données internes correspondent à des observations inaccessibles à la perception usuelle d'autrui. Ces données peuvent provenir de l'imagerie médicale comme l'Imagerie par Résonance Magnétique (IRM et IRMf pour fonctionnelle) [Beaucousin 2006], la Tomographie à Émission de Positons (TEP) ou l'Électro/Magnéto-Encéphalographie (EEG ou MEG). Ces données participent fortement à la compréhension des mécanismes neuronaux impliqués dans les processus émotionnels. Les données émotionnelles peuvent aussi provenir de mesures physiologiques : les variations de la tension artérielle, des rythmes cardiaques et de la conductivité de la peau ou encore, l'Electro-Glotto-Graphie (EGG) pour la parole, sont autant d'indices de nos états somatiques. Les données externes correspondent à des observations accessibles à tous, par le biais de nos modalités perceptives (l'ouïe et la vue, surtout). Elles consistent en des enregistrements de réactions émotionnelles faciales [Ekman 1999b], posturales et gestuelles [Bresin 2003] ou vocales [Murray 2008], et en les résultats d'analyses effectuées sur ces enregistrements. Les données externes sont les plus répandues car elles sont plus faciles d'acquisition. C'est pourquoi de nombreuses théories sur les émotions reposent sur ces données.

2.3.1 Degré de spontanéité des données émotionnelles

L'une des difficultés majeures de l'acquisition de données émotionnelles externes, réside dans l'évaluation du degré de spontanéité de l'émotion dont elles proviennent. Nous présentons dans l'ordre chronologique, les théories marquantes qui permettent d'appréhender la notion de contrôle des réactions émotionnelles.

Descartes rationalise l'étude des émotions par la description systématique des réactions émotionnelles. Il décrit celles-ci comme des mécanismes de réponse à des stimuli extérieurs, et annihile la notion de contrôle au profit d'un déterminisme rationnel [Descartes 1948]. Selon Le Brun [Brun 1698, Desjardins 2001], qui se fonde sur les théories de Descartes, notre cerveau, centre de la motricité et de la sensibilité, abrite la glande pinéale. Lorsque cette dernière est atteinte par des sensations, elle sensibilise l'âme et provoque des mouvements corporels déterminés.

Charles Darwin, en 1872, considère les émotions comme des vestiges de réactions prototypiques [Darwin 1965] dont les fonctions sont liées à la survie. Les réactions émotionnelles qu'il constate, sont issues d'une sélection naturelle effectuée lors du développement de notre espèce. C'est pourquoi ces réactions sont universelles et se retrouvent chez tous les êtres humains et même, chez certains de leurs ancêtres, les primates. Cette théorie possède un fort retentissement sur la manière dont les chercheurs représentent les émotions aujourd'hui (voir la section 2.2). En effet, l'universalité des réactions émotionnelles sous-jacente à la théorie darwinienne a conduit Ekman à définir un ensemble restreint d'émotions, dites *utilitaires* ou *primaires* [Ekman 1999a] : joie, tristesse, peur, dégoût, colère et surprise. Comme Le Brun,

cet ensemble d'émotions a été mis en évidence par l'observation poly-culturelle des réactions faciales associées [Ekman 1999b]. D'une certaine manière, Darwin bannit le contrôle personnel des réactions émotionnelles puisqu'il les présente en tant que "réflexes" systématiques, profilés par l'évolution.

Un autre courant de la psychologie, initié par William James, en 1884, considère aussi les réactions émotionnelles comme des réponses automatiques d'un organisme à son environnement, dans le but d'y survivre [James 2007]. L'apport de la théorie Jamesienne n'est pas tant sur les fonctions des réactions émotionnelles, mais plutôt sur les fonctions des émotions. En effet, il introduit une notion de causalité entre la réaction émotionnelle et l'émotion liée. Si notre corps réagit à l'environnement, de manière automatique par une réaction émotionnelle, à l'instar d'un réflexe, l'émotion, quant à elle, relève de la perception de cette réaction [Levenson 1990]. Cette différence subtile s'est avérée fondamentale puisqu'elle a entraînée de manière indirecte, la notion de processus d'évaluation chère à la psychologie cognitive.

En effet, des théories cognitives actuelles [Frijda 1986] considèrent que les émotions sont des évaluations ("appraisal") nous permettant d'interpréter l'adéquation de nos agissements à une situation. De là, naissent de nombreuses théories sur l'évaluation cognitive et ses primitives : certitude, agrément, responsabilité, effort [Lazarus 1991, Zajonc 1980] et contrôle. Scherer s'est particulièrement intéressé à la notion de contrôle. Il distingue deux effets responsables des variations observables dans la parole émotionnelle : c'est la théorie *push-pull* [Scherer 2006]. L'effet "push" (pousser) désigne tous les indices acoustiques liés à des changements physiologiques et somatiques (comme le jitter[3], par exemple) qui ne sont pas contrôlables. L'effet "pull" (tirer) est, quant à lui, contrôlable. Il est plutôt régi par des codes socioculturels. La théorie *push-pull* propose donc un cadre théorique permettant la séparation des effets contrôlés, de ceux qui ne le sont pas. Cette distinction se retrouve dans d'autres théories, sous d'autres terminologies comme spontané/acté ou encore spontané/symbolique [Buck 1984].

Le contrôle des réactions émotionnelles dans le but de satisfaire à des normes socioculturelles est aussi mis en avant dans la théorie social-constructiviste. Initiée par James Averill [Averill 1980] au début des années 1980, cette théorie se distingue fortement de ces prédécesseuses, par la négation du rôle des émotions dans la survie. En effet, les théories social-constructivistes placent la culture au centre de la formation des réactions émotionnelles et, plus largement, de l'organisation des émotions. Elles font donc de la société, l'élément fondateur des interactions émotionnelles, et par là, renforce l'hypothèse que nous contrôlons nos réactions émotionnelles (de manière collective). Le débat traditionnel entre la part de l'inné et celle de l'acquis semble donc être déterminant pour traiter la notion de contrôle sur les réactions émotionnelles. Si le radicalisme d'une telle théorie l'a rendue très controversée, certains pro-Darwininens, comme Ekman, admettent que des raisons socioculturelles peuvent tempérer/modifier nos réactions émotionnelles. Ils constatent effectivement

[3]Le jitter est la variation de la durée des périodes, de période à période, dans les segments voisés de la parole. Voir partie 3.2.4 pour plus d'explications...

des variations dans les manifestations de certaines émotions, dites *secondaires*, selon la culture.

La plupart de ces théories sur les états internes émotionnels et leurs fonctions, reposent sur l'observation de données émotionnelles externes. Les neurosciences permettent aujourd'hui l'observation in vitro d'un état émotionnel. Ainsi, la description des mécanismes des émotions et la recherche de leurs localisations (à l'instar de la glande pinéale de Le Brun) permettent de conforter certaines théories comme la théorie darwinienne. En effet, la mise en évidence de circuits neuronaux, non conscients, directs entre perception et action [LeDoux 2005], au niveau du système limbique, conforte les visions cartésiennes et darwiniennes, en ce qui concerne l'automatisme de nos réactions émotionnelles. A contrario, certaines théories plaident en faveur du contrôle sur les émotions et leurs réactions émotionnelles associées, puisque notre capacité à nous remémorer des situations peut agir sur nos états somatiques [Changeux 1983]. En effet, la théorie social-constructiviste, la théorie *push-pull* et la constatation d'Ekman semblent conforter l'idée que nous puissions exercer un contrôle partiel sur nos émotions, ou du moins, sur nos réactions émotionnelles.

Ce rapide aperçu des théories sur le contrôle des émotions, montre à quel point celles-ci sont basées sur l'observation de réactions émotionnelles. Ces dernières sont, à la fois, prédéterminées par des aspects physiologiques et modulées par notre parcours social (influence de la culture). Cela sous-entend que nous sommes capables, de manière collective ou individuelle, de plus ou moins contrôler certaines réactions émotionnelles, voire même de les simuler. C'est pourquoi l'adéquation entre l'émotion ressentie et l'émotion communiquée est difficile à évaluer. En d'autres termes, nous ne pouvons pas évaluer le degré de spontanéité d'une donnée émotionnelle.

Toutefois, l'absence de la connaissance de l'état interne du sujet d'observation est souvent compensée par la connaissance du contexte d'apparition de la réaction émotionnelle. Certains contextes favorisent ainsi l'apparition d'une émotion et de sa réaction émotionnelle associée, considérée alors comme "naturelle". Le degré de spontanéité de la réaction émotionnelle est alors apprécié suivant l'examen de son contexte d'apparition. Certains contextes favorisant l'apparition de certaines émotions semblent reproductibles. Satisfaisants alors la condition de reproductibilité, les chercheurs utilisent de tels contextes pour acquérir des données émotionnelles.

2.3.2 Méthodes pour l'acquisition de données émotionnelles

Plusieurs méthodes existent pour réunir des données émotionnelles. Ces méthodes sont généralement divisées, en trois sous-catégories : les données naturelles, les données induites et les données actées. Bien que ces méthodes semblent se distinguer par le degré de spontanéité des données qu'elles produisent, aucune d'entre elles ne fournit conjointement une mesure permettant d'évaluer si l'émotion est réellement vécue ou non par le sujet. Elles s'appuient sur le contexte d'apparition pour évaluer ce degré de spontanéité.

Pourtant, une distinction nette entre ces méthodes semble possible. En effet, les données émotionnelles sont dites "spontanées", dès lors que l'émetteur ne sait pas que l'objet de l'expérience réside dans l'acquisition de ces/ses réactions. Dans la terminologie des tests perceptifs psycho-physiques, un tel procédé emploie une méthode dite *indirecte*. Une méthode indirecte vise la mesure d'une variable produite par un sujet qui n'a pas la connaissance de cette mesure. Par exemple, le corpus E-Wiz [Aubergé 2004] est composé de réactions émotionnelles produites par des étudiants en situation d'apprentissage des langues. A contrario, une méthode directe implique des sujets qui connaissent la variable mesurée, comme des acteurs à qui l'on demande de simuler des émotions, par exemple. Dans cette distinction directe/indirecte des méthodes d'acquisition de données émotionnelles, on peut donc dire que les données actées sont issues de méthodes directes, tandis que les données naturelles et induites proviennent de méthodes indirectes.

2.3.2.1 Méthodes indirectes

Les méthodes indirectes sont aujourd'hui les plus utilisées par ceux qui désirent recueillir des réactions émotionnelles spontanées et non simulées/contrôlées.

Parmi elles, le recueil de données naturelles consiste à récolter des documents déjà existants, et à les réunir en corpus d'analyse. Certaines réactions émotionnelles apparaissent dans la vie courante et sont captées par différents matériels d'acquisition. Le contexte et les conditions d'enregistrement sont en général très variables. Ces données ne sont souvent pas reproductibles, compte tenu de la variabilité de leurs contextes d'apparition. Elles sont donc hétérogènes, ce qui rend difficile la constitution de grands corpus pour l'analyse statistique. C'est pourquoi ces données sont souvent analysées au cas par cas, à l'instar de Chung qui analyse des réactions apparues lors d'émissions de télé-réalité [Chung 2000].

De manière à rassembler des corpus conséquents et significatifs tout en en contrôlant la qualité, le recueil de données naturelles peut se faire en contexte prédéterminé. Des services (publics ou privés) sont proposés à des personnes. Les scénari limités de ces services et les raisons pour lesquelles les personnes les emploient, rendent ces derniers susceptibles d'être dans certains états émotionnels. C'est le cas des centres d'appels, par exemple, qui sont le terrain de situations sociales pouvant provoquer certaines émotions (colère, soulagement, inquiétude) [Devillers 2005]. Les réactions émotionnelles recueillies sont dites spontanées, car l'émetteur ne sait pas qu'il est enregistré, le plus souvent [Vidrascu 2005].

Toujours dans le souci de mieux contrôler le contexte d'apparition des réactions émotionnelles, les chercheurs ont crée des scénari artificiels, dans lesquels certaines réactions sont susceptibles d'apparaître spontanément. On parle alors de données induites. Une situation émotionnelle est artificiellement créée dans un contexte défini pour l'expérience. Les données sont alors recueillies dans de bonnes conditions d'enregistrement. Elles sont, là aussi, dites spontanées, car l'émetteur ne sait pas que l'objet de la situation réside dans l'acquisition de ces réactions émotionnelles [Aubergé 2004].

Toutefois, une personne en situation de laboratoire n'exprimera pas ses émotions de la même façon qu'elle les exprimerait dans sa vie de tous les jours. Les contraintes sociales amènent le sujet à un certain degré de contrôle de ses réactions émotionnelles. Aussi, une nouvelle génération de méthodes indirectes hybrides entre les données naturelles et induites voit le jour. Celles-ci reposent sur l'acclimatation du sujet à l'environnement de laboratoire ou aux contraintes d'enregistrement. Campbell a ainsi enregistré des conversations téléphoniques hebdomadaires entre deux personnes faisant connaissance [Campbell 2007a]. Si ces personnes savaient qu'elles étaient enregistrées, elles ne savaient pas, en revanche, que l'un des objets de l'étude était la mesure de l'évolution de leurs qualités vocales au fur et à mesure de leurs conversations. Cette méthode requiert une longue période d'analyse afin que les sujets "oublient" le contexte de laboratoire et les normes sociales associées.

Toutes ces méthodes indirectes visent l'acquisition de réactions émotionnelles spontanées. Elles varient selon le degré de connaissance/contrôle sur le contexte d'apparition de ces réactions. Plus le contexte est libre, plus les données sont considérées comme spontanées, mais plus elles sont hétérogènes. Au contraire, la détermination du contexte engendre une homogénéité parmi des données dont la spontanéité devient alors critiquable. En effet, à l'instar d'un acteur, tout un chacun exerce un contrôle sur ses réactions émotionnelles dans la vie courante, et peut les simuler sans qu'il soit pour autant le siège de l'état émotionnel associé. Ainsi, les données naturelles, dont le contexte est moins bien maîtrisé qu'en laboratoire, peuvent aussi provenir de simulations. Il en va de même pour les données induites où le contexte de laboratoire peut engendrer des modifications comportementales. En définitive, si le contexte d'apparition d'une réaction émotionnelle permet de faire des hypothèses quant à sa spontanéité, il ne permet pas, en revanche, d'établir la certitude scientifique que cette réaction est liée à l'état émotionnel correspondant. Si le niveau de connaissance du contexte d'apparition semble déterminant, il n'en reste pas moins que chacun semble potentiellement capable de simulation, de contrôle ou d'exacerbation de ses réactions émotionnelles. Cela est vrai pour toute les données émotionnelles externes, quelles soient issues de méthodes directes ou indirectes. De manière surprenante, cela est aussi vrai dans le cas des données actées (méthode directe), comme le montre la revue suivante des techniques employées par les acteurs[4].

2.3.2.2 Méthodes directes

Les méthodes directes emploient des personnes dont la mission est de fournir des réactions émotionnelles. Elles ont l'avantage d'une maîtrise parfaite du contexte, mais sont souvent critiquées pour le manque de spontanéité des données recueillies.

[4] L'étude scientifique de la mécanique des émotions, au sens de leurs dynamiques et de leurs interactions peux permettre d'isoler certaines séquences émotionnelles. Certains schémas semblent en effet redondants comme le montre la proximité de certains scénarios cinématographiques (blockbusters hollywoodiens, par exemple) ou la récurrence de formes musicales (sonates, par exemple). Ce faisant, ce type d'étude peut permettre d'en savoir plus sur la probabilité d'occurence d'une émotion et, ainsi, de maximiser les chances de son observation (voir chapitre 5).

En effet, les données produites par les acteurs sont toujours classées dans la catégorie "données actées" car elles sont considérées comme simulées et non spontanées. Pourtant, les acteurs et comédiens, pour exercer leurs professions, utilisent différentes techniques, leur permettant d'exprimer le mieux possible, les émotions et les états internes de leurs personnages. Parmi ces techniques, certaines d'entre elles reposent sur le principe de l'auto-induction d'états émotionnels. Nous relatons ici deux techniques majeures, afin de nuancer l'idée que les acteurs sont incapables de fournir des données émotionnelles spontanées.

"Actées I" : Traditionnellement, les "données actées" correspondent aux enregistrements de réactions émotionnelles externes entièrement simulées par un acteur, qui n'est pas, lui-même, le siège de l'émotion correspondante. La catégorie, nommée ici "actées I", regroupe les données produites par des acteurs dont la technique est basée sur la maîtrise de l'effet "pull". Usant de codes socio-culturels ou simulant/caricaturant certains traits des réactions émotionnelles spontanées, ces acteurs parviennent à restituer une réaction appropriée à une situation émotionnelle sans même l'avoir vécue auparavant. Ces techniques sont employées dans le théâtre "classique", de "boulevard" ou "Kathakali". Il y apparaît des réactions émotionnelles stéréotypées d'une grande "théâtralité". Ces formes culturelles facilitent notamment la compréhension de l'état interne du personnage, malgré des contraintes extérieures fortes (amplification des gestes vocaux, faciaux, posturaux due à la taille de la salle, le plus souvent...).

"Actées II" : La catégorie, nommée ici "actées II", regroupent les données produites par des acteurs, capables d'auto-induction de certains états émotionnels. Les techniques dites d'"Acteur Studio" ou de "la Méthode" sont toutes dérivées d'un courant initié par le metteur en scène Constantin Stanislavski, au début du XXème siècle. Ce système consiste en un entraînement de l'acteur à re-vivre, sur commande, des états émotionnels ou des sentiments déjà vécus et emmagasinés par sa "mémoire affective". Il s'agit de stimuler cette mémoire, ce matériau affectif, par le biais de la sensation, du souvenir ou de l'imagination, d'en réactiver les émotions recherchées, et de les utiliser pour nourrir le personnage à incarner. C'est donc à partir de sa propre matière humaine que l'acteur crée son rôle. Il n'est plus question de jouer, de "faire semblant", mais de vivre, ou de re-vivre sur la scène [Stanislavski 1966]. Le metteur en scène Peter Brook incite ainsi les acteurs à expérimenter des situations émotionnelles dans leurs vies personnelles, afin de construire un "dictionnaire" sensori-émotionnel propre, dans lequel ils peuvent réactiver certains états, pour une situation d'improvisation ou de performance scénique [5]..

[5] Il semble que cette dernière technique n'est pas encore été utilisée en laboratoire, ou du moins, que son usage n'ait pas été relaté. Les études ne mentionnent jamais si les acteurs impliqués utilisent une technique de simulation totale ou d'auto-induction. Or, ceux qui emploient cette dernière sont capables d'auto-induire des états émotionnels et par là-même, de fournir des données émotionnelles externes plus ou moins spontanées. De plus, ils peuvent, tout en maintenant le naturel de leur performance, maîtriser des variables nécessaires à l'étude scientifique, comme le texte qu'ils

2.3.3 Choix d'une méthode

Ce recensement des différentes méthodes d'acquisition de données émotionnelles, montre combien le degré de spontanéité des données produites, vis à vis des émotions vécues, est une notion déterminante. Or, aucune de ces méthodes ne mesure ce degré de spontanéité, de manière simultanée aux données qu'elle produit. Ce qui soulève une difficulté majeure quant à leur utilisation pour bâtir des théories de l'émotion. Toutefois, il est possible d'inférer un degré de spontanéité d'une réaction émotionnelle par l'observation du contexte d'apparition de celle-ci. La figure 2.3 recense les différentes méthodes d'acquisition présentées, dans un espace représentant, en ordonnée, le degré de spontanéité hypothétique des données qu'elles fournissent, et, en abscisse, la quantité d'information que nous connaissons sur leurs contextes d'apparitions.

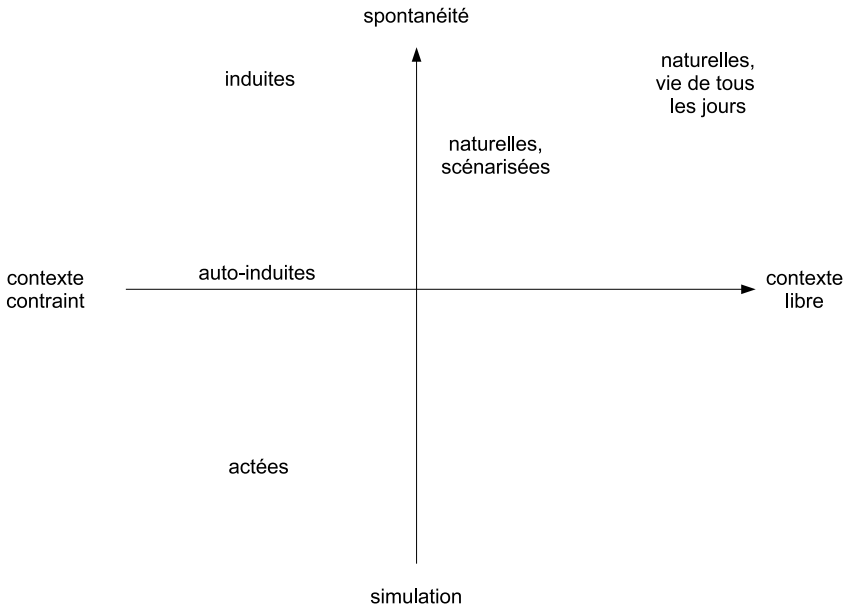

FIG. 2.3: Représentation des données émotionnelles provenant de différentes méthodes, dans un espace en deux dimensions. Abscisse : Degré de connaissance ou de contrôle sur le contexte d'acquisition. Ordonnée : Spontanéité des données émotionnelles produites.

Quelque soit la technique employée, celle-ci doit faire face au compromis suivant : produire des données émotionnelles semblant être les plus spontanées possibles, tout en assurant un contrôle sur la performance. Oscillant entre ces deux contraintes, certains acteurs sont parfois amenés à caricaturer certaines réactions émotionnelles ("Actées I"). Ces caricatures permettent alors aux spectateurs de mieux cerner l'état interne du personnage. Une étude intéressante de ce point de vue [Aubergé 2004], a

prononcent, par exemple. Le recueil de ces données peut aussi s'effectuer en laboratoire, dans de bonnes conditions d'enregistrement. Cette voie se révèle donc pertinente pour l'acquisition de données émotionnelles que l'on pourrait alors appelées "semi-spontanées"

montré que les réactions émotionnelles simulées menaient à de meilleurs taux de reconnaissance de l'émotion, que les réactions émotionnelles spontanées. Des acteurs, impliqués dans une expérience de mesure indirecte de leurs réactions émotionnelles, ont été amenés, par la suite, à simuler ces mêmes réactions. Ces dernières, parfaitement maîtrisées, semblaient contenir une information plus riche sur l'état interne de la personne, que les réactions spontanées, compte tenu des meilleurs taux de reconnaissance, à l'issu de tests perceptifs.

Compte tenu de cette remarque et puisque notre enjeu est la simulation du jeu d'acteur, nous préférons parmi toutes ces méthodes d'acquisition, une méthode directe impliquant des acteurs de la classe "Actées I". En effet, puisque l'on ne peut pas évaluer la spontanéité d'une donnée émotionnelle, et que dans certains cas, l'émotion communiquée est mieux perçue quand elle est jouée que quand elle est ressentie, alors il est préférable dans notre cas, d'utiliser des données actées. De plus, le contrôle sur leur contexte d'apparition permet une meilleure qualité d'acquisition. Enfin, les données peuvent être définies à l'avance et donc satisfaire à certaines contraintes comme l'usage d'un texte phonétiquement équilibré, par exemple.

2.4 L'expressivité de la parole

La difficulté d'évaluer le degré de spontanéité d'une donnée émotionnelle et le Paradoxe du comédien de Diderot montre à quel point il est nécessaire de dissocier l'émotion ressentie de l'émotion communiquée. C'est dans ce but que nous avons proposé une définition de l'expressivité dans la partie 1. Le corollaire de cette définition concerne l'expression neutre qui a pour avantage de posséder tous les autres niveaux d'information présents dans la parole :

- le message sémantique, porté par les aspects linguistiques
- l'identité du locuteur, portée par les caractéristiques de sa voix
- le style de parole emprunté
- les aspects pragmatiques relatifs à la parole en contexte
- l'expressivité

De manière à observer la variabilité qui n'incombe qu'à l'expressivité, il est nécessaire de contrôler ces autres niveaux d'information. Un aspect de la mis en place du corpus, consiste donc à fixer ou à faire varier ces autres variables dans le but de marginaliser leurs effets.

2.4.1 L'identité du locuteur

D'autres niveaux d'information sont véhiculés lors d'une situation de communication verbale. Tout d'abord, l'acte même de communiquer suscite un intervenant. Les différentes informations relatives à l'identité de cet intervenant sont alors sous-jacentes à la communication. Dans le cas de la modalité visuelle, un visage, une silhouette ou encore un avatar remplit cette fonction. Dans le cas de la "modalité textuelle", un nom, un pseudo ou une description sont souvent accessibles. Dans le cas de la modalité auditive, l'identité d'un locuteur est portée par sa voix. Son sexe, son âge, son état de santé, son origine (étrangère, accent régional, niveau social, niveau de langage ...) sont autant d'informations relatives à la personnalité d'un locuteur et rendues accessibles aux autres par sa voix.

2.4.2 Le style de parole

De nombreux facteurs contextuels peuvent affecter la manière de parler d'un individu. Tout d'abord, des contraintes physiques peuvent, par exemple, dans le cas d'un environnement bruyant, l'amener au forçage vocal ou à la répétition [Garnier 2007]. Des contraintes sociales entrent également en jeu. Par exemple, on chuchote près d'un bébé qui dort. Plus complexes, les contraintes professionnelles amènent à d'autres types de parole comme c'est le cas pour le discours politique ou l'enseignement. Enfin, d'importantes différences apparaissent entre la parole spontanée, la parole lue et la parole récitée "par cœur". Si tous ces différents facteurs externes influent grandement sur le style de parole, il n'en reste pas moins qu'ils ne sont pas des variables de l'expressivité. En effet, ces adaptations à des contraintes extérieures ne sont pas attribuables à une modification de l'état interne de leur locuteur. Qui plus est, le ou les interlocuteurs sont dans la connaissance de ces

contraintes et peuvent aisément en faire abstraction, ou au contraire, les prendre en compte. Dans tous les cas, le style de parole correspond à un niveau informatif de la communication, visant à décrire les situations externes et non les états internes, comme c'est le cas pour l'expressivité.

2.4.3 Le message sémantique

L'un des buts majeurs de la communication verbale réside dans l'échange d'un contenu sémantique, par l'usage des mots. Ces mots peuvent expliquer un état interne de manière directe ou de manière indirecte (métaphores dans le cas de la poésie). Ils reflètent donc un état interne. Mais ils sont rarement l'effet direct de cet état interne. De plus l'expressivité n'est pas dépendante d'un contenu sémantique, même si elle peut être explicitée verbalement ("tu me mets en colère."). L'expressivité possède d'autres supports, comme celui de la performance musicale (voir partie 6.5). De plus, les expressions peuvent être perçues par des personnes ne comprenant pas la langue de l'émetteur et, donc, le contenu sémantique prononcé [Burkhardt 2006], à l'instar des mimiques faciales universelles d'Ekman (voir chapitre 2.2.5). C'est pourquoi un même texte peut être prononcé par un acteur avec différentes expressions. Ce texte peut alors intégrer des marqueurs sémantiques de l'expressivité ou non. L'interaction et la possible incongruence entre ces marqueurs sémantiques et la façon dont ils sont prononcés, restent à confronter à la perception de l'expression d'un message parlé. Afin d'éviter ces éventuelles interactions, l'on peut choisir un texte dépourvu sémantiquement d'expressivité, comme c'est le cas pour les corpus présentés ultérieurement (voir chapitre 2.6.5). Un "texte dépourvu sémantiquement d'expressivité", aussi appelé "texte neutre", réfère à un texte pouvant être interprété avec tout type d'expression, indifféremment. Quoi qu'il en soit, malgré les interactions entre ces deux niveaux d'information, il n'en reste pas moins que le niveau d'information linguistique n'est pas le reflet d'un état interne potentiellement incontrôlable. C'est pourquoi nous le distinguons de celui de l'expressivité.

2.4.4 L'aspect pragmatique

Les aspects pragmatiques de la parole sont relatifs à la parole en contexte (un peu comme pour le style de parole). Par exemple, la modalité joue un rôle important dans la gestion du dialogue. La modalité se compose traditionnellement de l'assertion, de la question et de l'exclamation, et peut être marquée linguistiquement par la ponctuation (respectivement ".", " ?", et " !"). En situation dialogique normale, la modalité peut gouverner le tour de parole. A un point d'une discussion, elle reflète aussi la position du locuteur sur ce qu'il dit. Une assertion exprimée par une modalité interrogative sera considérée comme une question, tandis qu'une question exprimée par une modalité assertive n'engagera pas forcément l'interlocuteur à répondre. Ainsi, au fil de la parole, le locuteur exprime sa position, sa certitude ou son doute sur les mots qu'il prononce grâce à la modalité. Il peut alors

signifier le contraire de ce qu'il dit, cas généralement dénommé par : *ironie*. Il peut aussi exprimer le *doute* sur un propos, tout en l'exposant. Nous soumettons ici au lecteur, ces associations originales du ton ironique et du doute à la modalité. Ceci à cause de la forte relation existante entre le message linguistique et la façon dont il est exprimé. Ainsi, l'expression du doute ne fait pas partie de l'expressivité, mais il peut être accompagné d'un *embarras* ou d'une *confusion* qui traduisent un état interne probablement lié au doute.

2.4.5 L'expressivité

En ce qui concerne la parole, un acteur prononce un texte avec une certaine expression. Cette expression, ou l'information de l'état interne d'un personnage, est le plus souvent disponible via le canal linguistique. Les mots et les relations conceptuelles qu'ils provoquent peuvent, en effet, décrire des situations émotionnelles voire les engendrer. L'acteur est capable de déconceptualiser (au sens de "matérialiser") ces situations en jouant le texte avec l'expression appropriée. De plus, sa performance peut modifier ou ajouter, à ces situations, de l'expressivité selon la manière dont il prononce le texte. Enfin, il peut aussi créer, de lui-même, de l'information expressive à partir d'un texte sémantiquement neutre comme c'est le cas pour le corpus présenté dans la partie suivante.

Cette apparente division des niveaux d'information dans la communication verbale, n'exclut en rien la complexité provenant de leurs interactions. La relation entre le message sémantique et l'expressivité a été citée en exemple. De manière générale, le niveau d'information sémantique suffit souvent à lui seul pour convoyer tout ou partie des autres. Enfin, du point de vue de la perception, tous ces niveaux interagissent entre eux et avec les différents acquis de l'interlocuteur.

2.5 Corpus expressif : Combe2005

Le premier corpus expressif réalisé à l'IRCAM est nommé "Combe2005". C'est un corpus de test permettant d'affiner le protocole expérimental pour une campagne d'enregistrement plus conséquente. Cependant il a aussi permis d'obtenir certains résultats d'analyse. Il est constitué d'environ 1H30 de parole expressive et a été enregistré dans une chambre anéchoïque, en qualité CD (16bit/44KHz).

2.5.1 Support : Texte utilisé

Le texte prononcé est composé de 26 phrases de tailles variables :

1. *Bonjour.*
2. *Comment ?*
3. *Au revoir.*
4. *Assieds toi ici.*
5. *A toute a l'heure.*
6. *Quelle heure est-il ?*
7. *Je ne reconnais pas cet endroit.*
8. *On t'a dit ça, la veille au soir, par téléphone.*
9. *Tu m'obliges à parler, devant tout ce monde.*
10. *L'idéal est de trouver, un endroit assez élevé.*
11. *C'était bien avant, que je ne trouve du travail.*
12. *La veille du départ, ils ont dormi dans un hôtel.*
13. *Quelque chose, que tout le monde pourrait porter.*
14. *Ce ne sont que des lumières, et pas autre chose.*
15. *Tu veux me parler du travail, que j'avais avant d'être ici.*
16. *C'était dans l'enclos du fond, que ça devait se passer.*
17. *Une fois le rendez-vous fixé, il est impératif de s'y tenir.*
18. *Le mieux, pour regarder ce spectacle, c'est de s'asseoir.*
19. *Un costume qui ne soit, ni trop excentrique, ni trop neutre.*
20. *Ça fait un quart de page, avec une photo en noir et blanc.*
21. *Tout est déjà organisé, même la presse a été convoquée.*
22. *Dessous la photo, il y avait une légende, qui expliquait tout.*
23. *Il faut préserver une certaine dignité, autour de l'événement.*
24. *C'était le même nom, que la personne qui avait signé, cet article.*
25. *J'ai lu dans le journal, qu'on avait inauguré la statue, dans le village.*
26. *Il fallait prendre à gauche, au rond point principal, et ensuite, filer tout droit.*

2.5.2 Identité et style

Pour ce premier corpus, un acteur a été préféré à une actrice, notamment parce que la voix masculine se prêtait plus facilement aux outils d'analyse que la voix féminine (à l'époque ou le corpus a été enregistré), puisque les hommes possèdent généralement une fréquence fondamentale moyenne plus basse que celle des femmes. Le choix de l'acteur s'est dirigé vers une catégorie professionnelle dans laquelle le style de jeu n'est pas trop "marqué" (théâtre classique, conte...). Il s'est porté sur l'acteur, de catégorie "Actée I", Jacques Combe, qui est un comédien français d'une quarantaine d'années.

2.5.3 Expressivité

Ce texte est répété dans chacune des expressions suivantes :
- Neutre
- *colère*
- *joie*
- *peur*
- *tristesse*
- *ennui*
- dégoût
- indignation
- surprise positive
- surprise négative
- interrogation neutre

Pour les expressions écrites ci-dessus en italique, le texte est répété trois fois dans son intégralité, avec un degré d'intensité expressive croissant :
- *faible*
- *moyen*
- *fort*

2.5.4 Contenu du corpus

539 phrases ont été retenues après filtrage des mauvaises prononciations. De plus, l'acteur a éliminé les réalisations qu'il a jugé inadéquate vis à vis de l'expression désirée.

2.5.5 Avantages de ce protocole

La longueur variable des phrases du texte engendre une certaine variabilité prosodique, puisque le nombre de syllabes par groupe prosodique varie d'une phrase à l'autre. Plusieurs modalités sont aussi utilisées dans le même but : question, déclaration et exclamation. Les différentes expressions demandées sont présentées comme des catégories. Toutefois, la variation de l'intensité expressive autorise aussi

une représentation dimensionnelle des données (valence, intensité). L'enregistrement en chambre anéchoïque permet une qualité qui favorise les analyses acoustiques.

2.5.6 Inconvénients de ce protocole

2.5.6.1 longueur du texte

Un texte de 26 phrases est trop long pour ce genre d'expérience. En effet, le comédien éprouve des difficultés à se concentrer uniquement sur sa voix et son expressivité, et doit faire appel au texte avant chaque phrase de manière à ne pas se tromper. Une solution consiste en l'usage d'un texte plus court qui reste phonétiquement équilibré.

2.5.6.2 chambre anéchoïque

L'enregistrement a duré au total une demi-journée (6 heures) coupée de deux pauses de 30 min. Il s'est avéré particulièrement éprouvant pour l'acteur qui s'est dit surpris de son impuissance vocale dans la chambre anéchoïque. L'absence du retour de salle (effet de réverbération en situation écologique) a notamment été décrite comme particulièrement fatigante. De plus, 2×30 min de pause pour une demi-journée ne permettent pas de parer à la fatigue vocale qu'engendre un tel exercice. L'emploi d'un studio d'enregistrement semble donc plus adéquat pour le confort de l'acteur. L'idéal serait d'effectuer l'enregistrement sur le lieu de travail de l'acteur, s'il en possède un, ou tout au moins dans des conditions similaires à celles rencontrées dans l'exercice de sa fonction.

2.5.6.3 post-production

Les différentes combinaisons phrase/expression/intensité que l'acteur a dû gérer, ont été un obstacle dans la fluidité de son jeu. En effet, celui-ci a du se rappeler où il en était tout en jouant, ce qui a pu générer des répétitions et des ratés. La mise en place d'un système de communication entre la personne s'occupant de l'enregistrement (hors de la chambre anéchoïque) et l'acteur, a partiellement résolu le problème, bien que les indications de cette personne aient parfois déstabilisé l'acteur. Ces différentes erreurs ont provoqué des confusions dans les données qui ont nécessité une longue phase de post-production afin de les segmenter en phrases. L'emploi d'une interface d'aide à l'enregistrement semble souhaitable, afin de délester l'acteur d'une tâche de mémorisation supplémentaire, ainsi que pour permettre une segmentation et un étiquetage par phrase au fur et à mesure de la session d'enregistrement.

2.6 Corpus expressif : IrcamCorpusExpressivity

Fort de cette première expérience, une deuxième série de corpus a été enregistrée avec un nouveau protocole. Cette série de corpus [Beller 2008b], qui réunit les corpus Combe2006, Roullier2006, Olivia2006 et Daniele2006 a été baptisée : IrcamCorpusExpressivity.

2.6.1 Protocole

Chaque séance a été guidée par l'intermédiaire d'une interface informatique permettant la bonne conduite de l'enregistrement (texte, consigne, séquentialité). Cette interface possède un écran présentant la phrase, l'expression et l'intensité à réaliser. Le comédien déclenche et termine l'enregistrement grâce à une pédale. L'utilisation de cette interface-conducteur facilite le travail du comédien qui n'a plus à se concentrer sur la séquentialité des tâches. Cette interface facilite aussi la post-production puisqu'elle permet la synchronisation, l'annotation et la segmentation du corpus au fur et à mesure que celui-ci est enregistré. Ainsi l'acteur peut se tromper ou recommencer sans que cela n'entraîne de décalages. Les comédiens ont été enregistrés dans les mêmes conditions et dans un environnement qu'ils connaissent puisque il est leur lieu de travail. Le studio de doublage présente l'avantage d'une acoustique propre, tout en étant suffisamment réverbérante. Ainsi les acteurs ressentent moins de fatigue vocale que dans le cas de la chambre anéchoïque, qui possède une acoustique inhabituelle et particulièrement sèche. Un micro statique de qualité a permis l'acquisition des données en qualité ADAT (16 bit, 48000 Hz). Des données issues d'un laryngographe (EGG) ont aussi été enregistrées sur certaines parties du corpus.

2.6.2 Support : Texte utilisé

Le texte choisi pour ces corpus est plus court que celui utilisé précédemment. Il provient d'un corpus d'ensembles de dix phrases, phonétiquement équilibrés [Combescure 1981]. L'ensemble choisi parmi les 20 ensembles, regroupent des phrases sémantiquement neutres vis à vis de l'expressivité :

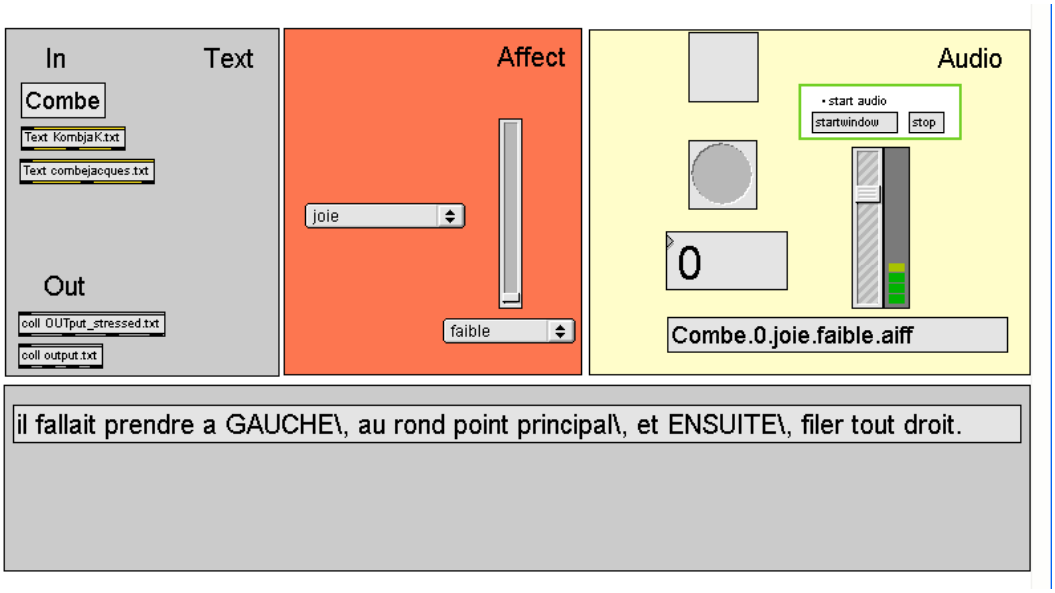

FIG. 2.4: Interface d'aide à l'enregistrement réalisée dans l'environnement Max/MSP.

$Stexte^{phrase}$

Description	:	Texte orthographique
Type	:	Symbolique et catégorielle
Unité	:	Phrase
Alphabet	:	"C'est un soldat à cheveux gris." "Alfred prit la tête de l'expédition." "Il ne pourra pas me voir si j'éteins la lampe." "Il entre avec sa chandelle, dans la vieille chambre." "Le nez du personnage s'abaisse, au-dessus de sa moustache." "Vous êtes vraiment obéissant !" "En attendant, c'est moi qui vais ouvrir." "Je ne pourrai jamais, me plier à son autorité." "Tout le monde sait que tu es la meilleure." "Je me demande, où se trouve cet endroit ?"
Cardinalité	:	10

L'accentuation de ces phrases a été annotée par la ponctuation et par des syllables en lettres capitales. Cela permet de varier les lieux de proéminence d'une phrase à l'autre, et donc la prosodie, malgré une longueur similaire. De plus, cela permet aussi de figer l'accentuation de la phrase, de manière à fixer la sémantique d'une répétition à l'autre (aspect pragmatique). Ces phrases ne perdent pas de sens, quelque soit l'expression avec laquelle elles sont prononcées. Elles sont donc "sémantiquement neutres" vis à vis de l'expressivité.

2.6.3 Identité et style

De manière à faire varier l'identité du locuteur (pour le respect de l'hypothèse N°4) et ainsi, d'observer les différences inter-acteurs, le corpus IrcamCorpusExpressivity est composé d'enregistrements de 4 acteurs : Jacques, conteur/comédien (environ 40 ans ; le même acteur que pour Combe2005), Philippe, comédien doubleur (environ 40 ans), Olivia, comédienne doubleuse (environ 25 ans) et Danielle, comédienne doubleuse (environ 50 ans).

$Sspeaker^{phrase}$

Description	: Nom de l'acteur
Type	: Symbolique catégorielle
Unité	: Phrase
Alphabet	: {"Combe", "Roullier", "Olivia", "Daniele"}
Cardinalité	: 4

Ils se distinguent notamment par le sexe, variable prise en compte pour obtenir une variation de l'identité du locuteur (et donc satisfaire à l'hypothèse N°4).

$Ssexe^{phrase}$

Description	: Sexe du locuteur
Type	: Symbolique et catégorielle
Unité	: Phrase
Alphabet	: {"homme", "femme"}
Cardinalité	: 2

Il a été formellement demandé aux acteurs de ne pas utiliser de style particulier (style théâtral ou autre).

2.6.4 Expressivité

De manière à pouvoir représenter les expressions enregistrées dans un espace dimensionnel dont les axes sont la valence (positif vs négatif), l'intensité (degré d'intensité de l'expression) et l'activation (introversion vs extraversion) [Schroeder 2003], et dans le but de fournir aux acteurs une description détaillée, sous la forme de vecteurs lexicaux, de l'expressivité (voir conclusion de la partie 2.2), nous leurs avons demandé d'exprimer des émotions utilitaires avec plusieurs degrés d'intensité et selon deux versions relatives à l'introversion et à l'extraversion. Les expressions choisies ont été désignées par le comité de pilotage du projet ANR-VIVOS[6] qui impliquait notamment des acteurs et de potentiels utilisateurs du système de transformation de l'expressivité (Studio de doublage Chinkel). Les expressions ainsi retenues sont :

[6]VIVOS : http://www.vivos.fr

$Sexpressivite^{phrase}$

Description	: Expressivité	
Type	: Symbolique et catégorielle	
Unité	: Phrase	
Alphabet	: “Neutre” :	état de référence
	“*colère introvertie*” :	colère contenue ou froide
	“*colère extravertie*” :	colère explosive ou chaude
	“*joie introvertie*” :	joie douce ou maternelle
	“*joie extravertie*” :	joie explosive ou enthousiaste
	“*peur introvertie*” :	peur contenue ou tétanisante
	“*peur extravertie*” :	peur explosive ou alarmante
	“*tristesse introvertie*” :	tristesse contenue
	“*tristesse extravertie*” :	tristesse explosive ou larmoyante
	“discrétion”	
	“dégoût”	
	“confusion”	
	“surprise positive” :	pour le locuteur
	“surprise négative” :	pour le locuteur
	“excitation”	
Cardinalité	: 15	

Pour ces dernières expressions (en caractère normal), les comédiens ont directement dit tout le texte avec le niveau d'intensité le plus fort possible. Pour les expressions en italique, le degré d'intensité a été varié selon 5 niveaux.

$Sdegree^{phrase}$

Description	: Degré d'intensité de l'expression
Type	: Symbolique et catégorielle
Unité	: Phrase
Alphabet	: {“1”, “2”, “3”, “4”, “5”, “6”}
Cardinalité	: 6

2.6.5 Contenu du corpus

Le déroulement de l'enregistrement est décrit par la procédure suivante. Pour une expression donnée, le locuteur lit la première phrase de manière neutre. Puis il répète 5 fois cette phrase, avec l'expression désirée, en accroissant son degré d'intensité. Ensuite il passe à la phrase suivante et recommence cette progression. Enfin, il réitère ce schéma avec les autres expressions. Cette procédure permet notamment d'obtenir une intensification de l'expressivité sans que le locuteur n'ait à relire le texte à chaque fois. D'une intensité à une autre, ni la phrase, ni son accentuation ne change, laissant apparaître seulement les variations imputables à l'intensité de l'expressivité. Les acteurs ont eu pour consigne explicite de ne pas varier la prononciation de leurs réalisations d'une même phrase. Ceci de manière à minimiser les variations dues aux phénomènes de coarticulation, de liaison et

d'élision. Dans le but de comparer les différentes réalisations, il leur a aussi été demandé de ne pas employer de restructurations et de sons non verbaux. Ces derniers ont été enregistrés séparément à la fin (car ils n'étaient pas l'objet direct de l'étude, au départ). Les fillers[7] suivants ont enregistrés avec chaque expression : *"Ah", "oh", "rire", "pleurs", "peur", "panique", "joie", "euh", "interrogation", "argh", "effort", "course", "hhh", "fff"*.

C'est sur ce dernier corpus que se fondent les analyses et les modèles de transformation qui sont présentés dans la suite de ce manuscrit. Les données collectées durant l'enregistrement consistent en un fichier audio pour chaque phrase et un fichier XML correspondant, contenant les annotations de l'expressivité (catégorie, activation et intensité), du texte déclamé, et des informations relatives à l'identité du locuteur (âge, sexe, nom). Au final, plus de 500 phrases ont été recueillies par acteur, formant un corpus d'une durée totale d'environ 12 heures de parole expressive.

[7] Le terme "filler", provenant de l'anglais, est largement employé dans le doublage de cinéma, pour décrire certains sons non verbaux.

2.7 Conclusion

Ce premier chapitre a permis de présenter les choix qui nous ont guidé dans la constitution de deux corpus expressifs. Tout d'abord, un mode de représentation de l'expressivité a été défini afin de servir de socle commun au chercheur, au comédien et à l'auditeur. Un aperçu des différents modes de représentation des émotions a permis l'établissement d'un nouveau mode de représentation des expressions sous la forme de vecteurs lexicaux. Puis, les différentes techniques d'acquisition de données émotionnelles existantes ont été recensées. La difficulté d'évaluer le degré de spontanéité d'une donnée émotionnelle, ainsi que l'enjeu de simuler le jeu d'un acteur, ont gouverné notre choix vers l'emploi de méthodes d'acquisition directes. Des acteurs-comédiens ont donc simulé un ensemble choisi d'expressions. Plusieurs corpus ont alors été enregistrés avec différents protocoles. Ces corpus de phrases constituant des paires d'expressions neutre-expressive ont notamment été définis de manière à isoler l'influence de l'expressivité sur la parole, de l'influence des autres niveaux d'information. C'est sur le dernier corpus, IrcamCorpusExpressivity, que vont se fonder les analyses et les modèles génératifs pour la transformation de l'expressivité qui vont être présentés dans les prochains chapitres.

Chapitre 3

Analyses du corpus

Sommaire

> ***En présence d'un être, on dirait que ce ne sont pas tellement les paroles qui comptent, mais leur musique.***
>
> G. Archambault, *Les Plaisirs de la mélancolie*, 1980

3.1 Résumé du chapitre

Ce chapitre est composé de deux parties complémentaires. La première partie permet d'appréhender le modèle utilisé pour représenter la parole. La deuxième fournit des résultats relatifs à l'application de ce modèle sur les corpus expressifs. La parole est un phénomène complexe qui combine plusieurs processus parallèles et séquentiels. De plus, l'appareil vocal permet de produire des sonorités qui ne sont pas (encore) traduites par la linguistique. Or, ces sonorités et la manière dont elles sont agencées semblent produire des marqueurs de l'expressivité. De la même manière, les restructurations possibles du discours comme les répétitions sont démonstratrices de certaines expressions. Enfin et surtout, l'expressivité dans la parole est véhiculée par la prosodie. Les analyses symboliques concernent les annotations phonétiques, paralinguistiques et de la proéminence. Les analyses acoustiques reflètent certaines tendances prises par les paramètres du modèle prosodique, vis à vis de l'expressivité. La mise en regard des analyses symboliques qui fournissent des unités discrètes et des analyses acoustiques qui produisent des données continues, permet l'analyse prosodique contextuelle. Cette partie fait office de préambule à la construction du modèle prosodique contextuel final, présenté dans le chapitre suivant.

3.2 Modèle de la parole

Afin d'observer l'influence de l'expressivité sur la parole, nous décrivons dans cette partie différents phénomènes que la communication verbale implique. Nous présentons ces phénomènes par le schéma de la figure 3.1 et les relions à la perspective du double codage de la parole, proposée par Fónagy [Fónagy 1983, Fónagy 1972a, Fonàgy 1972b].

3.2.1 Double codage de la parole

Le "double codage de la parole" proposé par Fonàgy [Fónagy 1983] différencie le canal linguistique du canal paralinguistique. Le canal linguistique est porteur du niveau d'information sémantique et peut être transcodé, sans perte d'information, en un texte. D'un point de vue acoustique, il est supporté par des séquences d'éléments sonores dits segmentaux, appelés phones. Ces phones sont des réalisations de phonèmes, qui constituent le dictionnaire symbolique fermé des sons d'une langue. Leur réalisation est très variable compte-tenu de plusieurs phénomènes segmentaux tels que la coarticulation, la liaison ou l'élision. Lors d'une prononciation intelligible, et pour quiconque en connaissance de ces précédents phénomènes et de la langue, chacun des phones est attribuable à la réalisation d'un mot, ou d'une suite de mots, permettant ainsi la compréhension du message linguistique (ou le décodage du canal linguistique).

La réalisation de ces phones est aussi influencée par des phénomènes suprasegmentaux. Ces phénomènes ont une portée dépassant le phone et n'affectent pas leur intelligibilité (c'est à dire qu'ils ne privent pas un phone de son appartenance à une catégorie phonétique). Ces phénomènes regroupés sous le nom de prosodie, sont les vecteurs du canal paralinguistique. Le canal paralinguistique véhicule les autres niveaux d'information que le niveau sémantique. Par exemple, l'intonation peut à elle seule, véhiculer la modalité et la proéminence.

D'un point de vue temporel, ces deux canaux d'information sont transmis simultanément. Les éléments segmentaux sont modulés par les phénomènes suprasegmentaux, définissant de concert, une forme acoustique dans laquelle, l'interlocuteur reconnaîtra les différents niveaux d'information attenant à la communication verbale (voir chapitre 2.4.5).

Si cette séparation permet de montrer une différence entre un texte et sa réalisation acoustique, il n'en reste pas moins que l'appareil vocal est susceptible de produire d'autres sons que les phones, même en situation de communication. Campbell relate qu'à peu près 30% d'un dialogue spontané enregistré est composé de sons non verbaux [Campbell 2007a]. Ces sons non verbaux sont aussi modulés par la prosodie, comme c'est le cas pour le rire (voir annexes D). De plus, bien que la syntaxe attenante au message linguistique apparaisse via le canal linguistique, de nombreuses restructurations non grammaticales entrent en jeu, surtout dans la parole spontanée, encore plus dans le dialogue spontané et d'avantage dans le cas de la parole expressive. C'est pourquoi nous introduisons ces phénomènes dans notre

modèle de la parole (voir figure 3.1).

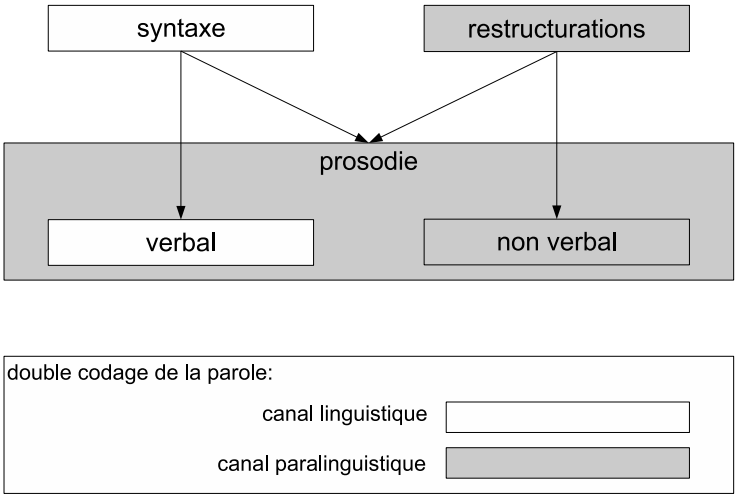

FIG. 3.1: Représentations de différents phénomènes de la parole. Les mots et la syntaxe constituent le canal linguistique. Les sons non verbaux, la prosodie et les restructurations sons les vecteurs du canal paralinguistique.

3.2.2 Mots verbaux et non verbaux

Un mot verbal possède un sens sémantique et une transcription linguistique. Il peut s'écrire grâce à l'usage d'un terme issu du dictionnaire des noms communs et des noms propres d'une langue. Il dépend donc de la langue et de normes socio-culturelles. Par opposition, un mot non verbal ne possède pas de transcription standard et est dépourvu de sens linguistique. Toutefois, il n'est pas rare de trouver des transcriptions phonético-orthographiques de ces sons comme "ah ah ah" ou "(rire)" pour écrire un rire dans un texte (de la bande-dessinée au roman, en passant par le script d'une pièce de théâtre). C'est à cause de cette fonction communicationnelle que nous parlons ici de mots et non de sons non verbaux.

Comme les mots non verbaux ne possèdent pas de transcription standardisée, ils sont difficilement verbalisables et qualifiables autrement que par reproduction. Malgré une grande variété, on distingue parmi les mots non verbaux, les "fillers" (rires, cris, pleurs,...), les pauses, les respirations (inspirations, reprises de souffle, expirations) [Beller 2006b] et d'autres bruits (gutturaux, nasaux...). Il semble que ces mots non verbaux soit de riches porteurs de sens pour l'expressivité [Schroeder 2006]. La tristesse peut être perçue seulement par un pleur et la peur, seulement par un cri, sans aucun autre mot verbal. Plus finement, une expérience perceptive informelle montre que le simple ajout local d'une respiration au milieu d'une phrase neutre peut changer l'expression perçue de toute la phrase (en l'occurence, l'expressivité perçue était la peur). Le pouvoir expressif

des mots non verbaux est tel, que les synthétiseurs de parole commencent à les générer (voir annexes D), de manière à accentuer le naturel et l'expressivité de la synthèse. Cela nécessite, entre autre, la définition de standard pour leurs transcriptions. Les récentes tentatives reposent en majorité sur des extensions du langage SSML[1] [Eide 2004, Blankinship 2001, Aubergé 2006].

3.2.3 Syntaxe et restructurations

La manière dont sont agencés temporellement les mots verbaux et non verbaux est informative. Ceci est bien connu dans le cas des mots verbaux dont l'agencement temporel est défini par des contraintes syntaxiques. Dans le cas d'une communication spontanée, ces mots peuvent toutefois ne plus respecter l'ordre régi par les règles grammaticales tout en conservant leur fonctions syntaxiques. En effet, la séquentialité entre sons non verbaux et verbaux forcent ces derniers à de possibles réorganisations temporelles, appelées *restructurations*. Ainsi, bien que la syntaxe du message linguistique organise a priori les mots et donc les séquences de phones, de nombreuses restructurations non grammaticales entrent en jeu, comme la répétition de syllables, de mots entiers ou bien même de propositions entières (resetting). En parole spontanée, la répétition qui est fréquente n'affecte pas forcément la compréhension des mots et de leurs relations syntaxiques. En revanche, elle peut révéler de l'hésitation ou de la confusion qui sont des expressions. D'autres restructurations sont porteuses de sens pour l'expressivité, alors qu'elles sont généralement considérées comme des disfluences pour la parole neutre [Piu 2007] et concerne la prononciation : la coarticulation, la césure, la liaison et l'élision en sont des exemples.

3.2.4 Prosodie

Le flux de parole est donc une séquence de mots verbaux et de mots non verbaux organisée par la double action des règles syntaxiques et des possibles restructurations. Dans le même temps, la réalisation acoustique de tous ces éléments sonores est modulée par la prosodie. Si ceci est bien connu en ce qui concerne les mots verbaux, cela reste vrai pour les mots non verbaux comme le rire, par exemple (voir annexes D). La prosodie comprend des traits phonologiques supra-segmentaux dont la portée dépasse l'horizon du phone (la syllable, le groupe accentuel, le mot, la clitique, le groupe de souffle, le groupe prosodique, la phrase...) et qui n'annihilent pas leur intelligibilité (c'est à dire qu'ils ne privent pas un phone de son appartenance à une catégorie phonétique). Cinq traits caractéristiques sont généralement cités dans la littérature comme les cinq dimensions de la prosodie [Pfitzinger 2006] :

- l'intonation : fréquence fondamentale, hauteur, pitch
- l'intensité : énergie, volume
- le débit de parole : vitesse d'élocution
- le degré d'articulation : prononciation, configurations du conduit vocal, dynamique des formants

[1] SSML : Speech Synthesis Markup Language : http://www.w3.org/TR/speech-synthesis/

– la phonation : excitation glottique, qualité vocale (voix pressée, normale, soufflée), mode vibratoire (fry, normal, falsetto), fréquence de voisement...

Durant un demi-siècle d'étude de la parole neutre, la prosodie a souvent été réduite à l'intonation. l'intonation a ainsi bénéficié de beaucoup d'attention et de modélisation, car, aisée à observer, elle a permis à elle seule, de faire émerger des fonctions de la prosodie (modalité, emphase...). Le cas de la parole expressive semble nécessiter plus fortement l'observation des autres dimensions [Campbell 2003]. Enfin, de part son caractère continu dans le temps, la prosodie accompagne la production de sons verbaux et non verbaux et interagit donc avec la syntaxe et les restructurations.

3.3 Modèle prosodique

Plusieurs modèles prosodiques ont été proposés dans la littérature. La majorité tente de relier des phénomènes continus (valeurs de la f0), après stylisation, à des catégories discrètes. Le modèle prosodique proposé dans cette thèse concerne la stylisation de paramètres prosodiques en fonction de catégories linguistiques. Ces modèles appliqués à des données réelles seront utilisés ultérieurement pour comparer la prosodie de plusieurs expressions. Cette partie montre les différentes analyses et les algorithmes de modélisation mis au point pour caractériser les cinq dimensions prosodiques que sont :

- l'intonation
- l'intensité
- le débit de parole
- le degré d'articulation
- la phonation

Une phrase d'un corpus est utilisée dans cette partie à titre d'exemple. Elle a été prononcée par l'acteur *Combe* avec une *joie introvertie* d'intensité 1 (voir chapitre 2.6.5). Elle correspond au texte "Il entre avec sa chandelle, dans la vieille chambre".

3.3.1 Unités/Groupes prosodiques

La définition d'un modèle prosodique dépend fortement de l'unité choisie (phrase, groupe de souffle, mot, mélisme, syllable...). La plus petite entité temporelle du modèle proposé est la syllable. Mais d'autres unités plus larges possèdent une influence sur les paramètres estimés des syllabes.

3.3.1.1 Phrase

L'aspect pragmatique et la modalité permettent au locuteur de situer sa position par rapport à ce qu'il dit. Le ton montant est traditionnellement attribué à la question, tandis que le ton descendant reflète généralement une affirmation (bien qu'il existe de nombreuses discussions à ce sujet). Ainsi, le ton indicatif de la modalité apparaît souvent en fin d'une phrase, mais il peut aussi être présent à un autre endroit ou bien réparti sur toute la phrase. Afin de le prendre en compte, il est donc nécessaire d'observer la phrase entière, qui devient de ce fait une unité d'observation. Parmi les modalités usuellement reconnues, on compte l'affirmation, l'exclamation et l'interrogation. Nous avons soumis l'idée (voir partie 2.4) d'y ajouter l'ironie et le doute. Ces 5 modalités forment le dictionnaire de la variable $Smodalite^{phrase}$.

$Smodalité^{phrase}$

Description	: Modalité d'une phrase
Type	: Symbolique et catégorielle
Unité	: Phrase
Alphabet :	{ "affirmation", "interrogation", "exclamation", "ironie", "doute"
Cardinalité	: 5

Une phrase courte est généralement déclamée dans une seule expiration, durant un seul cycle respiratoire. Lors de l'expiration, la pression sous glottique diminue progressivement. Il en résulte une déclinaison de la f0 [Gussenhoven 1988, Grobet 2001, Lacheret-Dujour 1999]. Cette déclinaison est réinitialisée à chaque prise de souffle. Lors d'une phrase plus longue, des prises de souffle peuvent intervenir et on observe des réinitialisations de la déclinaison dans la phrase. C'est pourquoi il est nécessaire de segmenter les phrases en groupe de souffle.

3.3.1.2 Groupe de souffle

Un groupe de souffle est défini entre deux inspirations. Il est donc constitué généralement d'une inspiration silencieuse et rapide, suivie d'une seule expiration. Après l'inspiration, une réinitialisation de la déclinaison de f0 peut apparaître (phénomène appelé "resetting" en anglais). L'unité groupe de souffle est pertinente pour modéliser l'expression car cette dernière peut-être accompagnée de modifications physiologiques ayant des conséquences sur la respiration. Ainsi, les modifications de la respiration induites par l'état interne responsable de l'expression, peuvent entraîner différents groupements en groupes de souffles. Dans le cas neutre, ces regroupements sont uniquement dépendants du débit de parole [Fougeron 1998]. Ces groupements ne dépendent pas seulement du contenu sémantique et il est important de les prendre en compte dans l'étude de l'expressivité.

3.3.1.3 Syllabe

La syllable joue un rôle primordial dans l'acquisition, la production et la perception de la parole. Lors de l'acquisition du langage, les premiers mots acquis sont généralement, mono ou bi syllabiques ("papa"). Il n'est pas possible de prononcer d'unité inférieure à la syllable. Dans cette acception, une voyelle prononcée seule est considérée comme le noyau d'une syllable dépourvue d'attaque et de rime. La syllabe joue aussi un rôle important dans la perception de la parole. Une caractéristique prosodique importante de la syllabe concerne le phénomène de proéminence.

Proéminence La proéminence est le résultat perceptif d'un contraste (culminance, distinction, démarcation) acoustique contrôlé, remplissant plusieurs fonctions. Tout d'abord, elle manifeste l'accentuation de certaines syllabes qui peut

être définie par des règles linguistiques [Caelen-Haumont 2004]. Cette accentuation dépend donc de la langue et on distingue classiquement deux grandes familles de langues : Les langues à accent déterminé comme l'anglais et les langues à accent libre comme le français. La proéminence joue parfois un rôle dans la désambiguïsation du sens (accent pragmatique). Enfin, la proéminence sert aussi à mettre en valeur certains éléments (accent de focus, d'emphase, d'insistance). Par ce rôle multi-fonctionnel, et parce que la proéminence résulte des interactions entre les différents niveaux d'information de la parole (niveau du sens sémantique, du style, de l'aspect pragmatique, de l'expressivité et de l'identité, voir chapitre 2.4.5), il semble nécessaire de prendre celle-ci en compte dans le modèle prosodique.

En effet, la proéminence joue un rôle particulier dans la communication verbale, car sa réalisation requiert une part plus importante de contrôle, lors de sa production. Cette même proéminence sera interprétée par l'interlocuteur comme un marqueur local de l'importance du message. D'une certaine manière, la théorie "push-pull" peut être, ici, déclinée dans une version locale et dynamique. Les syllabes non proéminentes manifestent plutôt l'effet "push", tandis que les syllabes proéminentes, plus maîtrisées, voient leurs réalisations marquées par l'effet "pull". Le caractère incontrôlable d'un supposé état interne se manifeste par une expressivité plus ou moins contrôlée (voir chapitre 5). La notion de proéminence d'une syllabe prend donc une part importante pour l'étude de l'expressivité, c'est pourquoi elle est introduite dans le modèle. La variable $Sproeminence^{syllable}$ permet de catégoriser les syllabes selon différentes catégories de proéminence perçue. Avec Anne Lacheret, Nicolas Obin et Jean-Philippe Goldman, nous avons déterminé les cinq catégories suivantes, parce qu'elle sont perceptiblement distinguables [Obin 2008] :

$Sproeminence^{syllabe}$

Description	: Proéminence d'une syllabe	
Type	: Symbolique et catégorielle	
Unité	: Syllabe	
Alphabet	:	"UN" : Proéminence non définie (silence par exemple) "NA" : Non proéminence "AS" : Proéminence secondaire "AI" : Proéminence intentionnelle "AF" : Proéminence finale
Cardinalité	: 5	

Enfin, les pauses intra-phrases, silencieuses ou non, de part leurs durées similaires à celles des syllabes, sont considérées comme des syllabes.

3.3.1.4 Phone

Le phone est une unité inférieure à la syllabe et liée au phonème. Si l'on peut la voir comme la "lettre" de la parole, son existence reste à définir. Ainsi, la tâche de segmentation phonétique est parfois rude car un phone n'existe pas localement alors qu'il est perceptible dans une séquence. Si le phone constitue une unité pratique car il est de petite taille et parce qu'il est censé appartenir à un dictionnaire

réduit de symboles (une trentaine de phonèmes), il s'avère que son utilisation pose des problèmes pratiques. C'est pourquoi l'utilisation du phone dans l'étude prosodique est réduite au traitement du degré d'articulation, bien que la manière dont ils sont prononcés soit aussi influencée par l'expressivité (voir chapitre 5. L'alphabet phonétique utilisé est le Xsampa, dérivé de l'API.

Sphoneme^{phone}

Description	: Phonème du phone	
Type	: Symbolique et catégorielle	
Unité	: Phone	
Alphabet	:	"e∼", "9∼","a∼","o∼" : voyelles nasales
		"i", "e", "a", "o", "u", "y", : voyelles orales fermées
		"E", "A", "O", "2", "9" : voyelles orales ouvertes
		"@" : schwa
		"j", "w", "H" : glides
		"p", "t", "k" : occlusives non voisées
		"b", "d", "g" : occlusives voisées
		"f", "s", "S" : fricatives non voisées
		"v", "z", "Z" : fricatives voisées
		"l", "R" : liquides
		"m", "n", "N" : consonnes nasales
		"###" : silences de début/fin
		"##" : pauses
		"*" : fillers, paralinguistiques...
Cardinalité	: 38	

3.3.2 Intonation

L'intonation est un paramètre prosodique primordial, à qui on a prêté de nombreuses fonctions linguistiques. L'intonation est relative à la variation de la fréquence fondamentale, nommée f0. Celle ci est relative à la vitesse de vibration des plis vocaux et n'existe que pendant la phonation. La f0 n'est donc estimable que sur les parties voisées du signal de parole. L'algorithme YIN [Cheveigné 2002] est utilisé pour estimé la f0 de la voix à partir d'un enregistrement audio. Quelques étapes de post-processing permettent d'obtenir une courbe de f0 qui correspond mieux à notre perception de l'intonation. Tout d'abord, les valeurs de la f0 estimées par l'algorithme sur des parties non voisées sont mises de côtés. Puis, un filtrage médian permet de réduire les fréquentes erreurs d'octave. Enfin, et ceci afin de produire des modèles de l'intonation indépendants du texte prononcé, l'algorithme interpole la f0 entre les segments voisés. L'intonation est donc représentée par une courbe continue évoluant le long de la phrase et appelée : $Af0^{phrase}_{interp}$ (courbe rouge de la sous-figure 1 de 3.5).

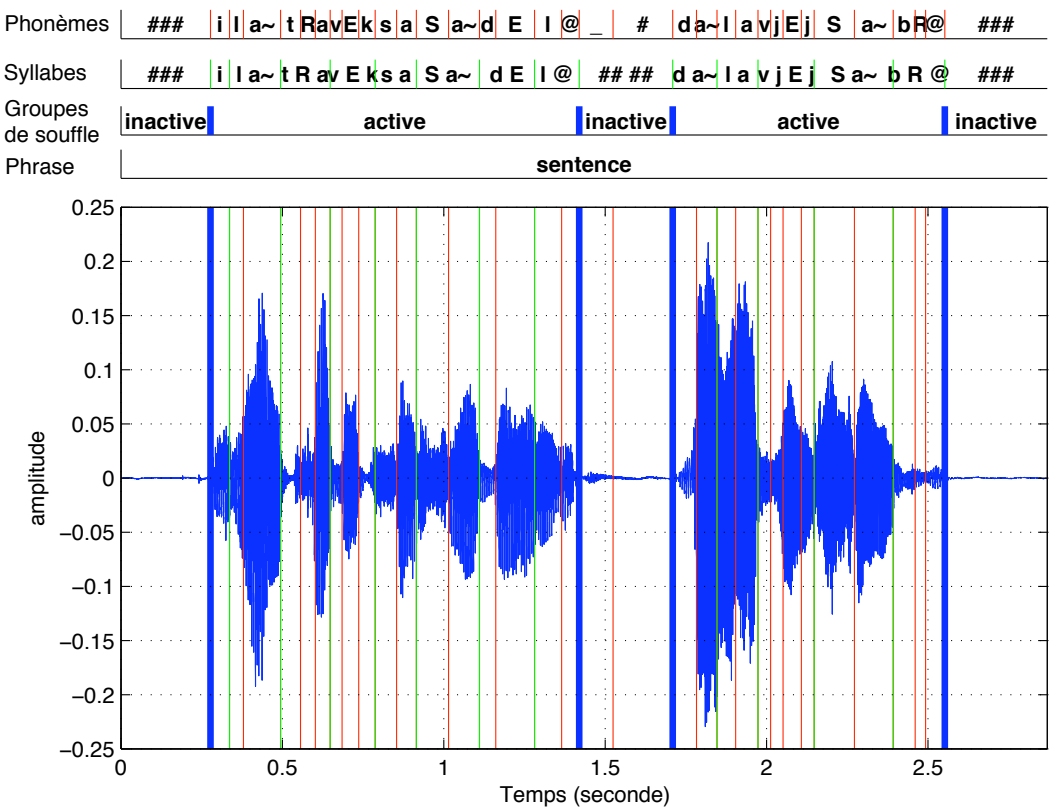

FIG. 3.2: Exemple de segmentation en plusieurs niveaux prosodiques de la phrase : “Il entre avec sa chandelle, dans la vieille chambre”.

3.3.2.1 Stylisation

Le terme de stylisation indique une forme simplifiée d’un paramètre prosodique continu [2], qui est censé préserver les phénomènes fonctionnels et audibles. La stylisation permet une réduction du nombre de paramètres du modèle et donc de sa complexité. Différents modèles existant dans la littérature sont présentés ci-dessous.

Modèles d’ordre 0 Un modèle d’ordre 0 stylise la courbe réelle de la f0 par des points (voir figure 3.4). Peu de modèles d’ordre 0 existent dans la littérature car ils ne sont pas assez complexes pour décrire la dynamique de l’intonation, élément qui semble capital dans la perception de la prosodie. Cela a été montré par une expérience informelle réalisée lors du séminaire littérature-musique, dirigé par François Nicolas et Michel Chaillou, organisé grâce à la collaboration du CDMC et du CNSMDP et qui a eu lieu le Samedi 17 mars 2007 au Conservatoire de Paris. Les participants, les poètes Michel Chaillou, Valère Novarina, Jacques Réda, Olivier Cadiot, Philippe Beaussant et Michel Deguy ont expliqué le rôle de la musique dans leurs œuvres et ont chacun déclamé un extrait choisi, d’une durée de trois

[2] f0 est l’exemple pris par la suite, mais d’autres descripteurs comme la “loudness” sont stylisés de la même façon

minutes. Ces extraits ont été enregistrés durant la journée, puis ont été mêlés à des extraits de voix de poètes connus (Arthaud, Claudel, Eluard, Jouve et Malraux, extraits du CD « voix de poètes »). Ensuite, à la fin de la journée, ces extraits ont été présentés au public d'une manière détournée, après transformation par un programme de musicalisation de la prosodie en temps réel.

Ce programme (voir 3.3) édite une partition à partir de la prosodie de la voix parlée, par segmentation automatique et estimation de notes. Une note est déclenchée au voisinage de chaque centre de stabilité acoustique. L'algorithme simplifié pour tourner en temps réel comprend une hiérarchie d'experts prenant en compte les signes des dérivées première et seconde de la f0 filtrée (estimée par l'algorithme Yin [Cheveigné 2002]), les signes des dérivées première et seconde de l'intensité filtrée et les signes des dérivées première et seconde d'un coefficient relatif au voisement (apériodicité). Ainsi, une note est déclenchée à peu près pour chaque syllabe, au milieu du noyau vocalique où la hauteur peut être relativement stable. La note produite possède une hauteur correspondant à la f0 arrondie sur un tempérament choisi, une vélocité relative à l'intensité en dB et une durée correspondant au temps entre deux "onset"). Si la correspondance note ↔ syllabe parait évidente par leurs durées similaires et par les mêmes notions structurelles d'attaque (attaque), de rime (tenue) et de coda (relâche), il n'en va pas de même pour la correspondance note ↔ intervalle entre noyaux. Malheureusement, la contrainte du temps réel ne permet pas de fournir une note, dès l'attaque de la syllabe, avec une hauteur relative à une f0 qui n'existera qu'après, dans le noyau vocalique. Toutefois, on peut considérer que le retard systématiquement introduit (d'une durée d'environ l'attaque et le début du noyau vocalique) est à peu près constant et qu'il n'influe donc pas sur le rythme perçu.

L'atelier ludique présenté à la fin de la journée consistait à demander au public et aux poètes, de reconnaître les prosodies de ces derniers, à partir des seules lignes mélodiques jouées par un piano (instrument volontairement lointain de l'instrument vocal). Seule les prosodies de Valère Novarina et de Paul Eluard ("Liberté") ont été reconnue à l'unanimité. Cette expérience conclut de manière informelle, à la nécessité de modéliser la dynamique de l'intonation par une élévation de l'ordre du modèle qui ne peut se résoudre à des paliers de notes successives.

Prosogramme Afin de tenir compte de l'aspect dynamique de l'intonation, il faut donc un modèle d'ordre supérieur à 0. Les modèles d'ordre 1 approximent la f0 par une suite de segments (voir figure 3.4). Un exemple est donné dans la figure 3.4. Le prosogramme [Mertens 2004] est un algorithme qui repose en partie sur un modèle de la perception tonale [D'Alessandro 1995] et qui nécessite une segmentation phonétique. Il introduit donc des hypothèses dans le procédé de stylisation. Pour chaque syllabe, le noyau vocalique est délimité comme la partie voisée qui présente une intensité suffisante (en utilisant des seuils relatifs au pic d'intensité local). Puis, pour chaque noyau, la F0 est stylisée en un ou plusieurs segments de droite, dans le domaine log-f0 (exprimé en demi-tons), définissant des tons

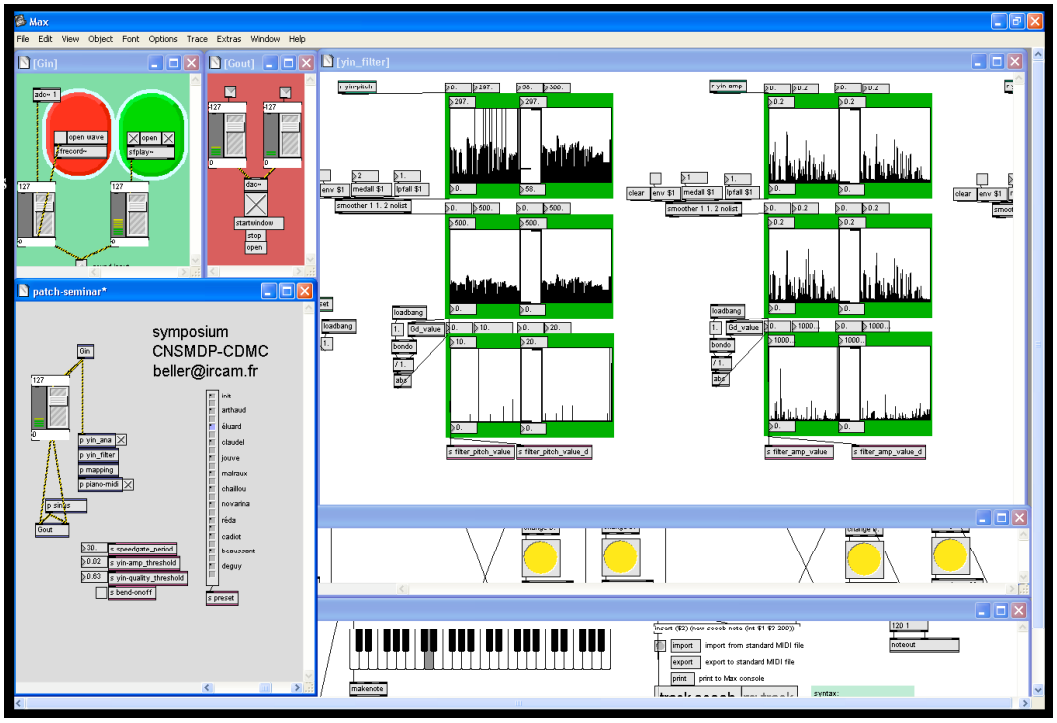

FIG. 3.3: Patch Max/MSP de musicalisation de la prosodie. En temps réel, une phrase parlée est analysée et transformée en notes musicales. Le programme fournit une partition ou commande un instrument midi.

syllabiques . Ces segments peuvent être stylisés comme plats ou avec une pente mélodique, selon des seuils perceptuels de glissando qui sont réglables.

MoMel MoMel pour MOdeling MELody [Hirst 1993] est un algorithme de stylisation automatique de l'intonation. L'algorithme MoMel repose sur l'acceptation de l'hypothèse suivante : la courbe mélodique peut être approximée par morceaux par une fonction quadratique (polynôme de degré 2). La nature relativement neutre de cet algorithme a permis son utilisation large. Il est ainsi utilisé comme instrument préliminaire de modélisation dans la formation de nombreuses représentations : modèle de Fujisaki [Mixdorff 1999], INTSINT [Hirst 1993, Hirst 2000], mais aussi ToBI [Maghbouleh 1998, Wightman 1995] après une réduction de l'ordre. En effet, l'usage de fonctions quadratiques place l'algorithme MoMel parmi les méthodes de stylisation de la f0 d'ordre 2 (voir figure 3.4).

Modèles log-quadratique : Fujisaki Le modèle de Fujisaki [Fujisaki 1981] est, en partie, fondé sur des considérations physiologiques. Il considère que le logarithme de la f0 correspond à la réponse d'un système masse-ressort constitué des cordes vocales, à un ensemble de commandes linguistiques. D'après l'observation physique de l'interaction entre le larynx et les cordes vocales, le système est modélisé par un

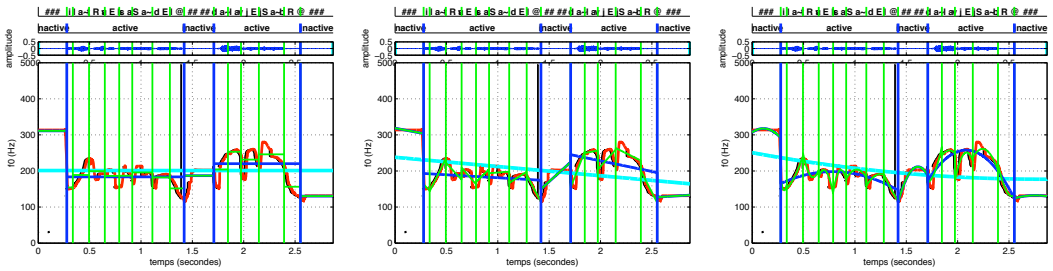

FIG. 3.4: modélisation d'ordre 0, 1 et 2 pour les unités de type phrase (bleu ciel), groupe de souffle (bleu foncé) et syllabe (vert).

système linéaire du second ordre. Tout comme pour le modèle Tilt [Taylor 2000], la stylisation de f0 correspond donc à un modèle d'ordre 2 sur lequel on applique la fonction logarithme (voir figure 3.4).

Modèles à base de splines Des méthodes d'ordre variable ou supérieur à 2 consistent à styliser automatiquement la courbe de f0 par des B-splines [Lolive 2006]. Dans le domaine mathématique de l'analyse numérique, une spline est une fonction définie par morceaux par des polynômes. La complexité du modèle dépend du nombre de nœuds (morceaux) et de l'ordre des polynômes. Cette complexité peut être définie de manière automatique par des méthodes visant à minimiser la distance entre la courbe réelle et la courbe stylisée (moindres carrés le plus souvent) tout en minimisant la complexité du modèle (critère BIC, AIC, MDL...).

Modèles superpositionnels Certains modèles proposent une modélisation de l'intonation multi-paramétrique [Bailly 2005]. Dans ces modèles, plusieurs niveaux prosodiques sont susceptibles d'influencer l'intonation. Ils reposent pour la plupart sur un principe d'indépendance entre les niveaux prosodiques [Morlec 1997]. Ainsi, ces modèles combinent par addition, multiplication ou modulation, différents effets attenant à différentes fonctions linguistiques qui reposent sur différents niveaux prosodiques, pour modéliser l'intonation résultante.

Modèle utilisé : ProsArchy Le modèle choisi pour la stylisation de f0 est un modèle tirant partie des modèles présentés précédemment. Il a été nommé ProsArchy car c'est un modèle hiérarchique additif de contours intonatifs log-quadratiques. L'algorithme utilise la segmentation phonétique pour définir les unités prosodiques : phrase, groupe de souffle et syllabe. Puis, de manière récursive, il convertit la courbe de f0 réelle en log, il estime un modèle quadratique sur unité "parent", retranche à la courbe log-f0 réelle la partie apportée par cette modélisation, et recommence avec les unités "enfants". Chaque unité de la phrase à styliser est modélisée par un contour log-quadratique, quelque soit sa durée. Les paramètres de ces polynômes convertis en polynômes de Legendre sont les paramètres de notre modèle. Les polynômes de Legendre, notés ici L_n sont des polynômes de degré n, vérifiant $L_n(1) = 1$ et définis

sur l'intervalle $[-1, 1]$ orthogonaux pour le produit scalaire. L'orthogonalité de cette famille permet une graduation de la complexité du modèle au fur et à mesure que l'on ajoute des coefficients. L'estimation d'un modèle quadratique donne trois coefficients de Legendre. Le premier coefficient $P0$ est relatif à l'ordre 0 et correspond à la moyenne ; Le second coefficient $P1$ est relatif à l'ordre 1 et correspond à la pente d'un segment de droite ; Le troisième coefficient $P2$ est relatif à l'ordre 2 et correspond à la courbure.

$$\begin{cases} P0(x) = & 1 \\ P1(x) = & x \\ P2(x) = & \dfrac{3x^2 - 1}{2} \end{cases} \tag{3.1}$$

La stylisation hiérarchique est réalisée par l'algorithme suivant (voir figure 3.5

pour des exemples) :

→Initialisation
- Estimation de la fréquence fondamentale $Af0_{reel}^{phrase}$ avec YIN [Cheveigné 2002] (courbe noire de la sous-figure 1 de 3.5);
- Binarisation de l'apériodicité par seuil (critère voisé / non voisé);
- Retraits des valeurs de $Af0_{reel}^{phrase}$ sur les parties non voisées;
- Filtrage médian des valeurs gardées (pour éviter les sauts d'octave);
- Interpolation quadratique de la f0 : $Af0_{interp}^{phrase}$ (courbe rouge de la sous-figure 1 de 3.5);
- Construction des unités de type phrase (phs), groupe de souffle (gds) et syllabe (syl) à partir de la segmentation phonétique;

→Niveau phrase
- Estimation du modèle log-quadratique relatif à la phrase : $Af0_{model}^{phs}$ (courbe cyan de la sous-figure 1 de 3.5);
- Soustraction de ce modèle à la courbe réelle : $Af0_{residuel}^{phs} = Af0_{interp}^{phs} - Af0_{model}^{phs}$ (courbe rouge de la sous-figure 2 de 3.5);
- Descente hiérarchique au niveau des enfants de type groupe de souffle, par les liens de parenté;

→Niveau groupe de souffle

foreach $g \in G$ **do** G est l'ensemble des groupes de souffle de la phrase

- Segmentation de $Af0_{residuel}^{phs}$ sur le groupe de souffle ;
- Estimation du modèle log-quadratique relatif au groupe de souffle : $Af0_{model}^{gds}$ (courbe bleue de la sous-figure 2 de 3.5);
- Soustraction de ce modèle à la courbe réelle : $Af0_{residuel}^{gds} = Af0_{residuel}^{phs} - Af0_{model}^{gds}$ (courbe rouge de la sous-figure 3 de 3.5);
- Descente hiérarchique au niveau des enfants de type syllabe, par les liens de parenté;

→Niveau syllabe

foreach $s \in S$ **do** S est l'ensemble des syllabes du groupe de souffle

- Segmentation de $Af0_{residuel}^{gds}$ sur la syllabe ;
- Estimation du modèle log-quadratique relatif à la syllabe : $Af0_{model}^{syl}$ (courbe verte de la sous-figure 3 de 3.5);
- Soustraction de ce modèle à la courbe réelle : $Af0_{residuel}^{syl} = Af0_{residuel}^{gds} - Af0_{model}^{syl}$ (courbe rouge de la sous-figure 4 de 3.5)

$Af0_{residuel}^{syl}$ est le résiduel de l'opération de stylisation ;

end

end

Algorithme 1 : Algorithme de stylisation de l'intonation ou de la f0.

Voici les variables du modèle résultant de cette stylisation et qui sont utilisées pour la transformation.

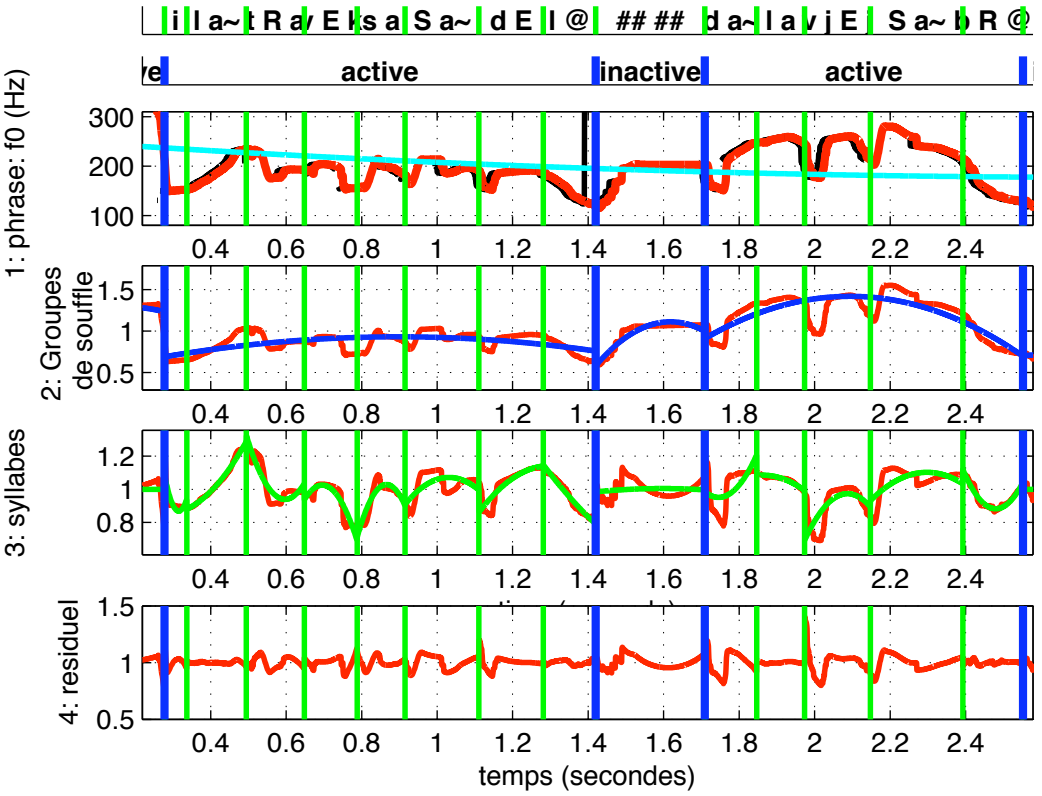

FIG. 3.5: Stylisation hiérarchique. Explications données par l'algorithme 1.

$Af0^{phs}_{model}$

Description	: Modèle quadratique de l'intonation
Type	: Acoustique et continue
Unité	: Phrase
Composant	: Moyenne [Hz], pente [Hz/s] et courbure [ND]
Dimensions	: 3

$Af0^{gps}_{model}$

Description	: Modèle quadratique d'un résiduel de l'intonation
Type	: Acoustique et continue
Unité	: Groupe de souffle
Composant	: Moyenne [ND], pente [ND] et courbure [ND]
Dimensions	: 3

$Af0^{syl}_{model}$

Description	: Modèle quadratique d'un résiduel de l'intonation
Type	: Acoustique et continue
Unité	: Syllabe
Composant	: Moyenne [ND], pente [ND] et courbure [ND]
Dimensions	: 3

3.3.2.2 Résiduel de l'intonation

Le résiduel de l'intonation permet d'observer plusieurs effets réunis, notamment les effets du jitter et de la micro-prosodie. La micro-prosodie correspond aux micro-variations de l'intonation due à l'articulation [Rossi 1981]. Le jitter est la variation de la durée des périodes, de période à période, dans les segments voisés de la parole[Horri 1982, Schoentgen 1995]. C'est un paramètre micro-prosodique important pour la parole expressive car il pourrait être relatif à l'effet "pull" (effet physiologique d'un état interne émotionnel, voir chapitre 2) [Johnstone 1999]. Cet effet se traduit par une micro-variation de la f0. Cette micro-variation est observable dans le domaine de Fourier. En effet, une voix possédant plus de jitter qu'une autre, aura tendance à posséder une fréquence fondamentale dont le contenu spectral est plus riche [Klingholz 1985]. C'est pourquoi l'observation du résiduel de l'intonation proposée ici est relative au centre de gravité du spectre de la f0 : Nous l'avons baptisé le centroïde de l'intonation, car il s'agit du barycentre spectral de l'intonation. De manière à éviter le contenu spectral provoqué par les variations macro-prosodiques citées précédemment, l'estimation du centroïde de l'intonation se fait à partir du résiduel de la stylisation : $Af0_{residuel}$ (courbe rouge de la sous-figure 4 de 3.5). Compte tenu de la faiblesse de la fréquence d'échantillonnage de la f0, une seule valeur du centroïde de l'intonation est estimée par phrase. L'estimation du centroïde de l'intonation repose du l'équation suivante :

$$Af0_{centroide} = centroid(abs(FFT(Af0_{residuel}))) \tag{3.2}$$

dans laquelle, la fonction FFT calcule la transformée de Fourier d'un signal, la fonction *abs* donne le module de cette transformée de Fourier et la fonction *centroid* permet d'estimer le centre de gravité :

$$centroid(x) = \frac{\sum_{f=1}^{F} f \times x}{\sum x} \tag{3.3}$$

$Af0_{centroide}$

Description	: Centroide de l'intonation
Type	: Acoustique et continue
Unité	: Phrase
Composant	: Barycentre spectral du résiduel de l'intonation [Hz]
Dimensions	: 1

3.3.3 Intensité

L'intensité est un paramètre prosodique peu étudié. En effet, celui-ci est souvent considéré comme un corrélat de l'intonation. Il semble en effet qu'à une augmentation de la f0 correspond une augmentation de l'intensité. La mesure de l'intensité, au sens prosodique est assez différente de la mesure de l'énergie instantanée. En effet, l'énergie instantanée est largement influencée par le contenu phonétique. Or, si une voyelle possède une énergie instantanée plus importante en moyenne qu'une plosive, cette-dernière n'en semble pas perçue de manière moins intense. Il faut donc ajouter un paradigme simulant la perception dans l'estimation de l'intensité prosodique. Ceci est réalisé grâce à l'utilisation d'une mesure d'intensité perçue ("loudness" en anglais) [Peeters 2004]. L'algorithme effectue la transformée de Fourier du signal fenêtré. Puis il effectue un filtrage du spectre d'énergie (module de la FFT) par les caractéristiques de l'oreille moyenne. Ensuite, il applique à ce spectre filtré, un second filtrage par bandes de Bark. Enfin, ces bandes permettent de calculer la sonie ou intensité perçue, ainsi que l'acuité et l'étendue timbrale. La figure 3.6 montre une visualisation de l'intensité ainsi mesurée. Le même algorithme de stylisation que pour l'intonation est utilisé pour modéliser l'intensité. Le même traitement du résiduel est appliqué en sortie de stylisation, pour définir le *centroïde de l'intensité* (micro-variation d'intensité souvent liée à la micro-prosodie et à au shimmer).

$Aint^{phs}_{model}$

Description	: Modèle quadratique de l'intensité
Type	: Acoustique et continue
Unité	: Phrase
Composant	: Moyenne [Phone], pente [Phone/s] et courbure [ND]
Dimensions	: 3

$Aint^{gps}_{model}$

Description	: Modèle quadratique d'un résiduel de l'intensité
Type	: Acoustique et continue
Unité	: Groupe de souffle
Composant	: Moyenne [ND], pente [ND] et courbure [ND]
Dimensions	: 3

$Aint^{syl}_{model}$

Description	: Modèle quadratique d'un résiduel de l'intensité
Type	: Acoustique et continue
Unité	: Syllabe
Composant	: Moyenne [ND], pente [ND] et courbure [ND]
Dimensions	: 3

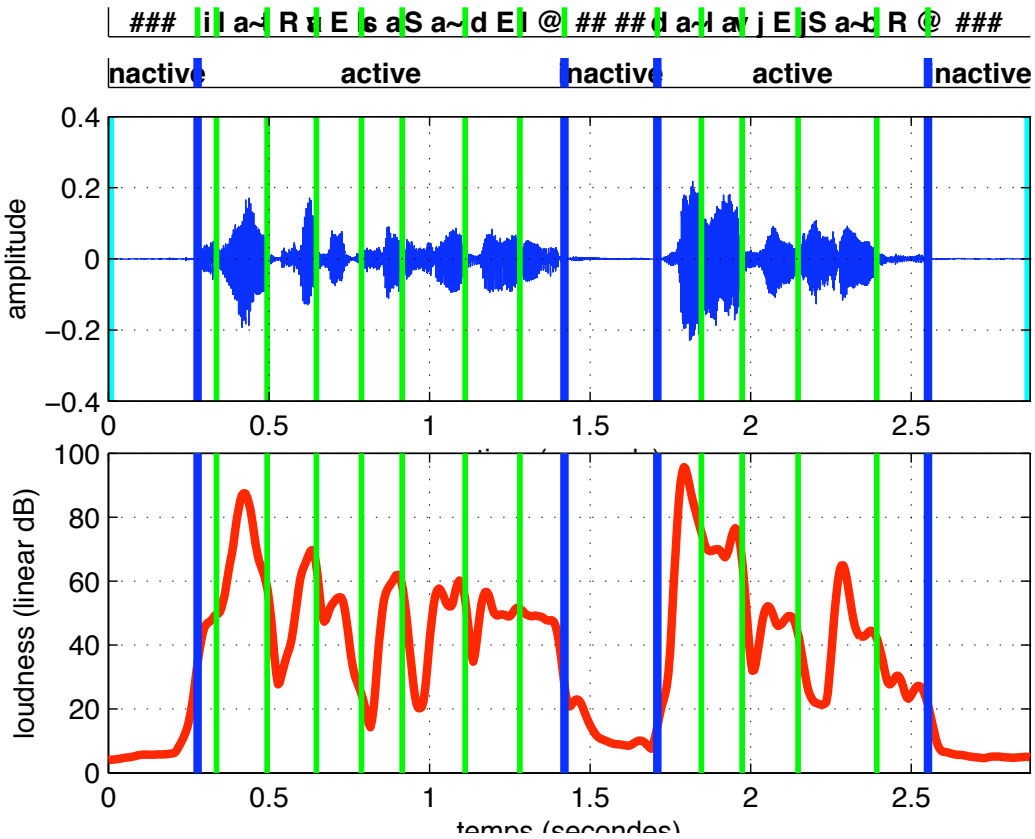

FIG. 3.6: Exemple de visualisation de l'intensité.

$Aint_{shimmer}$

Description	: Shimmer
Type	: Acoustique et continue
Unité	: Phrase
Composant	: Barycentre spectral du résiduel de l'intensité [Phone]
Dimensions	: 1

3.3.4 Débit de parole

Contrairement à l'intonation, le débit de parole est un paramètre prosodique difficile à définir. Il peut être défini "comme un mouvement global de ce qui est dit, prenant en compte les pauses, les arrêts, les accélérations et les décélérations." [Galarneau 2001]. Une langue à accent syllabique (dans la typologie de Pike ou Abercrombie) comme le Français construit son rythme sur des syllabes accentuées qui tendent à posséder une durée plus grande que les syllabes non accentuées, et cela, indépendamment du débit moyen de parole [Fougeron 1998]. Un consensus sur le rôle psycho-rythmique de la syllabe s'est établi et il semble que la perception du débit de parole est plus reliée aux syllabes qu'aux unités segmentales phonétiques

[Zellner 1998]. Ainsi, si une syllabe possède une durée plus longue, c'est parce qu'elle est mise en emphase prosodiquement, indépendamment de son contenu phonétique. Des études sur la production vocale ont renforcé l'hypothèse de l'existence d'un débit syllabique constant [Wheeldon 1994]. Nous utilisons donc les durées des syllabes comme indice acoustique du débit de parole. Ces durées varient selon l'accent et la prosodie, mais aussi en fonction des expressions [Beller 2006b]. A partir de la segmentation syllabique, l'algorithme d'estimation du débit syllabique interpole l'inverse des durées des syllabes. Cela permet la définition d'un débit de parole local qui montre les accélérations et les décélérations (voir figure 3.7).

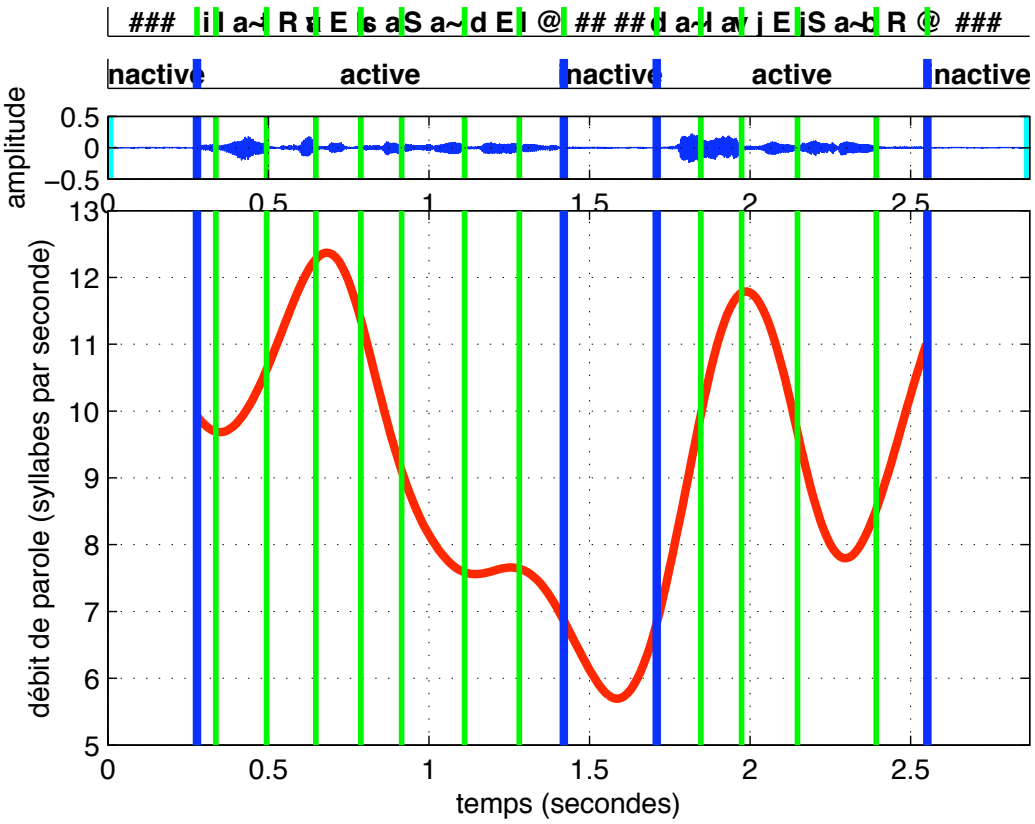

FIG. 3.7: Exemple de visualisation du débit syllabique.

Cette mesure du débit syllabique repose sur les durées des syllabes. Elle ne donne donc aucune information sur la dynamique intra syllabique. Toutefois, une telle mesure semble possible et est relatée dans le chapitre 5. La stylisation utilisée pour le débit de parole repose sur le même algorithme 1 que la f0, mais s'achève au niveau des groupes de souffle puisqu'il n'est pas possible de modéliser le débit à l'intérieur des syllabes à partir de cette mesure.

$Adebit^{phs}_{model}$

Description	: Modèle quadratique du débit de parole
Type	: Acoustique et continue
Unité	: Phrase
Composant	: Moyenne [syl/s], pente [ND] et courbure [ND]
Dimensions	: 3

$Adebit^{gps}_{model}$

Description	: Modèle quadratique d'un résiduel du débit de parole
Type	: Acoustique et continue
Unité	: Groupe de souffle
Composant	: Moyenne [ND], pente [ND] et courbure [ND]
Dimensions	: 3

3.3.5 Degré d'articulation

La théorie "H and H" de Lindblom [Lindblom 1983] propose deux degrés d'articulation de la parole : la parole *Hyper* qui s'oriente vers une clarté maximale du signal produit et la parole *Hypo* qui a comme objectif de produire le signal le plus économique possible. Le degré d'articulation renseigne ainsi sur la motivation/personnalité du locuteur vis à vis de ses interlocuteurs et sur son introversion/extraversion en situation de communication parlée. Cette position peut provenir de plusieurs facteurs contextuels dont l'état émotionnel du locuteur, ou l'expression avec laquelle celui-ci s'exprime.

Certaines représentation des émotions caractérisent les réactions émotionnelles par leurs degrés d'activation (voir partie 2.2.5). Cet indice relate si le locuteur est amené à agir ou à rester passif lorsqu'il est dans un état émotionnel. Ainsi, certaines émotions simulées par les acteurs sont jouées avec des stratégies simulant l'introversion ou l'extraversion. Nous montrons qu'un possible candidat pour la mesure du degré d'activation est le degré d'articulation [Beller 2007b, Beller 2008a].

3.3.5.1 Mesure du degré d'articulation

Le degré d'articulation est influencé par le contexte phonétique, le débit de parole et la dynamique spectrale (qui correspond à la vitesse du changement de configuration du conduit vocal). La mesure du degré d'articulation proposée est différente de la mesure traditionnelle du degré d'articulation [Lindblom 1983, Wouters 2001] qui consiste à définir des cibles formantiques pour chaque phonème, en tenant compte de la coarticulation, et à étudier les différences entre les réalisations et les cibles par rapport au débit de parole. Compte tenu de la difficulté à définir des cibles locales, la mesure présentée ci-dessous implique une mesure statistique du degré d'articulation.

Elle nécessite préalablement trois types d'analyse du signal de parole. Tout d'abord, ce dernier doit être segmenté phonétiquement afin de connaître à quelle

catégorie phonétique appartient chaque portion (trame) de signal. Puis, une segmentation syllabique permet la mesure dynamique du débit local de la parole [Beller 2006b]. Enfin, l'estimation des trajectoires de formant permet la mesure de l'aire du triangle vocalique (voir partie 3.3.5.7). La mesure du degré d'articulation d'une expression provient de l'observation conjointe des évolutions de l'aire du triangle vocalique et du débit de parole en fonction de l'intensité de l'expression (voir figure 3.25). La mesure de l'aire du triangle vocalique nécessite une segmentation phonétique ainsi que l'estimation de la fréquence des formants.

3.3.5.2 Définition d'un formant

Les formants sont relatifs aux modes de résonance du conduit vocal. Dans le modèle source/filtre de la production vocale [Fant 1960], ils correspondent aux résonances du conduit qui filtre la source glottique. Ils influent donc sur la répartition de l'énergie dans le spectre dans le cas des voyelles et plus largement, dans le cas des sons voisés et non voisés que produit la source glottique [Kaehler 2000]. L'examen de leurs trajectoires dans le plan temps/fréquence peut donner de nombreuses informations sur des phénomènes tels que la coarticulation ou le degré d'articulation [Bailly 1995]. Ils peuvent être modélisés par des filtres passe-bande en série ou en parallèle [Klatt 1980, Rodet 1997]. Leurs fréquences, amplitudes, phases et largeurs de bande varient au cours du temps et nous permettent de distinguer les voyelles de même hauteur. Ces paramètres sont fortement corrélés aux configurations du conduit vocal ainsi qu'à sa géométrie et peuvent varier fortement, par exemple, entre une voix d'homme et une voix de femme. On distingue généralement les résonances du conduit vocal en fonction de leurs régions fréquentielles, ce qui mène à une catégorisation des formants en fonction de leurs fréquences.

3.3.5.3 Estimation de la fréquence des formants

Il existe de nombreux outils permettant l'estimation de la fréquence des formants. La majorité de ces outils modélisent l'enveloppe spectrale du signal découpé en N trames temporelles et fenêtré, par un système autorégressif dont les pôles P correspondent aux résonances du conduit vocal. En filtrant ces pôles selon un ordre d'importance, ils sont capables de définir pour chaque trame (n), un ensemble restreint de pôles candidats $P_{can}(n)$ parmi lesquels certains correspondent aux formants. Mais ces outils n'attribuent pas un index de formant à ces pôles. C'est à dire qu'ils donnent un ensemble de candidats possibles mais qu'ils n'affectent pas ces candidats à un formant particulier. Or, l'étape d'attribution des pôles aux formants est nécessaire à l'observation du triangle vocalique puisque celle-ci requiert la connaissance de F_1 et F_2.

3.3.5.4 Attribution des pôles à des formants

En pratique, il se peut qu'un ensemble de pôles candidats $P_{can}(n)$ soit très différent de celui le précédant $P_{can}(n-1)$. Même le nombre de candidats peut changer d'une trame à l'autre. La figure 3.8 présente les ensembles de pôles estimés par Wavesurfer [Sjölander 2000] et PRAAT [Boersma 2001]. Si l'on attribue d'emblée ces ensembles rangés par ordre de fréquence croissante aux formants, on observe des sauts de fréquence importants en ce qui concerne leurs trajectoires. Il suffit que le pôle de plus basse fréquence disparaisse pour que tous les autres se voient affecter à un formant de rang supérieur (exemple sur la figure 3.8, à la 0,98 seconde). Traditionnellement, l'affectation des ensembles $P_{can}(n)$ aux formants est à la charge de l'utilisateur. Celui-ci trie les pôles selon leurs régions fréquentielles. Par exemple, pour une voix d'homme, un pôle dont la fréquence est située entre 800Hz et 1500Hz sera souvent nommé deuxième formant. Mais cet a priori peut biaiser les résultats si plusieurs formants sont présents dans une région fréquentielle ou si la fréquence d'un formant excède ces limites, ce qui est parfois le cas dans la parole expressive. De plus, la quantité de données utilisées nécessitent une automatisation de l'attribution des ensembles de pôles candidats $P_{can}(n)$ aux formants.

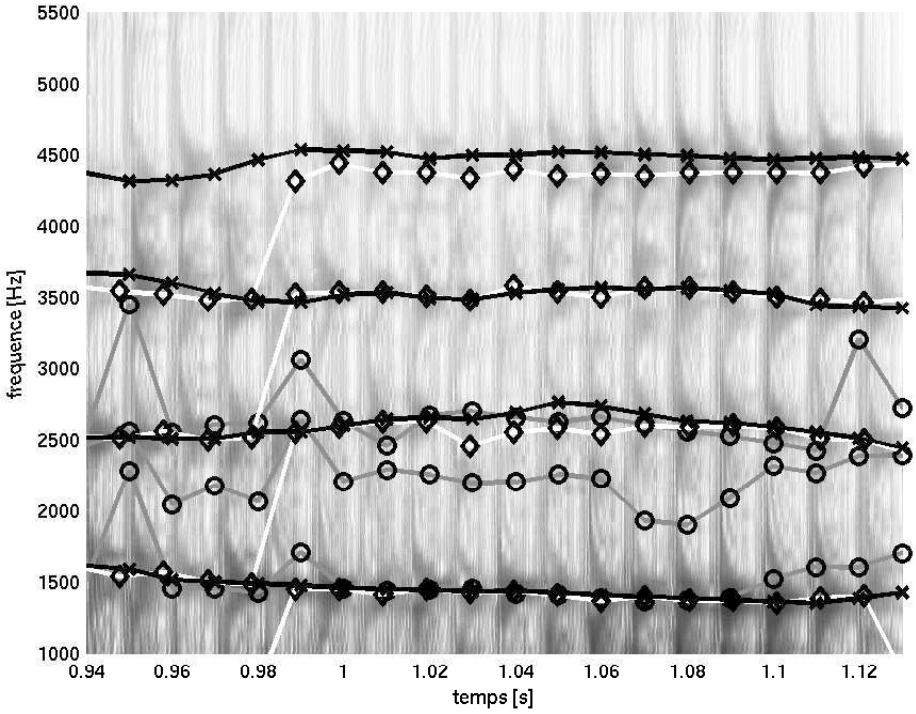

FIG. 3.8: Trajectoires de formants estimées par Praat (gris-cercle), par Wavesurfer (blanc-losange) et par la méthode formant-Viterbi proposée (noir-croix), tracées sur le spectrogramme d'une voyelle [A].

3.3.5.5 Algorithme formant-Viterbi

Un nouvel algorithme, baptisée formant-Viterbi, attribue à chaque formant, un des pôles candidats, sans aucun a priori sur sa région fréquentielle, et prend simultanément en compte une contrainte de continuité sur la trajectoire du formant.

Hypothèses Cet algorithme repose sur trois hypothèses :

- Hyp_1 : Les formants correspondent à des pôles proéminents de la modélisation de l'enveloppe spectrale par un système autorégressif.
- Hyp_3 : Ces pôles peuvent être classés selon des règles de proéminence et selon leurs places respectives les uns par rapport aux autres.
- Hyp_2 : La trajectoire d'un formant possèdent une certaine continuité dans le plan temps-fréquence.

La contrainte de continuité des trajectoires permet en pratique, de diminuer le bruit de l'estimation trame à trame.

Appartenance d'un pôle à un formant La première étape est une quasi-dérivation du signal découpé en N trames et fenêtré, par un filtre tout-zéros de la forme : $F(z) = 1 - az^{-1}$ où le paramètre "a" est proche de l'unité (0.98 étant une valeur empirique commune). Cela permet de supprimer une éventuelle composante continue et d'accentuer les hautes fréquences du signal. Le but de cette pré-accentuation est de tenter d'éliminer les effets de pente spectrales dus à la source (-12 dB/octave) et au rayonnement aux lèvres (+ 6 dB/octave) [Henrich 2001]. Puis une analyse linéaire prédictive (LP) du signal filtré est effectuée[3]. On évalue les racines de ce polynôme, constituant les P_{can} pôles candidats de l'enveloppe spectrale pour chaque trame n. Pour chaque pôle p d'une trame n, on mesure :

- F(p,n) : la fréquence correspondante (angle du pôle)
- Q(p,n) : la largeur de bande (proximité du pôle au cercle unité)
- Gd(p,n) : le retard de groupe du polynôme LPC à la fréquence du pôle
- A(p,n) : l'amplitude du polynôme LPC à la fréquence du pôle.

Les trois dernières grandeurs caractéristiques des pôles sont normalisées par rapport à l'horizon temporel correspondant à la phrase. Un poids (entre 0 et 1) est attribué à chacune de ces grandeurs caractéristiques. La favorisation du retard de groupe par l'attribution d'un poids plus conséquent permet d'obtenir de meilleurs résultats [Murthy 1989a, Murthy 1989b, Murthy 1989c, Murthy 1991a, Murthy 1991b, Murthy 2003, Bozkurt 2005, Zhu 2004, Duncan 1989]. Pour la trame n, la probabilité d'appartenance d'un pôle p à un formant Prob(p,n), découle de la somme pondérée de ses caractéristiques (Hyp_1). La matrice d'observation Prob(p,n) représente une phrase entière dans le plan temps-fréquence. A un instant n donné,

[3] Un travail à réaliser consiste à changer l'estimation de l'enveloppe par la LPC par la méthode TrueEnvelop-LPC [Villavicencio 2006], après neutralisation des effets de la glotte par une séparation source-filtre ARX-LF dans laquelle la source (X) est modélisée par un modèle LF [Fant 1960, Degottex 2008b, Vincent 2005b, Vincent 2005a, Vincent 2006, Vincent 2007].

Prob(p,n) renseigne sur la probabilité que la trajectoire d'un formant passe par une fréquence angle(p) donnée.

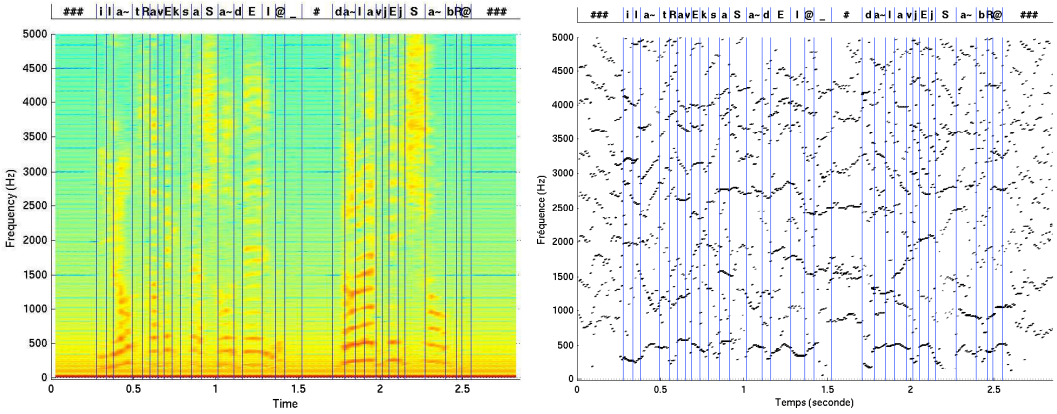

FIG. 3.9: A gauche : Spectrogramme et segmentation phonétique. A droite : Matrice de probabilité d'observation.

Trajectoires de formant La contrainte de continuité de la trajectoire spectro-temporelle d'un formant (Hyp_3) est représentée par une matrice de probabilité de transition Tran(p, p), de Toeplitz (symétrique et circulaire) [Hastie 2001] du type :

$$T(p,p) = \begin{bmatrix} 1 & \frac{2P-1}{2P} & \frac{2P-2}{2P} & \dots & 0 \\ \frac{2P-1}{2P} & 1 & \frac{2P-1}{2P} & \dots & \frac{1}{2P} \\ \frac{2P-2}{2P} & \frac{2P-1}{2P} & 1 & \dots & \frac{2}{2P} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \frac{1}{2P} & \frac{2}{2P} & \dots & 1 \end{bmatrix}$$

Les trajectoires des formants sont "décodées" récursivement, une après l'autre, par un algorithme de Viterbi qui prend en compte les N trames de la phrase. La programmation dynamique permet de tracer une trajectoire de formant sur la matrice Prob(p,n) tout en respectant la continuité Tran(p,p) à chaque trame. La trajectoire du premier formant est estimée en initiant sa fréquence à 0Hz à la première trame (t = 0). Les pôles correspondant à ce premier formant sont ensuite éliminés de la matrice de probabilité d'appartenance des pôles au formant Prob(p,n) (Hyp_2). Puis la trajectoire du second formant est évaluée de la même façon et ainsi de suite (voir figure 3.10). La tentative d'estimer la densité de probabilité conjointe de tous les formants en même temps a échoué à cause de la complexité à définir la matrice de transition T(p,p) de tous les formants en même temps. Les chemins empruntés par les trajectoires de formant permettent par cumulation des probabilités, d'obtenir un indice de confiance par formant, mais aussi un indice de confiance au total de l'estimation (voir figure 3.10).

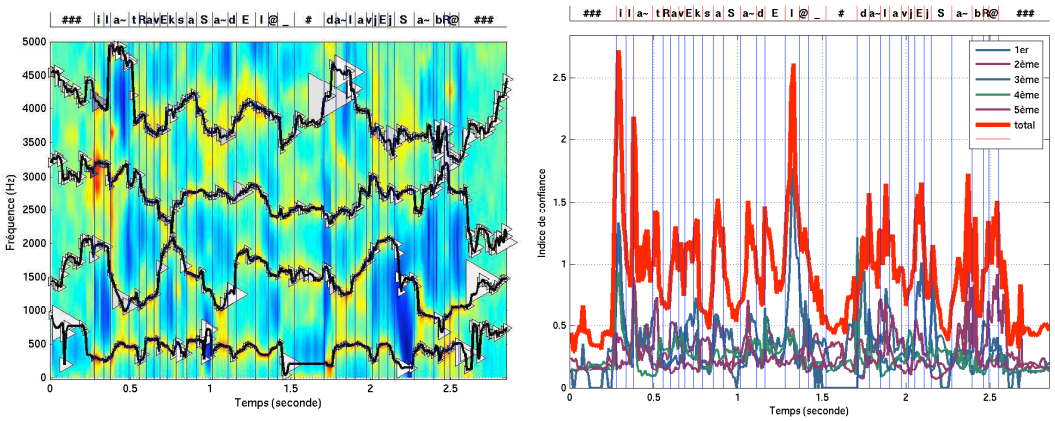

FIG. 3.10: A gauche : LPCgramme, trajectoires de formant et pôles choisis représentés par des triangles dont la couleur est relative à l'amplitude et dont la largeur est représentative de la largeur de bande. A droite : indices de confiance de l'estimation.

3.3.5.6 Fréquence des formants

Une de ces trajectoires estimées permet de connaître à chaque trame temporelle, la fréquence du formant correspondant qui évolue pendant la durée d'un phone. Afin de minimiser les erreurs d'estimation et d'obtenir une seule valeur représentative par phone appelée *fréquence caractéristique*, une mesure globale est effectuée sur toutes les trames temporelles de chaque phone, en utilisant la segmentation phonétique. Un polynôme de Legendre d'ordre 2 modélise l'évolution temporelle de la trajectoire d'un formant sur le phone (voir figure 3.11). Si l'évolution de la fréquence est linéaire, la fréquence caractéristique correspond à la valeur médiane des fréquences prises sur quelques valeurs avoisinant le milieu du phone. Si l'évolution de la fréquence est parabolique, ce qui montre que la fréquence d'un formant a atteint une cible puis s'en est écarté (coarticulation), la fréquence caractéristique correspond à la valeur médiane des fréquences prises sur quelques valeurs avoisinant l'instant où la fréquence a atteint sa cible (instant ou la dérivée du polynôme d'interpolation du 2^{nd} ordre s'annule). Cette mesure reflète mieux la cible "visée" par le locuteur lors de la prononciation de la voyelle et possède une variance inférieure à celle de la moyenne calculée sur tout l'horizon temporel du phone.

$Aformant1^{phone}_{model}$

Description	: Modèle quadratique de la fréquence du premier formant
Type	: Acoustique et continue
Unité	: Phone
Composant	: Moyenne [Hz], pente [Hz/s] et courbure [ND]
Dimensions	: 3

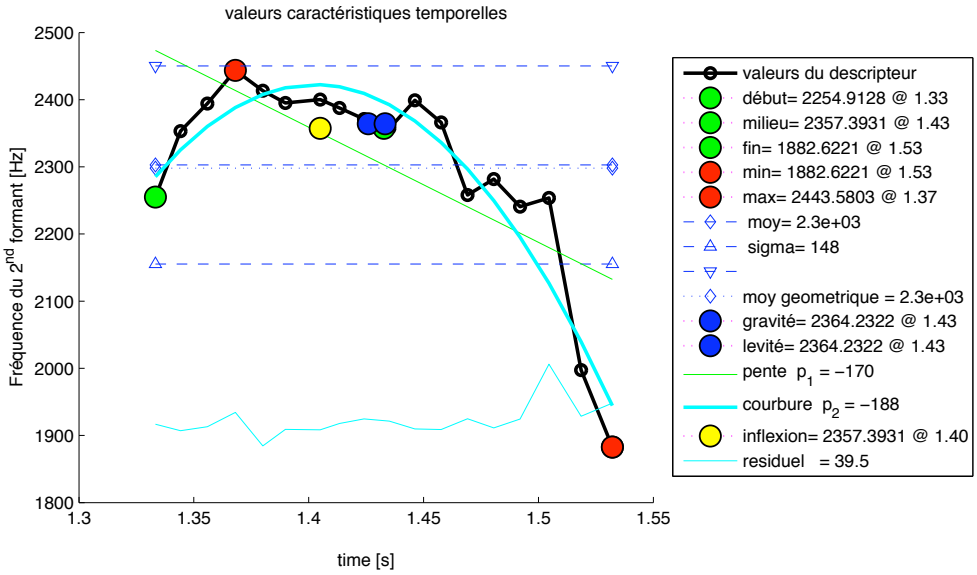

FIG. 3.11: Valeurs caractéristiques temporelles calculées sur la fréquence du second formant. Le point jaune pointe la valeur caractéristique de la fréquence retenue pour le phone entier.

$Aformant2^{phone}_{model}$

Description	: Modèle quadratique de la fréquence du deuxième formant
Type	: Acoustique et continue
Unité	: Phone
Composant	: Moyenne [Hz], pente [Hz/s] et courbure [ND]
Dimensions	: 3

$Aformant3^{phone}_{model}$

Description	: Modèle quadratique de la fréquence du troisième formant
Type	: Acoustique et continue
Unité	: Phone
Composant	: Moyenne [Hz], pente [Hz/s] et courbure [ND]
Dimensions	: 3

$Aformant4^{phone}_{model}$

Description	: Modèle quadratique de la fréquence du quatrième formant
Type	: Acoustique et continue
Unité	: Phone
Composant	: Moyenne [Hz], pente [Hz/s] et courbure [ND]
Dimensions	: 3

3.3.5.7 Triangle vocalique

Delattre [Delattre 1955] a montré qu'il était possible de catégoriser les voyelles dans un espace appelé cardinal, possédant deux dimensions liées aux fréquences des deux premiers formants, en français. Les différentes voyelles du français y forment un triangle appelé triangle vocalique. Les sommets de ce triangle sont définis par les fréquences des 2^{nd} (F_2) et du 1^{er} (F_1) formants des voyelles /a/, /i/ et /u/ (voir figure 3.12). Les fréquences moyennes ou les fréquences caractéristiques de ces formants sont utilisées pour placer chaque phone dans l'espace cardinal. La figure 3.12 permet de comparer ces deux estimateurs (fréquence moyenne et fréquence caractéristique) pour le même acteur pour l'expression neutre. Dans la figure de gauche, les phones sont placés à partir des moyennes de leurs fréquences, alors que dans la figure de droite, ils sont placés à partir de leurs fréquences caractéristiques. La juxtaposition des deux courbes montrent comment l'utilisation de la fréquence caractéristique (qui reflète mieux la cible visée) permet de mieux dissocier les phones centraux. De plus, en ce qui concerne les /u/, l'utilisation de la fréquence caractéristique permet de regrouper des familles qui se trouvent dans des contextes phonétiques différents.

3.3.5.8 Phénomène de réduction vocalique

Une étude sur l'influence du débit sur le triangle vocalique [Gendrot 2004], montre que les formants tendent vers une voyelle centrale pour les segments de courte durée. Sur la figure 3.12, la durée des phones est représentée par la taille des lettres et par leur luminosité (plus le phone est long, plus la lettre est grosse et foncée). On observe bien ce phénomène de réduction sur la famille des /i/. Ceci suggère que la réduction n'est pas un phénomène exclusivement linguistique, mais admet aussi une cause d'ordre physique ou physiologique. Or les émotions sont liées à des modifications sur les plans physique et physiologique. C'est pourquoi elles aussi, peuvent influencer ce phénomène de réduction/expansion du triangle vocalique. La mesure de l'aire du triangle joignant les voyelles limitrophes, associée à la mesure du débit de parole, permettent la comparaison entre différents degrés d'articulation correspondant aux expressions.

3.3.5.9 Mesure du degré d'articulation

Si l'observation des variations au sein d'un phone ne fait pas partie des modèles prosodiques, c'est parce que les attributs prosodiques nécessitent, par définition, une observation sur une fenêtre temporelle dépassant le phone. Toutefois, nous avons observé combien la prononciation des phonèmes peut être influencée par l'expressivité. De plus, certaines expressions semblent se démarquer des autres par la prononciation particulière de certains phonèmes. Fonagy explique ce phénomène par la psycho-linguistique [Fónagy 1983]. Quoiqu'en soit l'explication, il reste que certains phones possèdent un mode de prononciation spécifique pour chaque expression (notamment les consonnes) et que l'observation de ces variations phonétiques

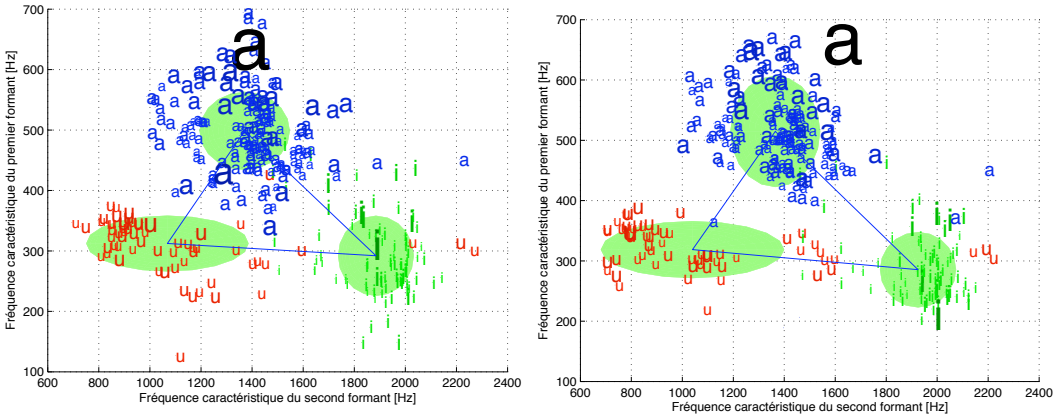

FIG. 3.12: Triangle vocalique dans l'espace cardinal de l'acteur *Roullier* pour l'expression neutre. La taille et la luminosité des lettres représentent la durée du phone (plus le phone est long, plus il est gros et foncé). Les centres des ellipses représentent les moyennes des trois classes et leurs largeurs, les variances selon les deux axes. A gauche : les fréquences d'un phone correspondent aux moyennes respectives sur la durée du phone. A droite : les fréquences d'un phone correspondent aux fréquences caractéristiques respectives (voir partie 3.3.5.6).

semble capitale pour caractériser les expressions et leurs différents modes d'articulations. Une comparaison manuelle entre les chaînes phonétiques réalisées par les acteurs et les chaînes phonémiques attendues (dérivant du texte), peut permettre une meilleure mise en évidence que l'approche statistique pour l'observation des tendances relatives aux expressions[4]. Compte tenu de la difficulté de définir des cibles localement, nous avons préféré une mesure statistique du degré d'articulation. Cette mesure s'appuie sur l'observation des voyelles et ne tient donc pas compte de l'articulation consonantique. L'influence de cette dernière a été réduite grâce à l'utilisation du même texte avec toutes les expressions.

La mesure du degré d'articulation proposée nécessite au moins une phrase possédant les trois phonèmes /a/, /i/ et /u/. Elle n'est donc pas dynamique au sens où elle n'évolue pas durant la phrase. De plus, de nombreux individus sont nécessaires pour obtenir une mesure statistique significative. C'est pourquoi le degré d'articulation sera estimé par expression et, non pas, par phrase. La mesure du degré d'articulation repose sur la confrontation de l'aire du triangle vocalique avec le débit de parole.

Enfin, si les expressions influencent le caractère dynamique de l'articulation (parole hyper ou hypo articulée), certaines d'entre elles (correspondant aux émotions utilitaires) s'accompagnent d'expressions faciales [Ekman 1999b] plus ou moins figées durant une phrase. A titre d'exemples, la parole souriante ou joyeuse émane d'un sourire, l'expression du dégoût est accompagnée d'une nasalisation provenant

[4]Un modèle de substitution phonémique dépendant de l'expression pourrait alors contribuer à la synthèse de parole expressive, si celui-ci est placé en amont de la sélection d'unités.

de réflexes ancestraux (qui permettent de bloquer l'odorat), la surprise provoque une arrondissement des lèvres. Ces configurations statiques du conduit vocal, spécifiques à certaines expressions, se retrouvent dans le signal de parole à travers la mesure du degré d'articulation.

FIG. 3.13: Expressions faciales du dégoût, de la joie et de la surprise tirées du projet d'Analyse Automatique de Visages de CMU : http://www.cs.cmu.edu/~face/

3.3.6 Phonation

La phonation semble constituer un indice remarquable de l'expressivité [Gobl 2003, Gendrot 2002, Campbell 2003]. Ce n'est que récemment qu'elle a été introduite en tant que dimension prosodique. Cette prise en compte "tardive" provient d'un manque de consensus sur sa définition, ainsi que de la difficulté à la mesurer. En effet, l'étude de la phonation comprend plusieurs phénomènes provenant tous de la même origine, la glotte, mais pouvant posséder des effets très différents. Parmi ces phénomènes, la *qualité vocale* permet la distinction entre des voix pressées, normales ou relâchées [D'Alessandro 2003b, D'Alessandro 2003a]. Elle est mesurable grâce à l'estimation du coefficient de relaxation Rd [Fant 1997, Degottex 2008b]. Le *voisement* permet de décrire l'efficacité de la phonation ou le degré de souffle glottique. C'est une sorte de rapport signal/bruit au niveau de la source glottique ; Enfin le *mode vibratoire* permet de distinguer des modes de production relatifs aux différents registres du grave à l'aigu : voix fry (mode 0), voix de poitrine (mode 1), voix de tête (mode 2) et voix de falsetto (mode 3). Si ces différents phénomènes sont connus, des confusions existent quant à leurs qualifications. En effet, le mode vibratoire fry est souvent associé à la voix "creaky" qui dénomme une voix serrée en mode fry. La figure 3.14 représente divers phénomènes liés à la phonation dans un espace physiologique dont les axes sont l'effort pulmonaire et la constrictions des plis vocaux. Ces deux axes traduisent des efforts qui peuvent être contrôlés ou non. Comme une évidence, ces efforts peuvent être ralliés aux expressions. Il est, en effet, désormais connu que l'expression de certaines émotions (utilitaires) s'accompagne de modifications physiologiques (battement cardiaque, sécheresse dans la bouche...). Ainsi, la phonation est une facteur important pour la caractérisation des expressions.

De manière à étudier la qualité vocale, une analyse de signaux EGG a été

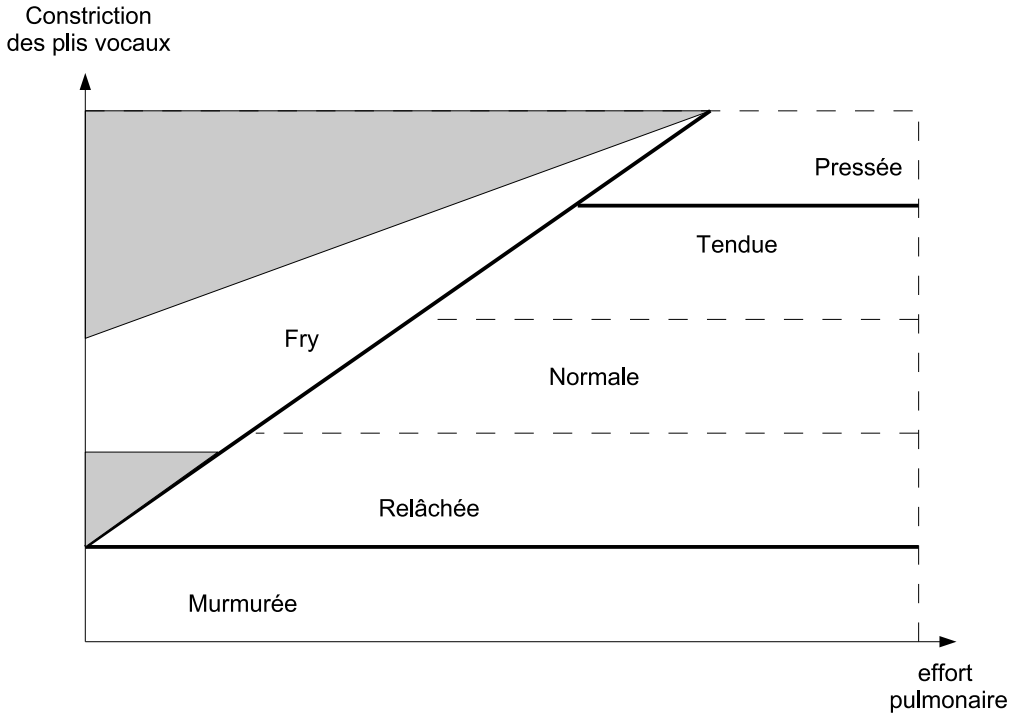

FIG. 3.14: Représentation de divers phénomènes phonatoires dans un espace physiologique dont les axes sont l'effort pulmonaire et la constriction de plis vocaux.

menée sur la base de données de parole expressive allemande emoDB[5]. A cette occasion, de nombreux traitements des signaux EGG ont été mis au point et sont reportés dans l'annexe C. Ces traitements permettent la mesure du quotient ouvert et de la fréquence fondamentale instantanée. En conclusion, les résultats n'étant pas exploitables pour la transformation sans enregistrements EGG, il a été préféré une solution ne s'appuyant que sur le signal audio. Un algorithme de séparation source/filtre [Degottex 2008a] s'inspirant d'une méthode d'inversion AR-X [Vincent 2005b, Vincent 2005a, Vincent 2006, Vincent 2007] dans laquelle la source "X" est modélisée par un modèle LF [Fant 1985] a été utilisé. Cet algorithme définit les instants de fermeture et d'ouverture de la glotte, par conséquent, la fréquence fondamentale instantanée, et enfin, les paramètres du modèle LF (Quotient d'ouverture, coefficient d'asymétrie, durée de la phase de retour et énergie) qui permet de modéliser au mieux la source. Les paramètres dynamiques du modèle LF estimé permettent d'obtenir le facteur de forme Rd ou coefficient de relâchement/relaxation [Lu 1999, Lu 2002]. Ce coefficient permet de caractériser la qualité vocale : une voix tendue possède un Rd tendant vers $\rightarrow 0.3$, une voix normale se situe dans le milieu Rd $\simeq 1.2$ et une voix relâchée possède un Rd tendant vers $\rightarrow$ 2,5. La figure 3.15 représente ces grandeurs estimées sur notre phrase exemple : la valeur moyenne du coefficient Rd sur le phone et la composante

[5]EmoDB : http://database.syntheticspeech.de/

de notre modèle prosodique qui permet de modéliser la qualité vocale.

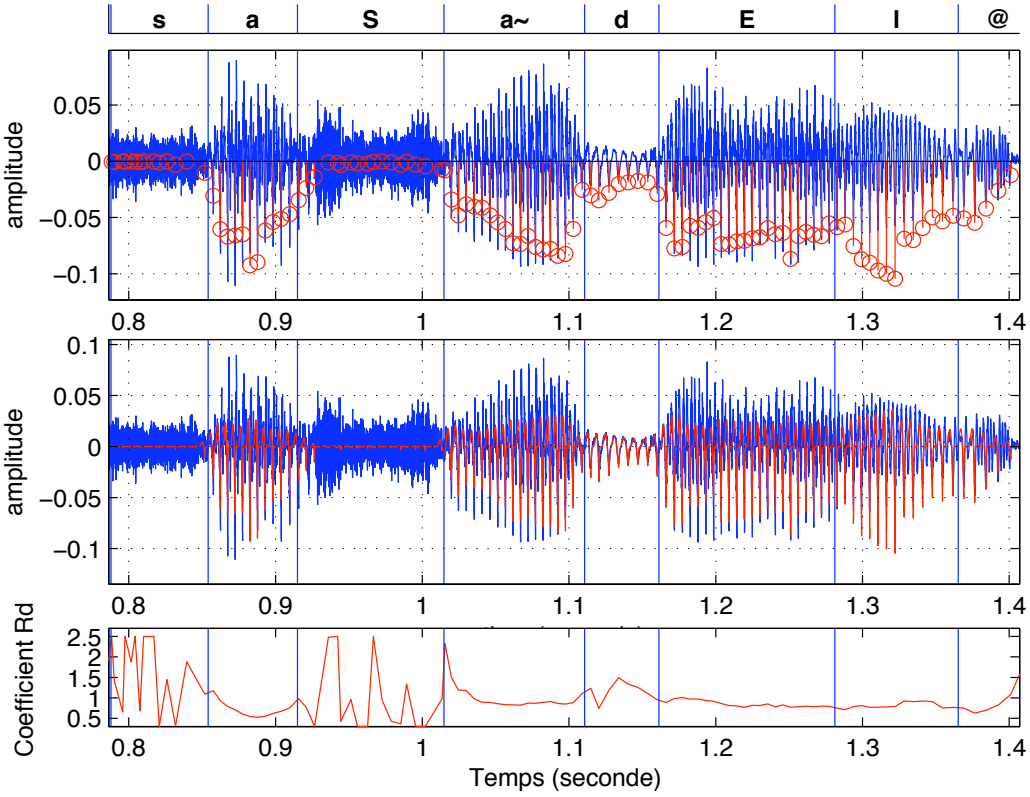

FIG. 3.15: En haut : Signal acoustique et marquage des instants de fermeture. Au milieu : Signal acoustique et model LF. En bas : Coefficient Rd.

ARd^{phone}_{model}

Description	: Modèle quadratique du coefficient de relaxation
Type	: Acoustique et continue
Unité	: Phone
Composant	: Moyenne [Hz], pente [Hz/s] et courbure [ND]
Dimensions	: 3

Les données collectées durant l'enregistrement des corpus consistent en un fichier audio pour chaque phrase et un fichier XML correspondant, contenant les annotations de l'expressivité (catégorie et intensité), du texte déclamé, et des informations relatives à l'identité du locuteur (âge, sexe, nom). A ces données de départ, sont ajoutées des annotations manuelles (segmentation phonétique, annotation paralinguistique et catégorisation de la proéminence), puis des analyses symboliques dérivées de ces annotations et, enfin, des analyses acoustiques. Toutes les données que nous allons décrire par la suite, sont stockées et mises en relation grâce à la plateforme IrcamCorpusTools (voir chapitre A).

3.4 Analyses symboliques

Les données recueillies durant l'enregistrement font partie des analyses symboliques du corpus. En effet, l'analyse symbolique consiste en l'attribution de catégories discrètes ou symboliques ou linguistiques à des portions de signal. Ainsi, l'étiquetage de l'expressivité ou du sexe du locuteur contribue à renseigner le signal acoustique enregistré, et par propagation, toutes les analyses acoustiques qui en découlent. Si ces catégories sont attribuées à la phrase entière, certaines d'entre elles possèdent des horizons temporels plus courts comme c'est le cas pour les catégories phonétiques par exemple.

3.4.1 Segmentation phonétique

La segmentation phonétique du corpus IrcamCorpusExpressivity est, en réalité, une segmentation semi-phonétique. Elle est donc plus précise car une frontière additionnelle est placée dans chaque phone. En effet, un phone est composé de deux semi-phones dont les frontières permettent aussi de constituer des diphones (pour la synthèse, par exemple). La méthode de segmentation automatique employée [Morris 2006, Lanchantin 2008] est classique et repose sur des chaînes de Markov cachées entraînées sur une base de données de parole neutre multi-locuteur [Lamel 1991]. Cette segmentation initiale a ensuite été corrigée manuellement avec l'outil wavesurfer [Sjölander 2000]. Non seulement les frontières ont été déplacées mais les labels ont aussi été changés quand cela s'est avéré nécessaire. Le code utilisé est le XSampa, qui est une version ASCII de la charte IPA[6]. La réalisation phonétique, dans le cas expressif, peut être assez lointaine de la chaîne phonémique automatiquement prédite à partir du texte[7].

[6] IPA : International Phonetic Alphabet

[7] chaîne phonémique et chaîne phonétique : Le phonème est l'étiquette abstraite que l'on donne à un phone, un son, pour le catégoriser. La chaîne phonémique peut être déduite du texte et regroupe la séquence de phonèmes qui est théoriquement attendue. Une fois ce texte prononcé, on peut segmenter phonétiquement la phrase produite. Cette segmentation manuelle ou automatique donne lieu à une chaîne phonétique décrivant la séquence de phones perçue. Parfois, la chaîne phonémique diffère de la chaîne phonétique pour différentes raisons : Restructurations, coarticulation, élision, expression...

Corrigeant la segmentation “bootstrap”, les phonéticiens ont effectivement constaté de nombreuses différences avec les prédictions de la machine (inférées après un entraînement sur des corpus neutres). Pour certaines expressions, des phones attendus ont été réalisés tellement différemment qu'il ont été annotés par d'autres phonèmes. De plus, certains ont disparu tandis que d'autres, inattendus, sont apparus. Certaines corrections ont aussi amélioré la précision de l'étiquetage (une voyelle ouverte changée en voyelle fermée, par exemple). C'est pourquoi, bien que toutes les expressions aient été exprimées avec le même texte, nous nous attendions à des disparités dans les distributions des phones reconnus, imputables à l'expressivité. Toutefois, tous acteurs confondus, les proportions moyennes d'apparition de phones regroupés par classe phonologique (voir figure 3.16), ne montrent pas de différences statistiquement significatives, fonction de l'expression. Seules la discrétion et la tristesse extravertie semblent montrer des différences significatives, mais cela est du aux nombreuses répétitions de consonnes (qui entraînent une proportion de voyelles moindre) pour la discrétion et la particulière inintelligibilité de la tristesse extravertie, comme le montre l'analyse des étiquettes paralinguistiques.

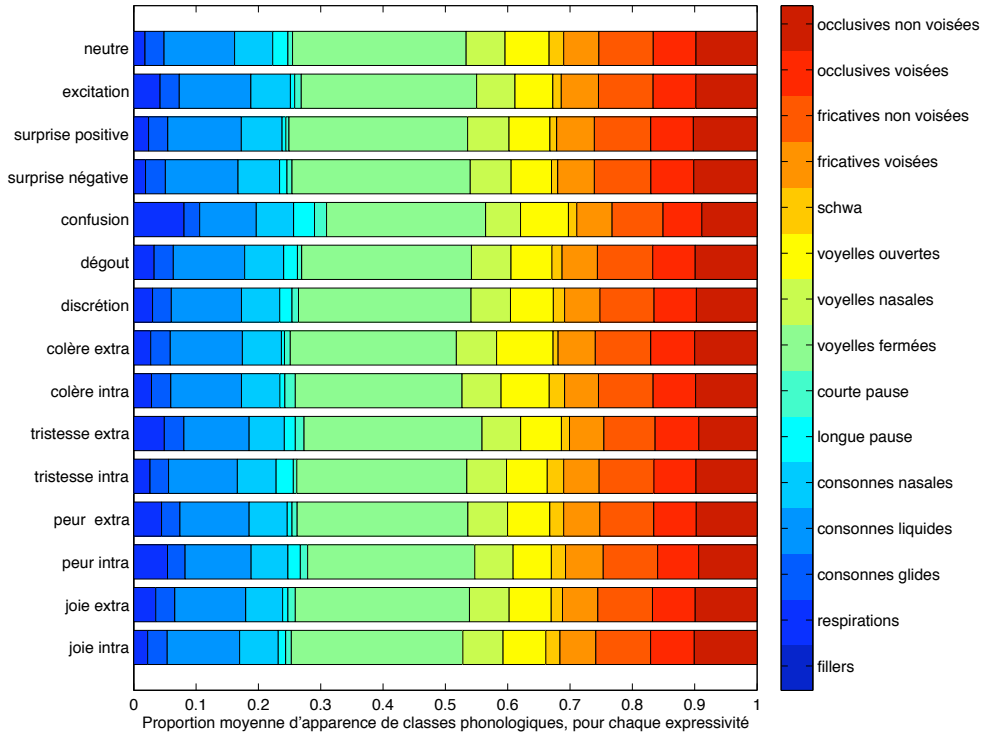

FIG. 3.16: Tous acteurs confondus. Proportions moyennes d'apparition des classes phonologiques, pour chaque expression.

3.4.2 Annotation paralinguistique

Conjointement à l'opération de correction manuelle de la segmentation phonétique, une couche d'annotation paralinguistique a été utilisée de manière à fournir des informations supplémentaires [Douglas-cowie 2000, Douglas-cowie 2007]. Ces informations décrivent notamment les sons non verbaux employés et les éventuelles restructurations, ainsi que divers phénomènes phonatoires ou prosodiques particuliers. Cette étape d'annotation a nécessité, une fois l'étape de segmentation phonétique terminée, une deuxième passe afin d'homogénéiser l'annotation de tous les corpus. En effet, puisqu'aucun dictionnaire d'annotation pour ce type d'éléments paralinguistiques n'a été défini a priori, le vocabulaire employé par les annotateurs a évolué d'un corpus à l'autre. Le dictionnaire que nous soumettons ici, a donc été l'objet de plusieurs discussions inter-annotateurs (et intra) et semble réunir les étiquettes les plus importantes :

- Sons non verbaux, respirations et voisement de la phonation :
  - $^{\circ}$: inspiration
  - $^{\circ\circ}$: expiration
  - nz : respiration nasale (reniflement)
  - bx : bruit indéfini risquant de gêner l'analyse du signal.
  - bb : bruits de bouche
  - ch : chuchotement, instabilité du voisement dans la phonation, les semiphones ne sont que partiellement dévoisés.
  - nv : non voisé : absence totale de voisement dans le semiphone
  - ph : transition : label indiquant une zone paraverbale en continuité avec la phonation verbale (zone souvent courte, mais déterminante pour l'expressivité). Dans la plupart des cas, on rencontre le phénomène [ph] soit juste avant le premier semiphone d'un groupe de souffle, soit juste après le dernier semiphone d'un groupe de souffle.
- Effets de pitch et gutturaux :
  - fp : variation de pitch : grand écart de pitch souvent vers l'aigu, concernant le plus souvent un seul des deux semiphones d'un phonème, changement de mode de vibration
  - fg : bruit guttural : coups de glotte, occlusive glottique, bruit guttural
  - fi : autres bruits provenant de l'appareil vocal
  - nt : non transcriptible : étiquette pour une phonation à vocation verbale mais non transcriptible en phonèmes du dictionnaire "labels de phonèmes".
- Restructurations :
  - lg : long : phonème spécialement long.
  - cu : césure : étiquette appliquée à un phonème silencieux par endroit(s) (phonation saccadée par l'émotion.)
  - rp : répétition : étiquette pour l'indexation numérotée d'un phonème ou d'un groupe de phonème répété.

Enfin, ces différents labels sont composables grâce à l'usage du symbole [/] qui permet de créer une description à partir des étiquettes de base données. Par exemple,

une inspiration voisée est annotée par [°/fi], tandis qu'une expiration nasale sera représentée par [°°/nz]. Enfin l'interaction avec la couche contenant la segmentation phonétique est forte, puisqu'une même étiquette mise en regard avec un silence ou une voyelle ne signifiera pas la même chose.

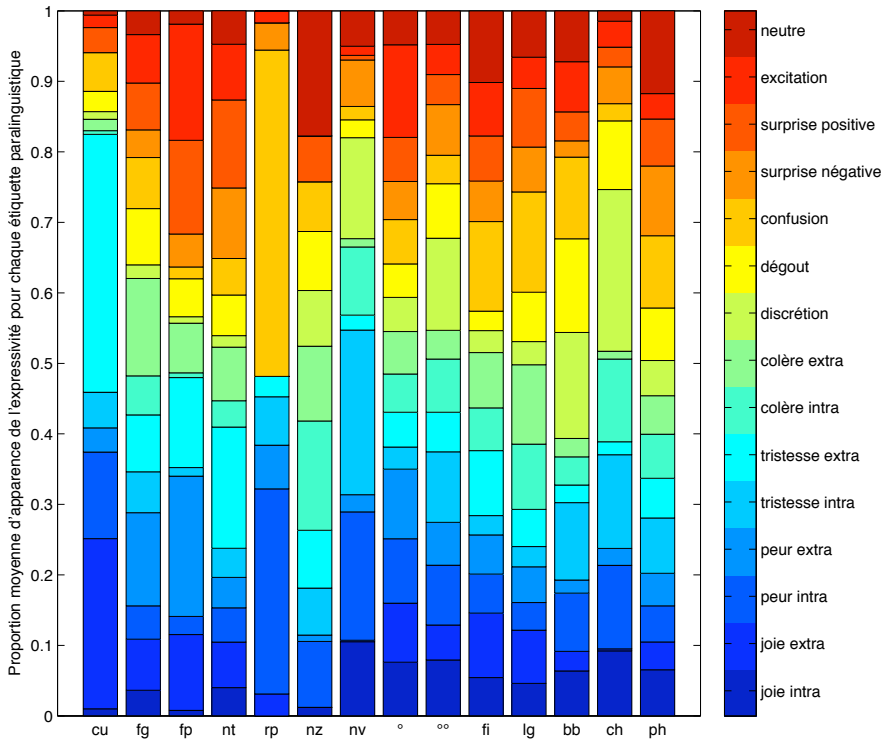

FIG. 3.17: Tous acteurs confondus. Proportion moyenne d'apparition de l'expression pour chaque étiquette paralinguistique.

La figure 3.17 présente pour chacune des étiquettes paralinguistiques, la proportion de chaque expression. Plus une expressivité contient une étiquette de manière récurrente (par rapport aux autres), plus la hauteur de son rectangle associé est grande. Ainsi, on observe que la césure a été fortement employé dans l'annotation de phrase exprimée avec la tristesse extravertie. Cette expression contient aussi beaucoup d'effet de pitch et de phonèmes non transcriptibles (certaines phrases sont presque inintelligibles à cause d'un trop faible degré d'articulation [Beller 2008a]). La colère extravertie semble se distinguer des autres expressions par l'emploi d'effets gutturaux et d'expirations nasales (comme la colère introvertie et le dégoût). Les répétitions n'apparaissent que pour la peur introvertie, la tristesse extravertie et la confusion (qui se démarque aussi par de nombreux phonèmes anormalements longs, comme les colères et les joies [Beller 2006b]. La discrétion et la tristesse introvertie contiennent de nombreux marqueurs de dévoisement ("nv", "ch" et "ph"). De plus,

les inspirations sont très courtes vis à vis des expirations dans le cas de la discrétion. Enfin la surprise négative semble "moins voisée" que la surprise positive et présente moins d'inspiration que d'expiration par rapport à celle ci. De nombreuses autres interprétations sont possibles et peuvent, là aussi être appuyée par des examens plus détaillés, au cas par cas.

3.4.3 Annotation de la proéminence

A partir de la segmentation phonétique, une étape de syllabification par règles produit une segmentation en syllables. La syllable joue un rôle particulier dans la prosodie, notamment parce qu'elle est le plus petit groupe prosodique prononçable. Dans le flux de la parole, certaines syllabes sont démarquées des autres et deviennent alors proéminentes. La proéminence joue un rôle fondamental dans la communication verbale, car sa réalisation requiert une part plus importante de contrôle, lors de sa production. Cette même proéminence sera interprétée par l'interlocuteur comme un marqueur local de l'importance du message. C'est pourquoi, la proéminence est très importante pour l'expressivité[8]. Un seul annotateur a étiqueté le degré de proéminence de toutes les syllables de tous les corpus. L'échelle sémantique mise au point par Anne Lacheret, Nicolas Obin, Jean-Philipe Goldman et moi-même, au préalable [Obin 2008], est composée de quatre niveaux perceptiblement distinguables :

- UN : Undéfini ou encore silence/pause (considéré comme une syllable)
- NA : Non proéminente
- AS : Proéminence secondaire
- AI : Proéminence avec emphase, insistance ou focus
- AF : Proéminence finale

Une étude similaire aux précédentes concernant la répartition de ces étiquettes en fonction de l'expression ne montre pas de variations significatives. Seule la colère extravertie semble se distinguer des autres expressions par une plus grande proportion d'étiquettes "AI". Ainsi, les syllables exprimées avec colère extravertie semblent perçues plus souvent comme proéminentes. Cela peut s'expliquer, en partie, par de l'hyperarticulation qui provoque un détachement des syllables, qui se retrouvent ainsi toutes proéminentes car démarquées.

Plusieurs étapes d'annotation et d'analyse ont permis l'examen du corpus IrcamCorpusExpressivity, sous l'angle des différents phénomènes appartenant au canal paralinguistique. Peu de différences attribuables à l'expressivité sont visibles dans les distributions phonologiques, ainsi que dans les distributions des niveaux de proéminence. Toutefois, ces résultats sont à minimiser puisque les acteurs avaient justement pour consigne explicite, de ne pas faire varier ces constituants, mais seule-

[8] D'une certaine manière, la théorie "push-pull" peut être, ici, déclinée dans une version locale et dynamique. Les syllabes non proéminentes manifestent plutôt l'effet "push", tandis que les syllabes proéminentes, plus maîtrisées, voient leurs réalisations marquées par l'effet "pull". Si l'on considère qu'un état interne peut se manifester de manière contrôlée ou pas, alors la distinction entre syllables proéminentes et non proéminentes peut aider à effectuer cette séparation.

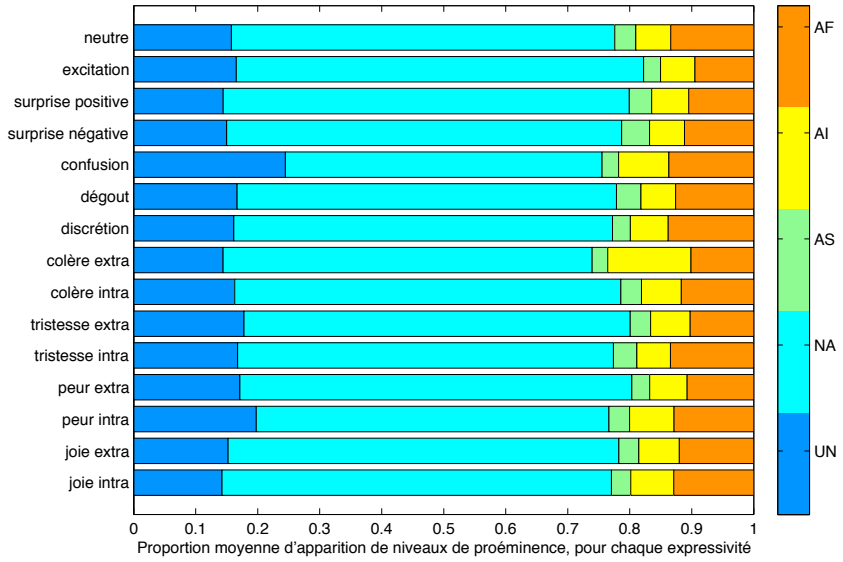

FIG. 3.18: Tous acteurs confondus. Proportion moyenne d'apparition des niveaux de proéminence, pour chaque expression.

ment la prosodie. En revanche, s'ils avaient aussi pour consigne d'éviter l'usage de sons nonverbaux et de restructurations, il n'en reste pas moins que ces constituants apparaissent fréquemment dans le corpus. Leur examen a permis de mettre en évidence que certaines expressions se distinguent par l'usage de ces constituants. Comme si certaines d'entre elles, nécessitaient l'emploi de sons nonverbaux et de restructurations, en plus des variations prosodiques, pour être exprimées.

3.5 Analyses prosodiques

Les résultats des analyses prosodiques présentées dans cette partie proviennent de l'observation statistique des paramètres du modèle prosodique présenté dans la partie 3.3, appliqué aux corpus. Ce modèle prend en compte les analyses symboliques mentionnées précédemment puisqu'il est hiérarchique et qu'il repose sur plusieurs niveaux de segmentation (syllabe, groupe de souffle et phrase). Il s'agit donc d'analyses prosodiques contextuelles dans le sens ou elles prennent en compte des données symboliques. Une exposition exhaustive de tous les résultats demanderait un nombre de pages conséquent tant la variété des contextes est grande. C'est pourquoi nous n'exposons dans cette partie, que les résultats les plus importants et qui justifient l'emploi d'un modèle contextuel.

3.5.1 Intonation

Les résultats concernant l'intonation sont en accord avec ceux rencontrés dans la littérature. Une contribution originale de cette étude consiste en la confrontation de la moyenne de la fréquence fondamentale par expression avec le degré d'activation. Le degré d'activation distingue les émotions associées à l'action ou à la passivité. En pratique, les émotions de degré d'activation négatif (passivité) et positif (activité) se manifestent respectivement par l'introversion et l'extraversion du locuteur. La figure 3.19 montre comment l'acteur Combe augmente la hauteur de sa voix en fonction du degré d'activation de l'expression. Même si ce degré d'activation n'a pas été mesuré par des tests perceptifs, la tendance est préservée chez les autres acteurs. Ce qui nous amène à penser que le rapport introversion/extraversion est peut-être mesurable par la moyenne de la fréquence fondamentale. D'un point de vue physiologique, augmenter la hauteur de sa voix correspond à un effort (une tension accrue des cordes vocales ou de la pression sous-glottique). D'une manière locale, la proéminence est souvent le fruit de cet effort et une syllabe proéminente se distingue souvent des autres non proéminentes par une hauteur supérieure. Il semble donc, avec toute la réserve sur la mesure du degré d'activation, qu'un locuteur allant vers l'extraversion, c'est à dire vers les autres, fournirait un effort supplémentaire qui se traduirait par une croissance de la hauteur de sa voix. Bien sur, tout ceci n'est que supposition et demande une étude approfondie.

3.5.1.1 Centroïde de l'intonation

La mesure du centroïde de l'intonation appliquée aux corpus montre des résultats mitigés. D'un côté, ils corroborent certaines tendances trouvées dans la littérature [Johnstone 1999], comme l'accroissement du jitter dans le cas de la tristesse extravertie (larmoyante), mais de l'autre, ils ne distinguent pas de manière significative les expressions. Ceci est certainement du à un biais de l'estimateur inadéquat à la mesure le jitter. La variable acoustique centroïde de l'intonation ne fera donc pas partie du modèle final.

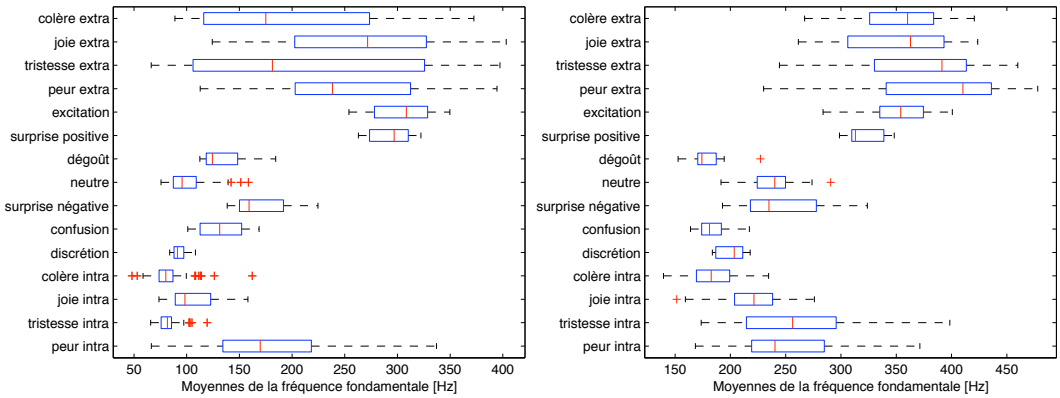

FIG. 3.19: "Boxplots" des moyennes de la fréquence fondamentale des phrases, classés par expression. L'ordre des expressions est relatifs à leur degré d'activation. A gauche : les deux acteurs ; A droite : les deux actrices.

La figure 3.19 montre aussi comment le sexe de l'acteur change la hauteur moyenne de la voix. Si cela est bien connu, il n'en reste que cela peut être aussi vu comme une source de variabilité de la prosodie.

Le modèle proposé doit tenir compte du sexe du locuteur.

3.5.2 Intensité

La figure 3.21 montre que l'intensité perçue, mesurée par la "loudness", est un paramètre fortement corrélé à la fréquence fondamentale. Les phrases sont symbolisées par des points dont la taille est proportionnelle à l'intensité de l'expression, elle-même distinguée par la couleur utilisée. Pour chaque ensemble de phrases de même expression, une gaussienne a été estimée, puis représentée par une ellipse (dont le centre et le contour sont respectivement relatifs aux moyennes et à la covariance de la gaussienne). Dans le plan "moyenne de la f0" vs. "moyenne de l'intensité", le caractère diagonal des axes principaux des ellipses traduit la corrélation entre la f0 et l'intensité.

3.5.2.1 Centroïde de l'intensité

Les conclusions des analyses du centroïde de l'intensité sont semblables à celles données pour le centroïde de l'intonation. La variable acoustique centroïde de l'intensité ne fera donc pas partie du modèle final.

3.5.3 Débit de parole

De multiples liens entre l'expressivité et le débit de parole ont été mis en évidence [Beller 2006b] et sont présentés ci-dessous ainsi que dans l'annexe B.

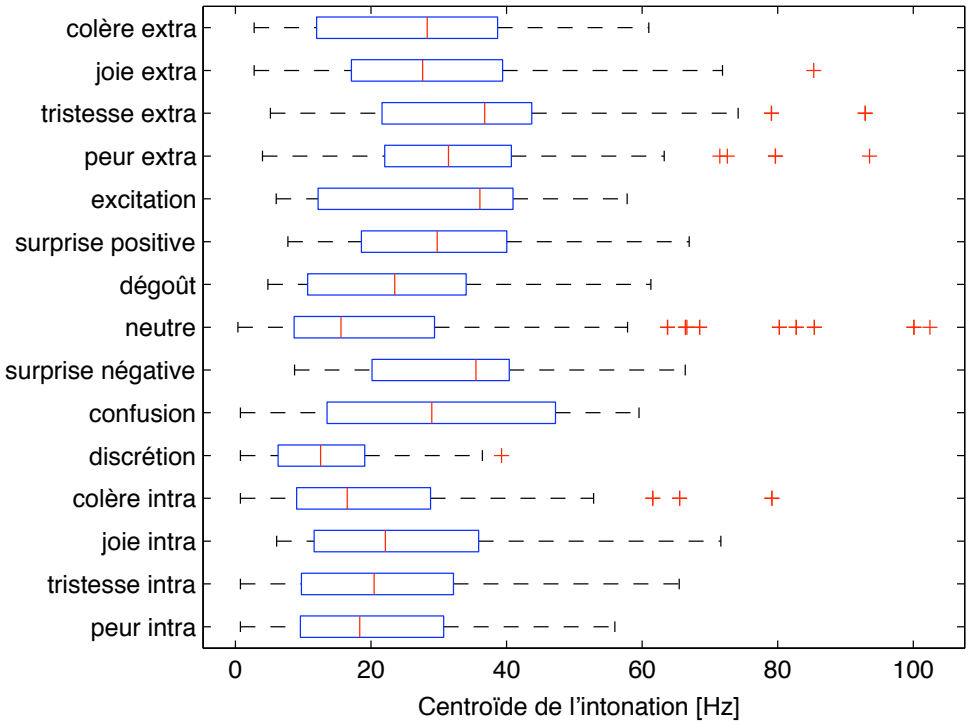

FIG. 3.20: Tous acteur confondus. Estimation du centroïde de l'intonation pour différentes expressions.

3.5.3.1 Distinction entre la joie et la colère

De manière surprenante, de nombreux traits prosodiques communs existent entre la joie et la colère [Chung 2000]. En effet, la figure 3.19 montre que les fréquences fondamentales sont en moyenne assez proches. Aussi, des algorithmes d'apprentissage ont parfois du mal à distinguer ces deux expressions s'ils surpondèrent les effets de la f0. La figure 3.22 présente les expressions du corpus Combe2005, modélisées par des ellipses dont les centres et les axes principaux sont respectivement relatifs aux moyennes et aux variances de la fréquence fondamentale et de la durée des syllabes. Les syllabes prononcées avec joie sont légèrement plus courtes que celles prononcées avec colère. Mais une différence sensible réside dans la variance des durées des syllabes. En effet, les syllabes proéminentes dans le cas de la joie semble durer beaucoup plus longtemps que dans le cas de la colère. Certaines dont la tenue est extraordinairement longue semblent chantées. C'est pour cela que nous avons affiché deux ellipses dans le cas de la joie, une représentant les syllabes proéminentes, et l'autre représentant les syllabes non proéminentes. Dans le cas de la colère, la différence de durée entre ces deux types de syllabes est sensiblement moins importante.

La figure 3.23 montre, tous corpus confondus, les distributions des rapports

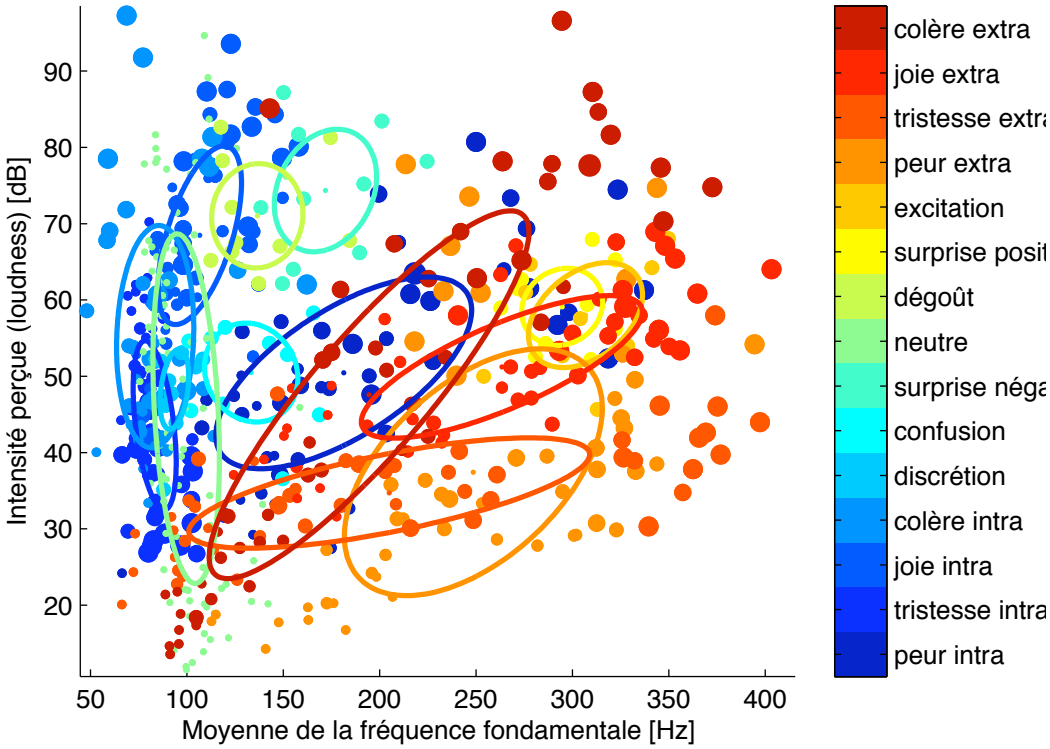

FIG. 3.21: Acteurs masculins. Phrases symbolisées par des points dont la taille est proportionnelle à l'intensité de l'expression, elle-même distinguée par la couleur utilisée. Pour chaque ensemble de phrases de même expression, une gaussienne est estimée, puis représentée par une ellipse (dont le centre et le contour sont respectivement relatifs aux moyennes et à la covariance de la gaussienne). Plan "moyenne de la f0" vs. "moyenne de l'intensité"

de durée entre les syllabes proéminentes et les syllabes non proéminentes, pour chaque expression. La confusion occupe une position marginale vis à vis des autres expressions, car le rapport est bien moindre et parfois inversé.

La proéminence est donc un facteur important de variation prosodique. Même dans le cas neutre, les syllabes proéminentes durent en moyenne, presque deux fois plus longtemps que les syllabes non proéminentes.

Le modèle proposé doit prendre en compte la proéminence.

3.5.3.2 Pauses et respirations

L'annotation des pauses et des respirations audibles dans le corpus Combe2005 a permis de mettre en évidence que la respiration est fortement influencée par l'expressivité. En effet, les respirations audibles sont autant d'indices pouvant faire basculer la perception de l'expression (voir chapitre 3.2.4). L'examen de l'audibilité des respirations et des pauses du corpus Combe2005 repose sur deux rapports. Le

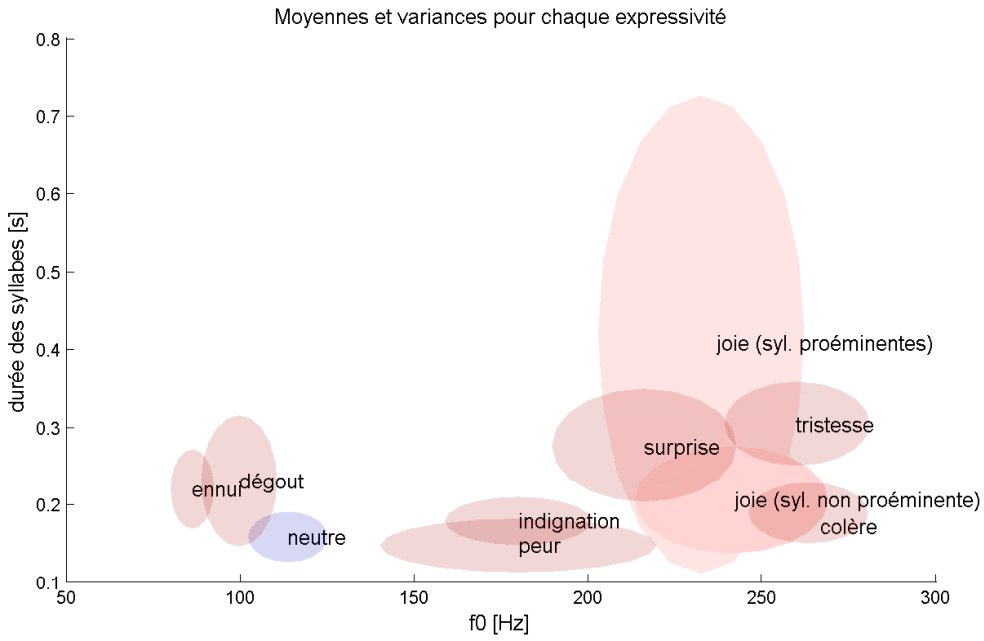

FIG. 3.22: Corpus : Combe2005. expressions modélisées par des ellipses dont les centres et les axes principaux sont respectivement relatifs aux moyennes et aux variances de la fréquence fondamentale et de la durée des syllabes. La joie a été divisée en deux ensembles de syllabes selon leurs degrés de proéminence.

premier divise le temps de pause sur le temps de la phrase (pause ratio). Le second rapport divise le temps de respiration sur le temps de la pause relative (respiration ratio). La figure 3.24 montre comment ces deux rapports permettent une distinction franche entre la surprise négative et la surprise positive.

Parce qu'il n'existe pas de méthodes non-invasives pour mesurer la respiration, la mesure de la pression sous-glottique n'est pas tellement utilisée pour caractériser l'expressivité. Or, il semble clair que de nombreux paramètres prosodiques (intonation, phonation...) sont reliés à la pression sous-glottique. L'introduction d'une nouvelle mesure non invasive de la pression sous-glottique pourrait permettre une avancée importante du côté de la modélisation de la prosodie et de l'expressivité. Une approche expérimentale consisterait à enregistrer la pression sous-glottique en même temps que la parole, au sein d'une même base de données expressive [9].

3.5.4 Degré d'articulation

Une étude concernant l'influence du débit sur le triangle vocalique en parole neutre [Gendrot 2004], montre que les formants tendent vers une voyelle centrale pour les segments de courte durée. Ce phénomène a bien été observé dans la partie 3.3. Une différence majeure existe entre le cas neutre et les autres expressions : le

[9] Cela pourrait aussi constituer un objet de mesure pertinent pour l'interprétation musicale.

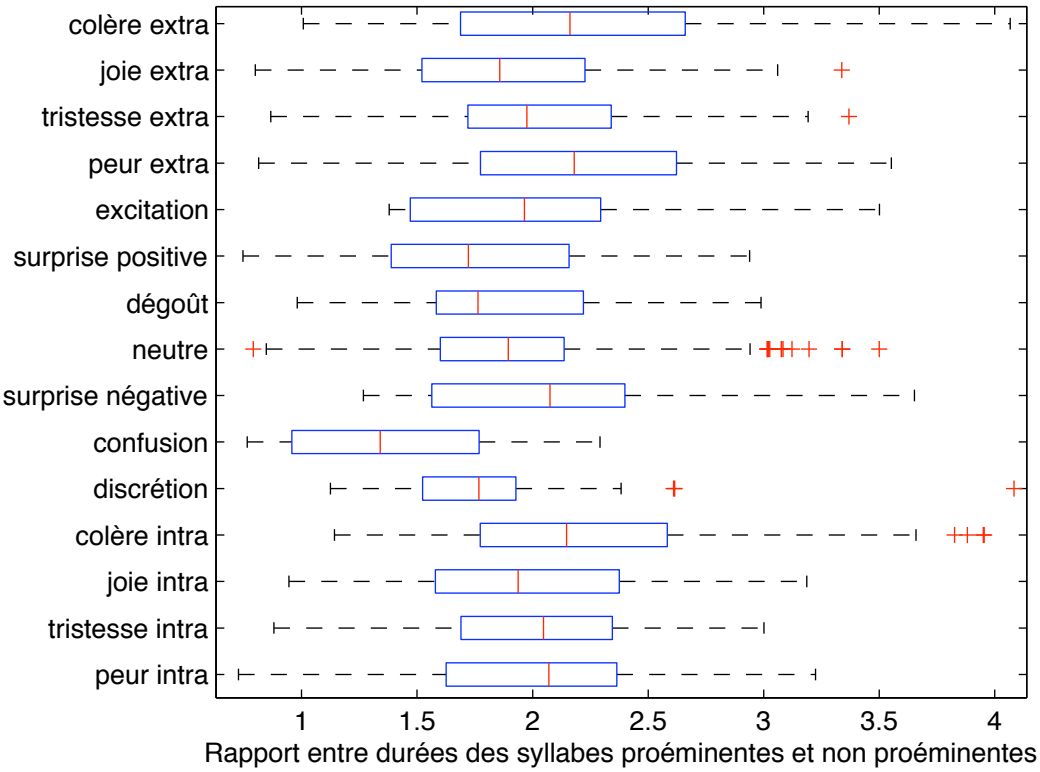

FIG. 3.23: Tous acteurs confondus. "Boxplots" des rapports (estimés par phrase) entre durées des syllabes proéminentes et durées des syllabes non proéminentes, classés par expression.

degré d'articulation n'est plus uniquement dépendant de la variable débit. Dans le cas neutre, une accélération et une décélération correspondent respectivement à une réduction et à une expansion du triangle vocalique. Il semble que cette tendance naturelle ne soit pas préservée dans le cas de certaines expressions.

3.5.4.1 Influence de l'intensité de l'expression

En effet, la figure 3.25 présente quatre triangles vocaliques superposés et mesurés dans le cas neutre (le plus grand) et dans le cas de la tristesse introvertie, pour trois degrés d'intensité différents (tristesse faible, moyenne et forte), sur le corpus Roullier2006. Les voyelles y sont représentées par des ellipses dont les coordonnées du centre et les largeurs sont définies respectivement par les moyennes et les variances des fréquences caractéristiques du 2^{nd} et du 1^{er} formant (voir partie 3.3). Cette figure montre une réduction du triangle vocalique au fur et à mesure que l'intensité augmente. D'un point de vue phonétique, cette réduction se manifeste par un rapprochement du /u/ vers le /a/. La conséquence est que les réalisations du /u/ se rapprochent du lieu acoustique du /o/. D'un point de vue articulatoire, ce rapprochement se traduit par une "délabialisation" et par une augmentation de

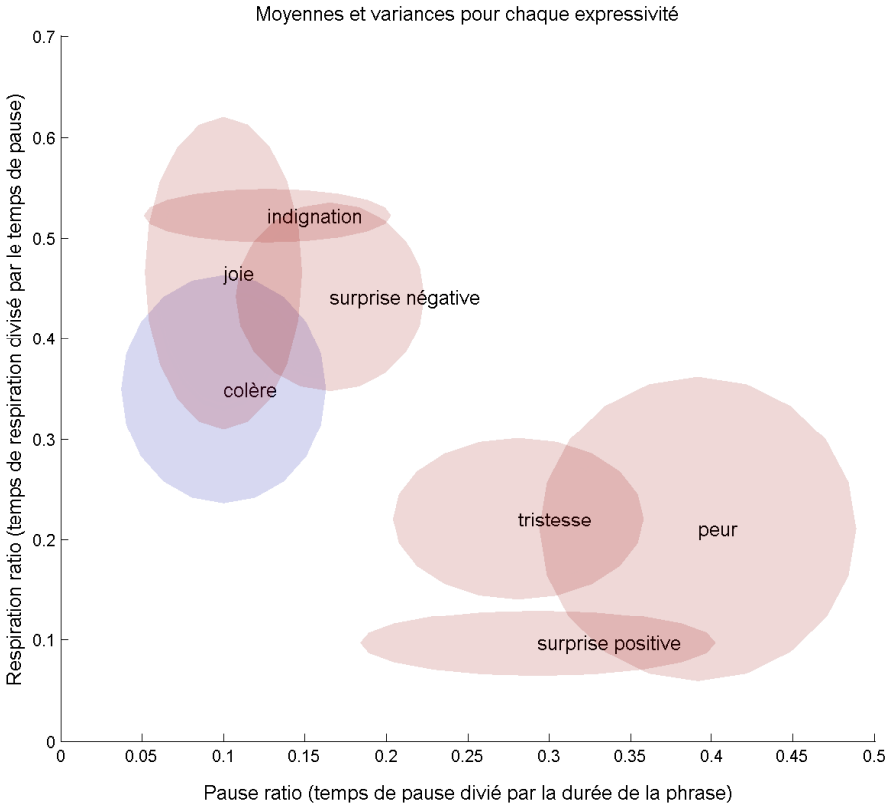

FIG. 3.24: Corpus : Combe2005. expressions représentées par des ellipses dont les centres et les axes sont respectivement les moyennes et les variances de deux rapports : le premier divise le temps de pause sur le temps de la phrase (pause ratio). Le second rapport divise le temps de respiration sur le temps de la pause relative (respiration ratio)

l'ouverture de la bouche. Il s'agit donc d'une version "moindre effort" du /u/ qui traduit une hypoarticulation.

Ce que la figure 3.25 ne montre pas, c'est que le débit syllabique diminue aussi selon l'intensité. Cette tendance dans le cas de la tristesse introvertie est donc inverse à la tendance du cas neutre qui proposerait une expansion du triangle vocalique au fur et à mesure que le débit diminue. Elle est donc bien démonstratrice d'un phénomène d'hypoarticulation de la tristesse introvertie.

3.5.4.2 Influence de l'expressivité sur le degré d'articulation

Ce phénomène et son contraire peuvent être observés selon l'expression. La figure 3.26 présente des résultats relatifs aux corpus Daniele2006 (à gauche) et Roullier2006 (à droite). Pour chacun des corpus, les expressions ont été représentées par des ellipses dont les centres et les largeurs sont respectivement relatives aux

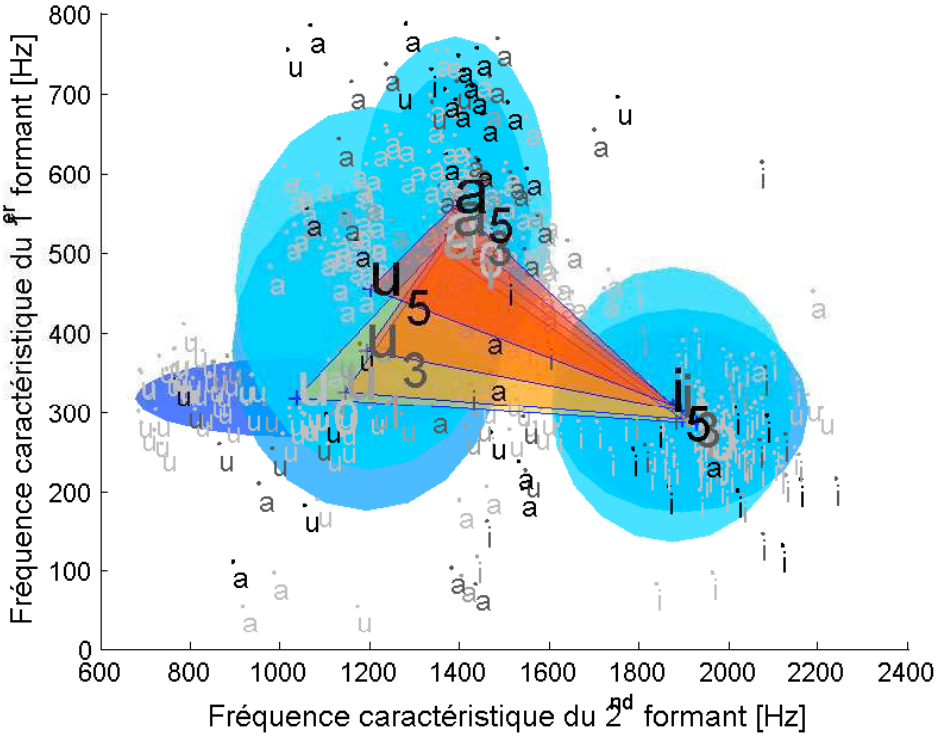

FIG. 3.25: Acteur : Roullier2006. Triangle vocalique neutre et selon trois niveaux d'intensité de la tristesse extravertie (de plus en plus foncé). Les voyelles y sont représentées par des ellipses dont les coordonnées du centre et les largeurs sont définies respectivement par les moyennes et les variances des fréquences caractéristiques du 2^{nd} et du 1^{er} formant.

moyennes et aux variances de l'aire du polygone vocalique (en abscisse) et de la durée des syllabes (en ordonnée). La couleur des ellipses est relative à la moyenne de la fréquence fondamentale (axe "z"). Le polygone vocalique est une extension du triangle vocalique aux autres voyelles (transitoires) fermées. Ses sommets sont définis par les lieux des voyelles /a/ /e/ /i/ /y/ /u/ et /o/, dans l'espace cardinal. La mesure de l'aire de ce polygone est plus robuste que la mesure de l'aire du triangle (/a/ /i/ /u/) car elle contient plus d'information sur la vocalité. Sur chacune des figures a été tracée une ligne arbitraire partant de l'origine et passant par le neutre. Les expressions à gauche de cette ligne sont arbitrairement appelées hypoarticulées, tandis que les expressions à droite sont appelées hyperarticulées.

La première conclusion que l'on peut tirer de cette étude, est qu'il existe des stratégies différentes selon les acteurs (et les sexes peut-être) pour exprimer les mêmes expressions. En effet, les lieux des expressions sont différents pour les acteurs Daniele (femme) et Roullier (homme). Seules certaines expressions semblent exprimées de la même façon par les deux acteurs. Par exemple, la colère extravertie et la surprise négative sont deux expressions clairement hyperarticulées. La peur,

la tristesse, la joie et la colère introverties sont des expressions hypoarticulées. Pour les autres expressions, il est difficile d'établir un consensus par cette étude. En effet, les expressions peuvent être exprimées selon plusieurs stratégies prosodiques, parfois très éloignées. Par exemple, la discrétion est hypoarticulée par l'actrice et hyperarticulée par l'acteur. Il peut y avoir, en effet, différentes stratégies pour la discrétion, selon les positions relatives des interlocuteurs (discrétion hypoarticulée si proximité et hyperarticulée si distance). Cette multiplicité des stratégies existantes pour véhiculer la même expression ajoute au modèle un degré de complexité. Au lieu de "moyenner" les stratégies de tous les acteurs de même sexe dans une seule et même stratégie, nous allons donc créer un modèle par acteur. En plus de l'information du sexe du locuteur,

Le modèle proposé doit tenir compte de l'identité du locuteur.

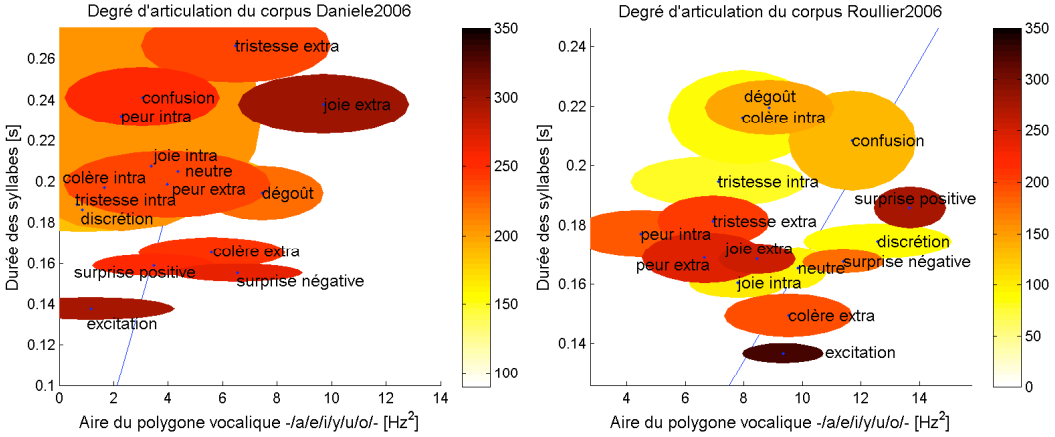

FIG. 3.26: Représentation des expressions dans un espace dont l'abscisse est l'aire couverte par le triangle vocalique [$\times 10^4 Hz^2$], l'ordonnée est la durée des syllabes [s], et la couleur est la fréquence fondamentale [Hz]. A gauche : Daniele2006 ; A droite : Roullier2006.

3.5.5 Degré d'articulation et degré d'activation

Les expressions possèdent des degrés d'activation différents. Si la moyenne de la fréquence fondamentale semble être un bon indice du degré d'activation de l'expression, en est-il de même pour le degré d'articulation ? En effets, les émotions de degré d'activation négatif (passivité) et positif (activité) se manifestent respectivement par l'introversion et l'extraversion du locuteur. Est-ce que l' intro/extra - version de l'expression est responsable d'une parole respectivement hypo/hyper - articulée ? Le tableau 3.1 met en vis à vis, le degré d'activation des expressions et le degré d'articulation moyen mesuré sur chaque corpus. Il semble difficile à la vue de cette confrontation de déduire une règle pour tous les acteurs. Il semble qu'au contraire, ceux-ci se démarquent par des stratégies parfois antagonistes. Toutefois, certains résultats semblent consensuels. Par exemple, les émotions utilitaires

expression	degré d'activation	degré d'articulation			
		Roullier	Combe	Olivia	Daniele
peur intra	passif	-	-	-	-
tristesse intra	passif	-	+	-	-
joie intra	passif	-	+	-	-
colère intra	passif	-	-	+	-
discrétion	passif	+	+	-	-
confusion	passif	-	-	-	-
surprise négative	passif	+	-	-	+
neutre	neutre	0	0	0	0
dégoût	actif	-	-	-	+
surprise positive	actif	+	-	0	0
excitation	actif	+	0	0	-
peur extra	actif	+	+	+	0
tristesse extra	actif	-	-	+	+
joie extra	actif	-	-	0	+
colère extra	actif	+	-	+	+

TAB. 3.1: Tableau recensant les expressions, leur degré d'activation et le degré d'articulation mesuré sur les corpus

introverties (tristesse, peur, joie et colère introverties) sont majoritairement hypo-articulées par les acteurs.

3.5.6 Phonation

Les moyens d'analyse de la phonation sont assez restreints au moment où cette thèse s'écrit, comme l'explique la partie 3.3. Cependant, certaines conclusions tirées de l'annotation paralinguistique et de la mesure du coefficient de relaxation Rd encouragent fortement les futures recherches à se diriger vers la mise en place de méthodes de mesure fiable du mode vibratoire, du voisement et de la qualité vocale.

3.5.6.1 Changements de mode vibratoire

Plusieurs changements de mode vibratoire ont été observés dans la base de données. Malheureusement, leur nombre est trop faible pour permettre une étude statistique du phénomène ou la mise en place d'un algorithme d'estimation du changement de mode vibratoire (modes 0 I II et III). Nous renvoyons donc le lecteur aux analyses symboliques puisque les changements abrupts de modes vibratoires ont été annotés durant l'annotation des phénomènes paralinguistiques (voir partie 3.4). Comme l'indique la figure 3.17 (étiquettes "fp" et "fg"), des changements rapides d'un registre à l'autre ont été observés majoritairement dans les expressions : peur extravertie, colère extravertie, surprise positive et excitation (des expressions à degré d'activation positif).

3.5.6.2 Voisement

De manière surprenante, certaines expressions nous paraissant antagonistes (tristesse/joie) ne se distinguent que par le degré de voisement. Par exemple, un rire et un pleur possèdent la même structure temporelle (succession de bursts (voir annexes D) et ne se distinguent que par le degré de voisement. Un rire vu comme la répétition d'une voyelle bien voisée (tendue) est ainsi facilement transformable en un pleur, par une opération de dévoisement. La colère introvertie et la peur introvertie sont aussi très semblables, excepté sur le plan du voisement (voir chapitre 5). De la même façon que pour les changements de mode vibratoire, le lecteur est invité à se reporter à la partie 3.4, dans laquelle sont recensés les différents types de voisement par expression (voir 3.17, étiquettes "ch" et "nv"). Le dévoisement total et le chuchotement ont été le plus souvent observés pour les expressions : peur introvertie, tristesse introvertie, colère introvertie et discrétion (des expressions à degré d'activation négatif).

3.5.6.3 Qualité vocale

Comme indiqué dans la partie 3.3, la moyenne sur le phone du facteur du coefficient de relâchement/relaxation Rd est utilisé pour modéliser la qualité vocale. Ce coefficient permet de caractériser la qualité vocale : une voix tendue possède un Rd tendant vers $\rightarrow$ 0.3, une voix normale se situe dans le milieu Rd $\simeq$ 1.2 et une voix relâchée possède un Rd tendant vers $\rightarrow$ 2,5. La figure 3.27 présente les distributions de ce coefficient par expression. Après estimation du coefficient Rd sur les corpus de voix d'homme (l'estimation sur les voix de femme n'étant pas encore opérationnelle), la segmentation phonétique est utilisée pour calculer la moyenne du coefficient sur chaque voyelle. Ces moyennes, classées par expressivion, permettent d'afficher les "boxplots" du coefficient Rd pour chaque expression.

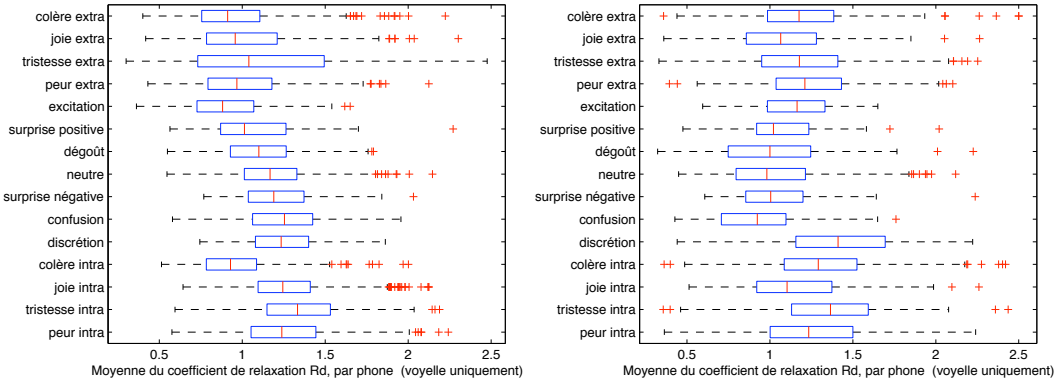

FIG. 3.27: "Boxplots" des moyennes du coefficient de relaxation Rd, estimées par phone (voyelle uniquement) et classées par expression. A gauche : Acteur Combe ; A droite : Acteur Roullier.

Chez les deux acteurs, on constate la même tendance. Le coefficient Rd se rapproche de 0 pour les expressions à forte activation. Ces expressions possèdent donc

une qualité vocale plus tendue que les autres. A l'inverse, la tristesse introvertie et la discrétion semble se manifester par une qualité vocale plus relâchée. Vis à vis de la figure 3.19, on peut se demander si cette mesure n'est pas tout simplement corrélée à la mesure de la moyenne de la fréquence fondamentale. Dans tous les cas, ces résultats sont à nuancer car l'algorithme utilisé a été paramétré sur de la parole neutre et qu'aucun test n'a été mené afin de confirmer sa possible application à de la parole expressive.

3.6 Conclusion

Un message vocal porteur de plusieurs niveaux d'informations est constitué d'une séquence de sons verbaux et de sons non verbaux, modulée par la prosodie et organisée par la double action de la syntaxe et des restructurations. Ces phénomènes paralinguistiques sont surtout observables dans la parole spontanée, encore plus dans le dialogue spontané et davantage encore dans le cas de la parole expressive comme le montre l'examen du corpus expressif IrcamCorpusExpressivity. L'analyse statistique des différentes annotations, ainsi que des paramètres du modèle prosodique appliqué aux corpus, a permis de dresser des conclusions préliminaires à l'élaboration du modèle final. D'autre part, ces résultats participent plus généralement à l'étude de l'expressivité dans la parole. Toutefois, une des conclusions prévient d'une éventuelle généralisation des résultats puisqu'il a été observé que différentes stratégies prosodiques ont été utilisées pour véhiculer la même expression. Ainsi, des variations notoires ont été constatées d'un acteur à l'autre. C'est pourquoi le modèle final comprendra, en réalité, plusieurs modèles, un par acteur. En effet, si le sexe influe fortement sur des "constantes prosodiques" comme les hauteurs moyennes de la voix et des formants, il distingue aussi certaines stratégies prosodiques. La prise en compte de la proéminence dans le modèle final est aussi souhaitée, tant celle-ci distingue profondément les caractéristiques prosodiques des syllabes.

Chapitre 4

Le système Expresso

Sommaire

4.1 Résumé du chapitre

Ce chapitre présente le système Expresso de transformation de l'expressivité de la parole. Plusieurs paradigmes existent pour contrôler la transformation. Ils sont présentés de manière à expliquer le choix d'un modèle de transformation dépendant du contexte. Ce modèle permet de générer des paramètres de transformation, afin de conférer une expression désirée à une phrase neutre, parlée ou synthétisée. La transcription phonétique, le niveau de proéminence, l'identité du locuteur ainsi que les autres niveaux d'information doivent être préservés. Pour cela, le modèle prend en compte ces différentes informations sous la forme de variables symboliques d'entrée. Ces variables symboliques constituent des contextes pour chacune des unités traitées par le modèle : phrase, groupe de souffle, syllabe et phone. Chaque contexte est ensuite utilisé pour inférer des distributions de variables acoustiques, dans le cas neutre et dans le cas de l'expression désirée, grâce à la prise en compte d'une base de données d'exemples constituée des corpus expressifs enregistrés. Enfin, ces distributions sont comparées de manière à fournir des fonctions de transformation. Ces fonctions sont appliquées au signal de phrase à transformer, puis comparées à l'original, dans le but de définir des paramètres de transformation des dimensions prosodiques de la parole.
Trois approches sont proposées. La première repose sur des relations arbitraires entre les paramètres de transformation et les contextes, par le biais d'un jeu de règles cumulatives. La seconde permet d'accroître la complexité du modèle, grâce à l'estimation fréquentiste des distributions des variables acoustiques à partir de la base de données d'exemples. La troisième et dernière approche permet de transformer progressivement le modèle à base de règles en un modèle guidé par les données lors d'une phase d'apprentissage, en reposant sur l'approche bayésienne. Une nouvelle phrase à transformer pouvant présenter un contexte qui n'aurait pas été déjà observé dans les corpus, cette partie s'attache à décrire des algorithmes permettant la généralisation du modèle.
Une fois que les paramètres de transformation sont générés, ils sont appliqués à la phrase par un vocodeur de phase évolué. L'intonation est modifiée par transposition, le débit de parole, par dilatation/compression temporelle, l'intensité, par un gain variable, le degré d'articulation, par un nouvel algorithme permettant la dilatation/compression non linéaire de l'enveloppe spectrale, et la qualité vocale, par filtrage dynamique. Tous ses opérateurs possèdent des paramètres variant dans le temps de façon à modifier les dimensions prosodiques de manière dynamique.
Enfin, ce chapitre présente les résultats d'une évaluation préliminaire du système Expresso, sur la base d'un test perceptif. Conformément à notre définition de l'expressivité, ce test perceptif de mesure directe repose sur la catégorisation de l'expression de stimuli actés et transformés. Une interface web a été implémentée pour permettre une diffusion large du test et ainsi récolter une population importante de participants. Les résultats du test sont encourageants bien que les taux de reconnaissance soient assez faibles. Malheureusement, cette évaluation reste partielle, car elle présente trop peu de stimuli, et problématique, car elle présente trop de classes expressives. Toutefois, elle permet de préparer le terrain d'évaluations futures, et de pressentir certains comportements du système. En moyenne, les performances d'Expresso sont de moitié celles des acteurs. De plus, les matrices de confusion permettent de mettre en évidence qu'un modèle hybride, basés sur différents acteurs selon l'expression, pourraient donner de meilleurs résultats.

4.2 Présentation du système

Expresso est un système de transformation de l'expressivité, dont le but est de conférer une expression choisie, avec un degré d'expression désiré, à une phrase neutre donnée. Cette phrase peut provenir d'un enregistrement de voix réelle ou bien être le produit d'une synthèse TTS (Text To Speech). Dans le premier cas, si la segmentation phonétique n'a pas été effectuée manuellement, elle est fournie par la segmentation automatique. Cette étape nécessite la connaissance du texte pour de meilleurs résultats. Dans le second cas, la segmentation phonétique et le texte sont disponibles comme produits dérivés de la synthèse. La phrase à transformer, appelée phrase source, se présente donc au système sous la forme d'un fichier audio, du texte correspondant et de la segmentation phonétique associée.

La figure 4.1 présente sous la forme d'un schéma, le fonctionnement du système Expresso.

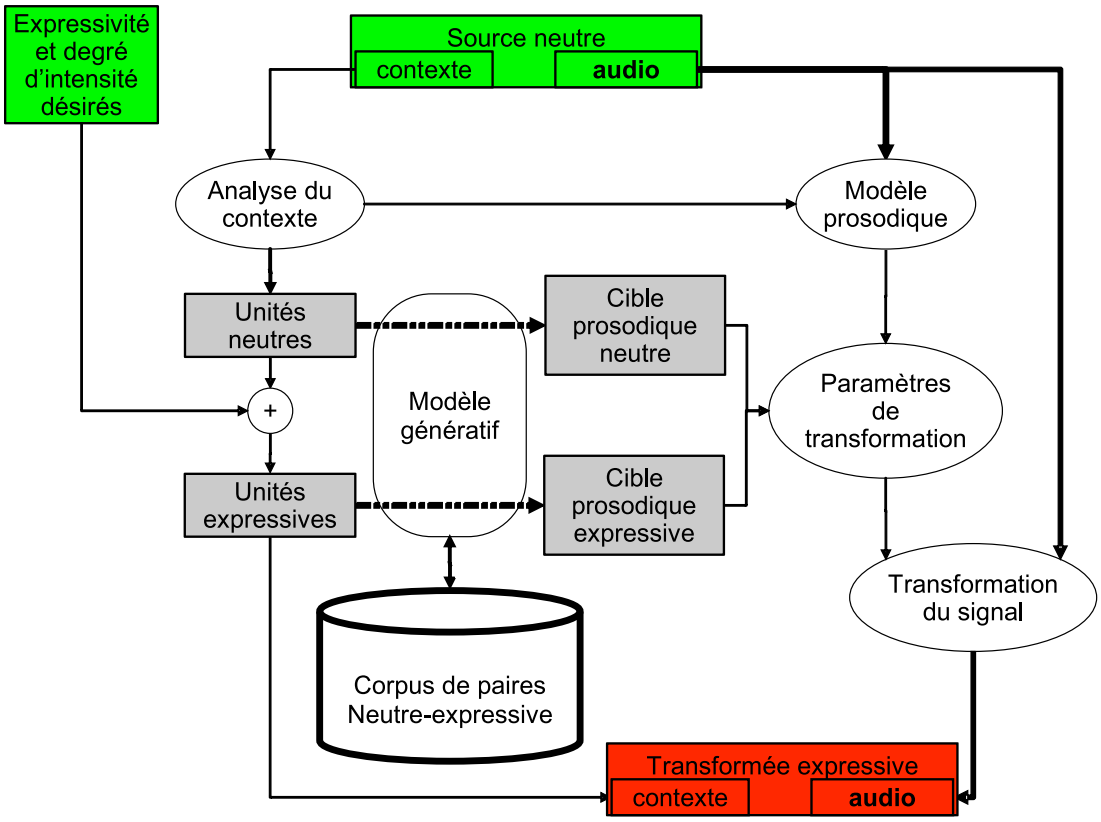

FIG. 4.1: Schéma de présentation du système Expresso.

En haut et en vert, sont présentés les entrées du système. La phrase neutre source à transformer ainsi que l'expression et le degré d'intensité désirés. En bas et en rouge figure la sortie, une phrase transformée se présentant sous la forme d'un

fichier audio muni d'une description largement héritée de celle de la phrase neutre source. De sorte de ne pas modifier les autres niveaux d'information de la parole, Expresso compare les versions neutres et expressives des acteurs, pour ne modéliser que la variation due à l'expression. Le modèle génératif permet de prédire la façon dont un des acteurs enregistrés aurait prononcé la même phrase que la phrase neutre à transformer, dans une version neutre et dans une version expressive. La différence entre ces deux versions permet de définir la variation qu'apporte l'expression. Cette variation est alors appliquée à la phrase neutre source, par des algorithmes de traitement du signal, dans le but de lui ajouter l'expression désirée. Les étapes d'analyse préliminaires permettant de fournir un ensemble d'unités en contexte, ainsi qu'un modèle prosodique, tout deux représentatifs de la phrase à transformer ont été largement décrits dans les précédents chapitres 3.4.3 et 3.2.4. De même, le corpus de paires neutre-expressive a été présenté dans le chapitre 2.6.5. Cette partie s'attache donc à décrire le modèle génératif permettant de fournir des paramètres de transformation, ainsi que les algorithmes de traitement du signal utilisés pour conférer à la phrase neutre, l'expression souhaitée.

4.3 Paramètres des transformations

La définition des paramètres de transformation peut s'effectuer de plusieurs manières, selon qu'elle prend en compte des caractéristiques de l'entrée, des règles arbitraires, des contraintes, des cibles, des modèles ou bien une combinaison de ces différentes informations. En effet, une transformation peut être vue, soit comme une opération invariable, définie par des paramètres absolus, soit comme une opération dépendante de l'entrée, définie par des paramètres relatifs à certaines caractéristiques de l'entrée, soit par la différence entre une source (entrée) et une cible (provenant d'un exemple ou bien d'un modèle), dont les paramètres héritent d'une comparaison entre deux stimuli, l'entrée d'un côté et la cible de l'autre. Ces différents procédés de transformation ne sont que plusieurs visions de la même chose, puisque la comparaison de la source à une cible déduite d'un modèle dépendant de la source, conduit à un ensemble de paramètres de transformation, au même titre qu'un ensemble de règles définies de manière heuristique. La différence entre ces techniques de transformation réside uniquement dans la génération des paramètres, étape que nous appellerons par la suite, le contrôle.

4.3.1 Contrôle

Le contrôle est l'étape de définition des paramètres de transformation permettant de conférer à une phrase neutre X_{neutre}, l'expression désirée $S_{expressivite} = E$ avec un certain degré d'intensité $S_{degree} = D$.

4.3.1.1 Contrôle absolu

Le contrôle peut être défini de manière heuristique et indépendante des caractéristiques de l'entrée. Notre première approche a été de déduire des valeurs de transposition, de dilatation/compression temporelle et de gain, de manière arbitraire, dans un premier temps, puis, de manière experte, à partir de l'observation manuelle du corpus expressif Combe2005. Un nombre restreint de règles a alors été élaboré, puisque les paramètres sont invariants dans le temps et indépendants de l'entrée (donc 3 paramètres à définir par expression et par degré). Par exemple, toute phrase neutre est transposée d'une octave vers le haut pour lui conférer de la joie avec un degré moyen. Bien entendu, cette réduction du problème à un petit ensemble de valeurs simplificatrices ne produit pas de résultats satisfaisants, puisque la parole est un phénomène dynamique et que, par conséquent, le contrôle doit évoluer dans le temps.

4.3.1.2 Contrôle adaptatif

Le contrôle adaptatif permet de faire évoluer les paramètres de transformation, selon certaines caractéristiques de la phrase à transformer X_{neutre}. L'analyse des corpus expressifs nous montre que deux types de variables peuvent caractériser l'entrée, appelée aussi la source.

variable	unité	card	Description
$Sspeaker^{phrase}$	phrase	4	Nom de l'acteur
$Ssexe^{phrase}$	phrase	2	Sexe du locuteur
$Smodalite^{phrase}$	phrase	5	Modalité d'une phrase
$Sproeminence^{syllabe}$	syllabe	5	Proéminence d'une syllabe
$Sphoneme^{phone}$	phone	38	Phonème du phone
$Stexte^{phrase}$	phrase		Texte orthographique
$Sexpressivite^{phrase}$	phrase	15	Expression
$Sdegree^{phrase}$	phrase	6	Degré d'intensité de l'expression

TAB. 4.1: Noms, unités, cardinalités et descriptions des variables symboliques S

D'un côté, des étiquettes symboliques permettent de la caractériser de manière symbolique et/ou catégorielle. Le sexe du locuteur $Ssexe^{phrase}$, la proéminence d'une syllabe $Sproeminence^{syllabe}$ ou encore le phonème prononcé $Sphoneme^{phone}$, en sont des exemples. Ce type de variables est détectable par le préfixe notationnel S, qui traduit le caractère symbolique. Elles peuvent prendre un nombre d'états fini, si leurs valeurs appartiennent à un vocabulaire prédéterminé. On les nomme alors, des variables catégorielles.

De l'autre côté, des grandeurs acoustiques sont estimées par analyse du signal de parole. Ces grandeurs sont des variables continues, c'est à dire qu'elles peuvent prendre une infinité de valeurs bien qu'elles soient le plus souvent bornées. Elles sont précédées d'un A qui signifie, grandeur acoustique. La moyenne de la fréquence fondamentale sur la phrase $Af0^{phrase}_{moyenne}$, la pente du débit de parole sur un groupe de souffle $Adebit^{gps}_{pente}$ ou la courbure de l'intensité sur un phone $Aint^{phone}_{courbure}$, en sont des exemples. Ces variables peuvent posséder plus d'une dimension.

La prise en compte de ces différentes variables d'entrée rend le contrôle des paramètres de transformation, adaptatif. Trois approches se distinguent alors selon que l'on utilise les variables symboliques uniquement, les variables acoustiques uniquement, ou bien la réunion des deux.

Adaptivité symbolique Les approches ne prenant en compte que les variables discrètes sont appelées approches contextuelles. Selon les différentes configurations des variables symboliques, ces approches génèrent des paramètres qui deviennent dynamiques parce que le contexte évolue le long de la phrase à transformer. Cette approche est aussi utilisée par la synthèse concaténative qui sélectionne des unités acoustiques, sur des critères symboliques issus de l'analyse du texte à synthétiser.

Le tableau 4.1 recense les différentes variables symboliques présentées dans ce manuscrit, leur cardinalité, ainsi qu'une courte description pour rappel.

Transformation en temps réel L'utilisation d'information de type symbolique n'est pas chose courante en temps réel. Si leur utilisation par un algorithme temps réel est tout à fait possible, leur génération appartient encore au temps différé.

Deux raisons permettent d'expliquer pourquoi les données symboliques ne sont pas générées en temps réel.

La première raison est de nature théorique. Un segment n'est définissable que dès lors que ces deux frontières temporelles sont intervenues. C'est à dire que l'étiquette d'un segment, n'est définissable que lorsque ce segment s'est déroulé dans son intégralité. Or, cette contrainte va à l'encontre de la définition même du temps réel. Un système temps réel convoqué par une entrée, doit produire une sortie correspondante, au plus tard, au bout d'un temps fini, et défini a priori. Or un segment (une voyelle par exemple) posséde une durée variable qu'il est difficile de borner (surtout en voix chantée). Une alternative est de fixer une limite de durée des segments, a priori, tout en sachant que cela peut engendrer des effets de bord (plusieurs voyelles segmentées au sein d'une seule et même voyelle). Dans tous les cas, plus la durée maximale autorisée est longue, plus le délai entre l'entrée et la sortie du système temps réel l'est aussi. Un compromis existe donc entre la taille maximale des segments et le délai acceptable pour leur traitement.

La deuxième raison qui limite la génération des données symboliques en temps réel, provient de l'aspect subjectif de l'annotation. La plupart des variables symboliques permettent la catégorisation subjective des segments, à divers degrés de subjectivité. Si le sexe d'un locuteur parait un critère objectif, l'identité phonétique d'un phone peut engendrer des divergences entre annotateurs et la proéminence des syllabes encore plus. Cependant, de nombreuses applications permettent l'annotation automatique : les systèmes de reconnaissance de parole [Lanchantin 2008] produisent une séquence de phonèmes censée traduire la prononciation d'une phrase ; les outils d'annotation automatique de la proéminence [Obin 2008] distinguent les syllabes proéminentes sur différents degrés de proéminence. Toutefois, l'utilisation de ces systèmes automatiques reste toujours suivie d'une étape de vérification manuelle. Non seulement parce qu'ils font des erreurs, mais surtout car ils fournissent un résultat qui se veut le reflet d'une catégorisation subjective. Or l'annotation semi-automatique, qui comprend l'étape manuelle de correction d'un résultat automatique, n'est pas possible en temps réel pour des raisons évidentes (à moins que les unités soient très longues par rapport au temps du temps réel). La difficulté à générer des données symboliques, à la volée, réduit considérablement leurs utilisations en temps réel. Toutefois, des systèmes exploitent les résultats d'annotations effectuées en temps différé, pour des traitements en temps réel [Schwarz 2006].

Adaptivité acoustique La deuxième approche ne considère que les variables acoustiques dérivées du signal d'entrée. Les résultats des analyses acoustiques du signal guident alors certains paramètres de transformation qui évoluent tout au long de la phrase. Ces méthodes sont utilisées, par exemple, en conversion de voix, où la phrase source à transformer, est convertie en phrase prononcée par un locuteur cible, grâce à un modèle fondé sur le "mapping" des variables acoustiques de la source à celles de la cible (mapping souvent effectué grâce à un mélange de modèles gaussiens, GMM).

variable	unité	dim	Description
$Af0^{phs}_{model}$	phrase	3	Modèle de l'intonation
$Af0^{gps}_{model}$	gps	3	Modèle d'un résiduel de l'intonation
$Af0^{syl}_{model}$	syllabe	3	Modèle d'un résiduel de l'intonation
$Aint^{phs}_{model}$	phrase	3	Modèle de l'intensité
$Aint^{gps}_{model}$	gps	3	Modèle d'un résiduel de l'intensité
$Aint^{syl}_{model}$	syllabe	3	Modèle d'un résiduel de l'intensité
$Adebit^{phs}_{model}$	phrase	3	Modèle du débit de parole
$Adebit^{gps}_{model}$	gps	3	Modèle d'un résiduel du débit de parole
$Aformant1^{phone}_{model}$	phone	3	Modèle de la fréquence du 1^{er} formant
$Aformant2^{phone}_{model}$	phone	3	Modèle de la fréquence du 2^{eme} formant
$Aformant3^{phone}_{model}$	phone	3	Modèle de la fréquence du 3^{eme} formant
$Aformant4^{phone}_{model}$	phone	3	Modèle de la fréquence du 4^{eme} formant
ARd^{phone}_{model}	phone	3	Modèle du coefficient de relaxation

TAB. 4.2: Noms, unités, dimensions et descriptions des variables acoustiques A

Le tableau 4.2 recense les différentes variables acoustiques présentées dans ce manuscrit, leur cardinalité, ainsi qu'une courte description pour rappel.

Transformation en temps réel Une application de ce paradigme de contrôle adaptatif acoustique a été implémentée, pour modifier l'intonation d'une phrase source, en temps réel. Au fur et à mesure que le signal se présente (sous la forme d'une séquence de trame), la hauteur $Af0^{trame}_{reel}$ est estimée. Une fonction (possiblement non-linéaire) produit, à partir de la hauteur mesurée, une hauteur cible $Af0^{trame}_{cible}$. La comparaison de ces deux hauteurs source et cible produit un facteur de transposition qui est appliqué, en temps réel, par une version temps-réel du vocodeur de phase SuperVP. Quelques exemples de fonctions utlisées sont visibles sur la figure 4.2. Les deux premiers exemples (cas A et B) montrent comment des fonctions non linéaires permettent de réduire et d'augmenter le registre de l'intonation, mais aussi, de réduire et d'augmenter le phénomène de proéminence. En effet, l'utilisation d'une fonction non-linéaire (mais linéaire par morceaux ici), permet un traitment différent selon la valeur de f0. En changeant la pente pour les valeurs extrêmes de f0, le système amplifie de manière différentes les valeurs extrêmes de l'intonation qui sont associables au phénomène de proéminence. La fonction présentée dans le cas C inverse la courbe d'intonation, ce qui peut produire l'effet d'un changement de modalité (ton descendant transformé en ton montant et vice-versa). Enfin, la fonction ludique du cas D permet de discrétiser l'intonation sur une échelle musicale choisie (ici, la gamme tempérée). L'effet produit est un rapprochement perceptif de la parole traitée à du chant (principe de l'harmoniser) puisque les faibles variations de f0 sont écrasées sur des paliers (considérées alors comme des notes). D'autres exemples de fonctions non linéaires produisent des effets intéressants comme la fonction échelon

qui “binarise” l'intonation ou les fonctions faisant intervenir des pentes positives et négatives et qui ont pour effet de rajouter des contours mélodiques à la courbe intonative initiale. De manière à adapter ces fonctions à la phrase à transformer, en particulier au locuteur, il est nécessaire de centrer celles-ci sur la fréquence fondamentale moyenne de la phrase à transformer, et de les borner à l'intervalle de hauteur correspondant. Ces grandeurs peuvent être définies a priori mais cela va à l'encontre d'un fonctionnement en temps réel. Pour éviter ceci, le procédé de transformation utilise des valeurs de moyenne courantes et d'intervalles locales, sans cesse réactualisées. Le choix de la taille de la fenêtre pour le calcul de ces valeurs courantes devient critique et peut engendrer de mauvais fonctionnements si celle-ci est trop courte ou trop longue. De plus, les dilatations/compressions temporelles sont impossibles en temps réel (et cause des problèmes de synchronisation), car on ne peut raccourcir ce qui n'a pas encore été prononcé. Or l'expressivité possède une forte influence sur le débit de parole et celui-ci doit être pris en compte pour une bonne transformation. Ainsi, malgré des résultats convaincants, la transformation de l'expressivité en temps réel ne peut être que moins performante que la transformation de l'expressivité en temps différé. C'est pourquoi, le système Expresso mis au point n'est pas temps-réel.

Adaptivité hybride Enfin, il existe quelques approches utilisant à la fois les variables symboliques et les variables acoustiques, pour transformer la phrase source. Certains synthétiseurs concaténatifs associent à la distance symbolique, une distance acoustique entre unités consécutives, qui réduit certains défauts de la concaténation. Certaines des nouvelles approches de transformation par HMM [Yamagishi 2005], utilisent à la fois les informations symboliques issues du texte, et les informations acoustiques, pour transformer la phrase source. Notre méthode se place dans le cadre de ces approches hybrides dont les paramètres de transformation varient selon des variables symboliques et des variables acoustiques caractérisant l'entrée. Ainsi, le système présenté par la suite, produit des paramètres de transformation par l'utilisation conjointe des variables symboliques du tableau 4.1 et des variables acoustiques du tableau 4.2.

4.3.2 Paradoxe de la transformation paralinguistique

L'utilisation du contrôle adaptatif hybride des paramètres de transformation présente l'intérêt de pouvoir manipuler les variables acoustiques tout en en définissant un contexte d'apparition, grâce aux variables symboliques. Nous avons vu, en effet, qu'une difficulté majeure de l'analyse et du traitement des traits paralinguistiques résident dans l'influence du contenu verbal sur la prononciation. Paradoxalement, l'étude de phénomènes para-verbaux demande une analyse conjointe du phénomène verbal, de manière à dissocier les effets segmentaux des effets suprasegmentaux [Beller 2009a]. Par exemple, le contenu phonétique $Sphoneme^{phone}$ permet de construire le triangle vocalique qui renseigne sur le degré d'articulation. Le sexe du locuteur $Ssexe^{phrase}$ permet de dissocier deux catégories de stimuli qui

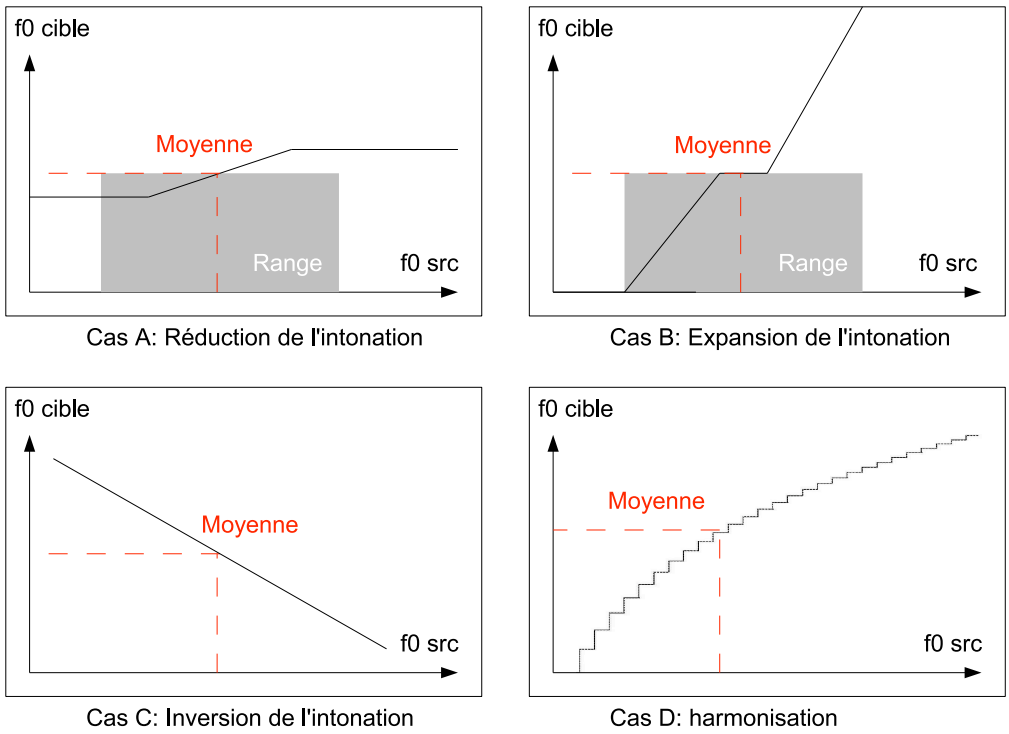

FIG. 4.2: Courbes de transformation non-linéaire pour le temps réel : cas A : réduction de l'intonation, cas B : expansion de l'intonation, cas C : inversion de l'intonation, cas D : harmonisation.

possèdent des fréquences fondamentales moyennes différentes. La connaissance de la nature phonologique associée au phonème permet de n'observer la qualité vocale que sur les segments voisés. Aussi, la plupart des traitements prosodiques reposent, d'abord, sur des classifications phonétiques sur lesquelles vont être définies les unités prosodiques telles que les syllabes ou les groupes de souffle. L'étape de segmentation phonétique est donc cruciale et demande, parfois, une vérification manuelle (voir la partie 4.3.1.2). Toutefois, comme il s'agit de transformer l'expression de phrases neutres, les outils de segmentation phonétique automatique donnent de bons résultats, surtout si le texte correspondant est donné. D'autre part, si la phrase a été synthétisée en amont par un synthétiseur de parole neutre par concaténation d'unités, la segmentation phonétique est alors accessible comme "produit dérivé" de la synthèse. Enfin, il est toujours possible de corriger celle-ci ou bien de la modifier délibérément pour obtenir certains effets désirés.

4.3.3 “Neutralisation” des paramètres

Transformer l'expressivité d'une phrase nécessite, de manière indirecte, la préservation des autres niveaux d'information de la parole. Une phrase enregistrée ou synthétisée comporte, en plus de l'expressivité, l'identité du locuteur, le style de parole, le message sémantique, l'aspect pragmatique et la proéminence (voir partie 2.4 pour le détail). Aussi, la transformation de l'expressivité ne doit pas entamer la perception de ces différents niveaux d'information, jugés indépendants de l'expressivité. C'est pourquoi le modèle fournissant les paramètres de transformation doit comprendre une représentation de ces différents niveaux d'information de la phrase neutre. De plus, si ce modèle est fondé sur des données, il doit utiliser la comparaison de deux versions égales au regard de ces niveaux d'information et dont seule l'expressivité peut les distinguer : une version neutre et une version expressive (voir partie 1.3.2.2). Ceci afin de produire des paramètres de transformation de l'expressivité applicable à une nouvelle phrase, d'un nouveau locuteur, bref, d'un nouveau contexte. Ainsi, les paramètres de transformation permettant de transformer une nouvelle phrase, sont générés par la comparaison d'exemples les plus proches de cette phrase, par rapport aux niveaux d'information de celle-ci, pris dans leur version neutre et dans leur version expressive. Cette étape est appelée “neutralisation des paramètres” et permet de modifier l'expressivité d'une nouvelle phrase, tout en préservant les autres niveaux d'information.

4.4 Modèle génératif

L'utilisation d'un modèle de génération des paramètres de transformation permet d'automatiser la tâche tout en la complexifiant. Plusieurs modèles sont présentés. Tout d'abord, un modèle heuristique simple à base de règles. Puis, un modèle fréquentiste entièrement guidé par les données. Enfin, un modèle permettant la prise en compte, à la fois des règles et des données est décrit. Cette partie s'attache à décrire pour chacun de ces modèles, les phases d'apprentissage, d'inférence et de généralisation.

4.4.1 But d'un modèle génératif

Le but du modèle génératif est de fournir des paramètres de transformation, à partir d'informations extraites de la phrase à transformer. L'adaptivité hybride tient compte à la fois des variables symboliques et des variables acoustiques issues d'analyses de la phrase à transformer (voir partie 4.3.1.2). Une phrase se présente donc à la fois comme une séquence de "contextes", et comme une réalisation acoustique de cette séquence. La segmentation phonétique permet la mise en concordance de ces deux visions d'une seule et même phrase.

4.4.2 Définition du contexte

Un contexte est défini par une combinaison d'état de variables symboliques. C_i est un exemple de contexte qui décrit les caractéristiques d'une syllabe particulière, dont le noyau est un schwa, d'une phrase neutre prononcée par un homme en modalité affirmative :

$$C_1 = \begin{cases} Ssexe^{phrase} & = \text{“homme”} \\ Smodalité^{phrase} & = \text{“affirmation”} \\ Sexpressivite^{phrase} & = \text{“neutre”} \\ Sdegre^{phrase} & = \text{“1”} \\ Sproeminence^{syllabe} & = \text{“AS”} \\ Sphoneme^{phone} & = \text{“/@/”} \end{cases} \tag{4.1}$$

Le tableau 4.1 montre le nombre d'états (*cardinalité*) que peuvent prendre les variables symboliques et catégorielles. Ces variables symboliques sont supposées indépendantes puisque les niveaux phonétique, de proéminence, de l'expressivité et de l'identité du locuteur peuvent se réaliser dans n'importe quelle combinaison (ceci est discuté dans le chapitre 2.4.5 et des résultats le confirmant sont présentés dans le chapitre 3.4.3). Ainsi l'*Univers* $\mathcal{U}$ de notre modèle dépendant du contexte est

composé de :

$$
\begin{aligned}
Card(\mathcal{U}) &= Card(Ssexe^{phrase}) && (2)\\
&\times Card(Smodalité^{phrase}) && (5)\\
&\times Card(Sexpressivite^{phrase}) && (15)\\
&\times Card(Sdegre^{phrase}) && (6)\\
&\times Card(Sproeminence^{syllabe}) && (5)\\
&\times Card(Sphoneme^{phone}) && (38)\\
&= 171000
\end{aligned}
\tag{4.2}
$$

4.4.3 Inférence des paramètres de transformation

Une fois que chacune des unités, composant la phrase neutre à transformer, est renseignée par ses informations contextuelles correspondantes (de manière automatique, supervisée ou manuelle), la phrase est représentée par une séquence temporelle et hiérarchique de contextes C. Puis, deux ensembles de descripteurs acoustiques sont prédits, l'un dans le cas neutre (source) et l'autre dans le cas expressif désiré E (cible). Deux inférences fournissent deux possibles réalisations acoustiques d'une même phrase prononcée de manière neutre ou prononcée de manière expressive. La comparaison de ces deux ensembles de valeurs acoustiques fournit des fonctions de transformation des dimensions prosodiques. L'application de ces fonctions aux valeurs acoustiques, mesurée sur la phrase source à transformer, fournit des valeurs cibles. La comparaison des valeurs source et cible permet alors la définition de facteurs de transposition, de dilatation/compression temporelle, de gain, de réassignement spectral et de modification de la qualité vocale, qui évoluent durant la phrase, puisque le contexte change à chaque phone. Après une phase de lissage de ces paramètres de transformation, un vocodeur de phase [Bogaards 2004] transforme le signal de la phrase neutre selon ces paramètres dynamiques. Ainsi, le problème revient à inférer des ensembles de valeurs acoustiques A correspondant à

un contexte donné : $S = C_i$, i.e. d'évaluer $P(A|S = C_i)$.

→**Initialisation**

- modèle crée;
- phrase source neutre : [N] (audio, texte et segmentation phonétique);
- expression [E] et degré d'intensité expressive [D], désirés;

→**Analyses**

- analyses symboliques de la source [N] :

⇒ création de P unités : phones, syllabes, groupes de souffle et phrase;

- définition des contextes caractérisant la source : $\{C_N(i)\}_{i\in[1:P]}$;
- définition des contextes caractérisant la cible :

$$\forall i \in [1:P], C_E(i) = C_N(i) \begin{cases} Sexpressivite^{phrase} & = \text{E} \\ Sdegre^{phrase} & = \text{D} \end{cases}$$

- analyses acoustiques de la source : A_{source} ;

→**Inférence des variables acoustiques**

for $i \in [1:P]$ **do**

- Inférence des variables acoustiques $A_N(i)$ correspondant à la source : $P(A_N(i)|C_N(i))$;
- inférence des variables acoustiques $A_E(i)$ correspondant à la cible : $P(A_E(i)|C_E(i))$;

end

→**Paramètres de transformation**

- Calcul des fonctions de transformation $T_{N\to E}$ à partir de A_E et A_N ;
- Application des fonctions de transformation à $A_{source} \Rightarrow A_{cible}$;
- Génération des paramètres de transformation par comparaison de A_{source} et A_{cible};
- Lissage des paramètres de transformation;

→**Transformation du signal de parole**

- transposition dynamique ;
- dilation/compression temporelle dynamique ;
- gain dynamique ;
- réassignement spectral dynamique ;
- modification de la qualité vocale dynamique ;

Algorithme 2 : Algorithme de transformation d'une nouvelle phrase.

4.4.4 Modèle à base de règles

Notre première approche dans la transformation de l'expressivité a été basée sur l'écriture directe des paramètres de transformation, uniquement contrôlés par des variables symboliques (adaptivité symbolique), sous la forme de règles [Schroeder 2001, Hozjan 2006]. Les variations prosodiques représentées par des facteurs de transposition, de dilatation/expansion temporelle et de gain on été réglées de manière heuristique. "Afin de transformer une phrase neutre en phrase joyeuse, transposer les parties voisées d'une octave vers le haut" est un exemple de règles

qui ont été écrites à la main et appliquées. Cet ensemble de règles cumulatives a été créé manuellement après observation des exemples enregistrés (voir chapitre 2.6.5). Par exemple, le débit d'une syllabe *neutre* (contexte C_1), est accéléré d'un facteur 1,5 pour transformer celle-ci en *tristesse extravertie* (contexte C_2), puis à nouveau par un facteur 1,8 si elle est accentuée. Une telle règle peut s'écrire :

$$\begin{aligned} Adebit^{syllabe}(C_3) &= 1.8 \times Adebit^{syllabe}(C_2) \\ &= 1.8 \times 1.5 \times Adebit^{syllabe}(C_1) \end{aligned}$$

Pour des utilisations artistiques, le modèle nécessite une certaine flexibilité et doit permettre à l'utilisateur des réglages simples et compréhensibles. Outre l'avantage de ne rien nécessiter au départ (données ou autres), le modèle à base de règles est flexible car ses paramètres sont directement créés par l'utilisateur/expert. La contrepartie est son manque de précision, ainsi que son manque de variété vis à vis de la phrase à transformer. En effet, la mise en place manuelle d'un certain nombre de règles ne permet pas de fournir assez de combinaisons pour couvrir la cardinalité de l'Univers $\mathcal{U}$. C'est la raison pour laquelle nous avons choisi un paradigme d'apprentissage impliquant des données.

4.4.5 Modèle fréquentiste

Les modèles statistiques sont aujourd'hui largement utilisés pour les applications en parole. La reconnaissance de parole [Lanchantin 2008, Gauvain 1996], la conversion de voix [Hsia 2007], et la synthèse de parole [Bulut 2007, Yamagishi 2005, Beller 2004] s'appuient sur des systèmes dont les paramètres sont souvent modélisés par des mélanges de gaussiennes (GMM en anglais, pour Gaussian Mixture Model). Outre de nombreux avantages sur le plan mathématique, ces distributions présentent l'intérêt d'être interprétables puisqu'elles "résument" les données sous la forme de moyennes et de variances. Aussi, certaines caractéristiques des données peuvent se représenter sous la forme de gaussiennes, modèle déjà utilisé pour interpréter certains résultats d'analyse (voir chapitre 3.4.3). S'il existe plusieurs cas présentant le même contexte (plusieurs syllabes dans le contexte C_1, par exemple), on peut définir des moyennes et des variances sur les variables acoustiques qui vont rendre le modèle plus robuste et produire des valeurs plus fiables[1]. Le modèle fréquentiste repose sur l'estimation de ces gaussiennes, modélisant le comportement des variables acoustiques, compte tenu d'une classification par contexte, (issu des données symboliques). En d'autres termes, le modèle se compose d'autant de gaussiennes que de contextes possibles, multiplié par le nombre de variables acoustiques à définir (1 gaussienne par dimension). Soit au total $Ngaussiennes$ gaussiennes pour ce modèle.

$$Ngaussiennes = Card(\mathcal{U}) * \sum_i dim(A_i)) = 171000 * 51 = 8721000 \quad (4.3)$$

[1] Cependant, ce "moyennage" présente l'inconvénient de faire disparaître les données particulières au profit des données les plus fréquentes, c'est à dire proches de leurs barycentres.

Afin d'estimer les paramètres de ce modèle et parce que notre système doit être capable de changer l'expression d'une phrase comme le ferait un acteur, nous utilisons les corpus enregistrés (voir chapitre 2.6.5) en guise de base de données d'exemples. Nous nous plaçons dans le cas où toutes les variables sont complètement observées, c'est à dire qu'il existe au moins une donnée acoustique pour chacun des contextes possibles. Une classification des données acoustiques, par contexte, permet alors de mesurer la moyenne et la variance des observations. Cette mesure est effectuée en utilisant l'estimateur du maximum de vraisemblance (MV). L'estimation du maximum de vraisemblance est une méthode statistique courante utilisée pour inférer les paramètres de la distribution de probabilité d'un échantillon donné.

Soient :

- C_i un contexte donné
- A un descripteur acoustique vu comme une variable aléatoire : $a_{i,j}$ $(j = 1..N)$ est une valeur particulière des N_i données observées dans le contexte C_i
- μ_i et σ_i les paramètres du modèle (respectivement moyennes et variances) que l'on cherche à estimer pour représenter le comportement de A dans le contexte C_i

Alors on définit la fonction de vraisemblance f_i pour le contexte C_i, telle que :

$$f_i(\mu_i, \sigma_i|A) = \begin{cases} f_i(\mu_i, \sigma_i|A) & \text{si A est continue} \\ P_i(\mu_i, \sigma_i|A = a_{i,j}) & \text{si A est discrète} \end{cases} \tag{4.4}$$

$f_i(\mu_i, \sigma_i|A)$ représente la densité de probabilité de A et $P_i(\mu_i, \sigma_i|A = a_{i,j})$ représente une probabilité discrète. Pour la distribution normale $\mathcal{N}(\mu_i, \sigma_i^2)$ qui possède la fonction de densité de probabilité suivante :

$$f_i(\mu_i, \sigma_i|A) = \frac{1}{\sqrt{2\pi}\ \sigma_i} \prod_{j=1}^{N_i} e^{-\frac{(a_{i,j}-\mu_i)^2}{2\sigma_i^2}} \tag{4.5}$$

L'estimateur du maximum de vraisemblance donne accès aux paramètres qui maximisent la fonction $f_i(\mu_i, \sigma_i|A)$:

$$\hat{\mu}_{iML} = \arg\max_{\mu_i} f_i(\mu_i, \sigma_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} a_{i,j} \tag{4.6}$$

$$\widehat{\sigma}_{iML}^2 = \arg\max_{\sigma_i} \ f_i(\mu_i, \sigma_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} (a_{i,j} - \hat{\mu}_{iML})^2 \tag{4.7}$$

L'estimateur du maximum de vraisemblance permet donc d'aboutir à un modèle fréquentiste qui favorise les valeurs les plus fréquentes. Ces estimations faîtes pour chaque contexte, permettent la représentation d'un descripteur acoustique par une gaussienne dont les paramètres (moyenne et variance) varient selon le contexte.

A l'inverse du modèle par règles, le modèle fréquentiste, plus complexe, est incontrable par l'utilisateur. En revanche, il confère au modèle une plus grande complexité et ce, de manière automatique.

4.4.6 Modèle bayésien

De manière à profiter de l'approche experte qui permet la flexibilité et l'introduction de règles arbitraires, et de l'approche fréquentiste qui donne la complexité au modèle, le paradigme de Bayes est appliqué. Après une phase d'apprentissage, le modèle initial à base de règles est, en partie, transformé en un modèle guidé par les données, selon le nombre d'exemples observés disponibles. Les paramètres de transformation dépendants du contexte sont alors inférés par un modèle statistique paramétrique représenté par un réseau bayésien [Naïm 2004].

Ce formalisme nous permet de raisonner sur des probabilités selon des conditions de certitude. Similaire à l'approche fréquentiste, il prend en compte un critère objectif supplémentaire d'optimisation qui incorpore une distribution a priori sur la quantité que l'on souhaite estimer. Cela consiste à chercher les paramètres du modèle les plus probables, sachant que les données ont été observées, et en incluant des "a priori" sur ces paramètres. Cette approche conduit aux estimateurs de l'espérance a posteriori (EAP) ou du maximum a posteriori (MAP). L'estimateur du maximum a posteriori (MAP), tout comme la méthode du maximum de vraisemblance, est une méthode pouvant être utilisée afin d'estimer un certain nombre de paramètres inconnus, comme par exemple les paramètres d'une densité de probabilité, reliés à un échantillon donné. Cette méthode est très liée au maximum de vraisemblance mais en diffère toutefois par la possibilité de prendre en compte un a priori non uniforme sur les paramètres à estimer.

La méthode du maximum a posteriori consiste à trouver les valeurs $\hat{\mu}_{iMAP}$ et $\widehat{\sigma}^2_{iMAP}$ qui maximisent la grandeur $f_i(\mu_i, \sigma_i, |A)p_i(\mu_i, \sigma_i)$ où $f_i(\mu_i, \sigma_i|A)$ est la fonction de vraisemblance pour le contexte C_i et $p_i(\mu_i, \sigma_i)$ la distribution a priori des paramètres μ_i et σ_i dans ce même contexte C_i.

Supposons que nous avons un a priori sur la moyenne μ_i donnée par $N(\mu_{\mu i}, {\sigma_{\mu_i}}^2)$. La fonction à maximiser s'écrit alors :

$$f_i(\mu_i, \sigma_i, |A)p_i(\mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_{\mu i}} e^{\frac{1}{2}\left(\frac{\mu_{\mu i}-\mu_i}{\sigma_{\mu_i}}\right)^2} \prod_{j=1}^{N_i} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\left(\frac{a_{i,j}-\mu_i}{\sigma_i}\right)^2} \tag{4.8}$$

Ce qui revient à minimiser la quantité suivante :

$$\sum_{j=1}^{N_i} \left(\frac{a_{i,j}-\mu_i}{\sigma_i}\right)^2 + \left(\frac{\mu_{\mu i}-\mu_i}{\sigma_{\mu i}}\right)^2 \tag{4.9}$$

Dans ce cas, l'estimateur MAP est donné par :

$$\hat{\mu}_{iMAP} = \frac{\frac{\mu_{\mu i}}{{\sigma_{\mu_i}}^2} + \frac{1}{\sigma_i^2}\sum_{j=1}^{N_i} a_{i,j}}{\frac{1}{{\sigma_{\mu_i}}^2} + \frac{N_i}{\sigma_i^2}} \tag{4.10}$$

On remarque que, dans le cas ou l'on ne possède pas d'information a priori, c'est à dire que $p_i(\mu_i, \sigma_i)$ est uniforme, ce qui se traduit, dans ce cas, par : $\sigma_{\mu i} \to \infty$, l'estimateur du MAP tend vers l'estimateur du MV : $\hat{\mu}_{iMAP} \to \hat{\mu}_{iMV}$. A l'inverse,

si les données sont peu nombreuses, voire manquantes, alors $N_i \rightarrow 0$ et l'estimateur du MAP tend alors vers l'a priori : $\hat{\mu_i}_{MAP} \rightarrow \mu_{\mu i}$. Ainsi, $\sigma_{\mu i}$ et N_i peuvent alors être vus comme des paramètres permettant au modèle de se diriger plutôt vers l'a priori (règles arbitraires) ou plutôt vers les données. Le rapport entre les variances σ_i des données et $\sigma_{\mu i}$ de l'a priori agit en tant que balance entre données de départ (a priori) et données observées.

4.4.6.1 Réseau bayésien

Les réseaux bayésiens ont été utilisés dans différents domaines du traitement de la parole. Des classifieurs bayésiens naïfs et des réseaux bayésiens dynamiques sont utilisés en reconnaissance des émotions [Ball 2003]. L'approche bayésienne permet, dans notre cas, de réunir l'approche par règles et l'approche fréquentiste. Le modèle à base de règles est toujours utilisé pour l'initialisation de la phase d'apprentissage. L'application de la loi de Bayes permet le calcul de distributions des paramètres acoustiques qui concordent mieux aux données selon le nombre d'observations disponibles par contexte. La boîte à outil Matlab® pour les réseaux bayésiens [Murphy 2001] calcule efficacement les distributions de probabilités conditionnelles des variables discrètes et continues. L'hétérogénéité de la nature de ces variables confère au modèle la capacité d'être dépendant du contexte. Un réseau bayésien se compose d'une description qualitative représentée par un graphe et d'une description quantitative représentée par une fonction de densité de probabilité généralisée (ou jointe) $GPDF$. Dans notre cas, cette fonction réunit les dépendances entre variables symboliques contextuelles et variables acoustiques issues de la modélisation :

$$GPDF = P(A, S) \tag{4.11}$$

4.4.6.2 Partie qualitative : modèle graphique

La structure du modèle graphique est ici donnée arbitrairement (elle peut être apprise) et présentée dans la figure 4.3. Cette vision qualitative du modèle statistique montre les variables impliquées durant les phases d'apprentissage et d'inférence. Les rectangles représentent les variables discrètes composant le contexte symbolique. Les cercles représentent les variables continues qui sont des vecteurs composés des trois coefficients issus de la modélisation sous forme de polynômes de Legendre d'ordre 2 (voir chapitre 3.4.3). Les rectangles arrondis et grisés représentent les dimensions acoustiques qui sont concernées par les variables acoustiques. Les flèches représentent les dépendances entre les variables.

Le premier graphe utilisé était un graphe complètement connecté, c'est à dire que chaque variable acoustique dépendait de toutes les variables symboliques et toutes les variables étaient connectées à toutes les autres. Certaines dépendances ont ensuite été enlevées lorsque deux variables ont été jugées indépendantes. L'intelligence du modèle bayésien repose sur la définition de ces indépendances qui va simplifier le calcul de la partie quantitative. Nous nous sommes donc interrogés sur l'indépendance de chaque variable acoustique vis à vis des variables du contexte.

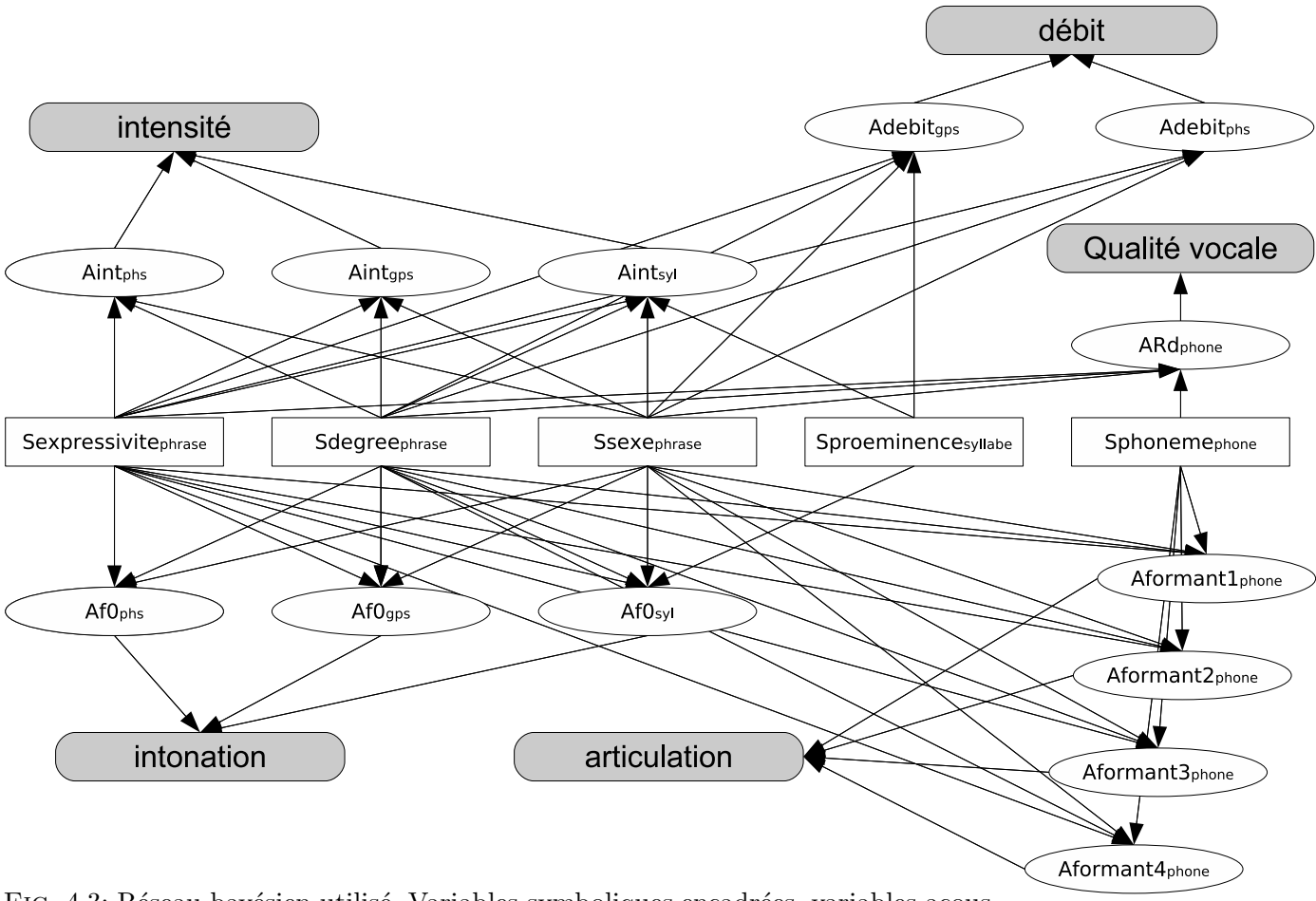

FIG. 4.3: Réseau bayésien utilisé. Variables symboliques encadrées, variables acoustiques entourées et dimensions prosodiques affectées en gris. Les flèches représentent les dépendances.

Par exemple, la séquence phonétique n'est pas une information pertinente pour déduire le débit syllabique de parole. La flèche reliant les variables $Sphoneme_{phone}$ et $Adebit_{gps}$ a donc été ôtée. Les flèches restantes représentent les dépendances entre les variables. Par exemple, le modèle de l'intonation estimé sur un groupe de souffle $Af0_{gps}$ est non seulement dépendant de l'expression et du degré d'expression, mais aussi du sexe du locuteur. C'est le même cas au niveau de la syllabe, excepté que $Af0_{syl}$ est aussi dépendant du niveau de proéminence de la syllabe.

4.4.6.3 Partie quantitative : densité de probabilité généralisée

La fonction de densité de probabilité généralisée $GPDF$ quantifie toutes les dépendances entre les variables du réseau bayésien. Chaque variable continue se voit attribuer une distribution gaussienne linéairement conditionnelle (LCG), dépendante de la configuration de ses variables parentes discrètes. La $GPDF$ est

estimée grâce à la loi de Bayes :

$$P(A,S) = P(A|S)P(S) \tag{4.12}$$

La phase d'apprentissage du réseau bayésien consiste à estimer la $GPDF$. La phase d'inférence utilise cette même $GPDF$, une fois estimée. Les distributions LCG $P(A|S = C_i)$ des variables acoustiques sont inférées en utilisant l'inverse de l'équation 4.12. Si la phase d'apprentissage peut se révéler assez longue, la phase d'inférence est, elle, relativement immédiate.

4.4.6.4 Généralisation

A l'instar de la méthode fréquentiste, la phase d'apprentissage utilise les données des corpus présentées dans le chapitre 2.6.5. Malheureusement, la taille de ces corpus ne produit pas assez d'exemples pour couvrir tout l'*Univers* $\mathcal{U}$. C'est à dire que pour certains contextes, des données acoustiques sont manquantes. Afin de permettre l'estimation de tous les paramètres dans tous les contextes possibles, deux solutions s'offrent à nous : la première consiste à réduire la taille de l'Univers $\mathcal{U}$. En effet, les corpus ne présentent pas, par exemple, une variation de la modalité. La variable $Smodalite^{phrase}$ est donc ôtée de la définition du contexte, ce qui permet d'en réduire la cardinalité d'un facteur 5 ($Card(\mathcal{U}) = 34200$). La deuxième solution permettant l'estimation de tous les paramètres pour tous les contextes, repose sur la capacité du modèle à généraliser, c'est à dire à inférer des grandeurs acoustiques pour certains contextes qui n'ont pas été observés. Plusieurs cas se distinguent alors : l'observation complète, l'observation partielle et l'observation incomplète. Le schéma 4.4 présente ces différents cas de figure.

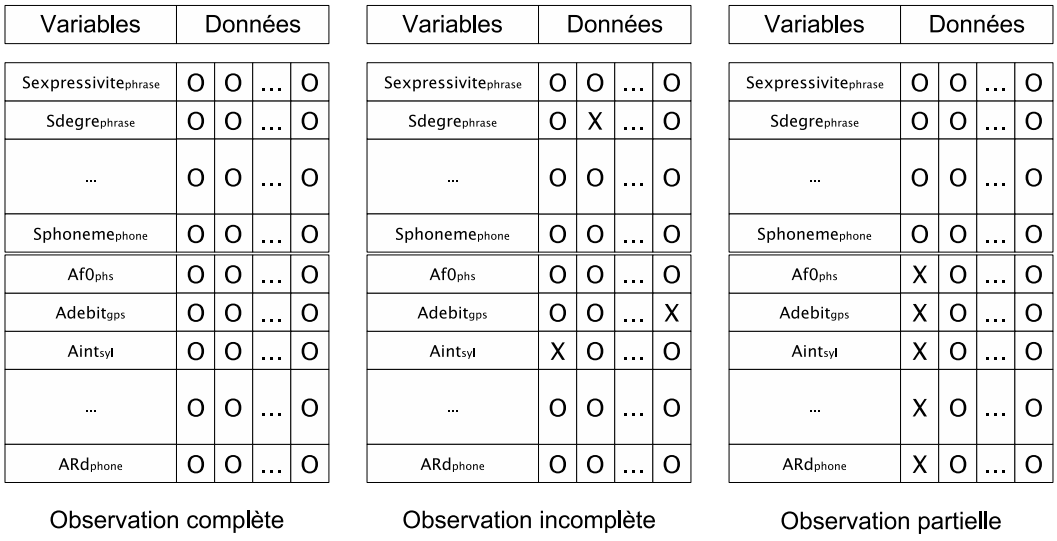

FIG. 4.4: Trois cas d'observation. Les cercles décrivent la présence d'une donnée tandis que les croix manifestent leur absence.

Observation complète Si la base de données permet une couverture complète de l'Univers $\mathcal{U}$, c'est dire que tous les contextes ont été observés acoustiquement ou encore que nous possédons une mesure des variables acoustiques A dans chaque contexte $S \in \mathcal{U}$, alors il est possible d'estimer la distribution des variables acoustiques pour chacun des contextes. Une nouvelle phrase (n'appartenant pas à la base de données) présente alors une nouvelle séquence de contextes dont chacun a été au préalable observé séparément dans les corpus. Dans le cas du réseau bayésien présenté, les types de contexte varient selon les variables acoustiques, et, par conséquent, les cardinalités des Univers aussi. Par exemple, l'Univers $\mathcal{U}_{Af0^{syllabe}}$ de la variable $Af0^{syllabe}$ possède une cardinalité :

$$\begin{aligned} Card(\mathcal{U}_{Af0^{syllabe}}) &= Card(Ssexe^{phrase}) && (2) \\ &\times Card(Sexpressivite^{phrase}) && (15) \\ &\times Card(Sdegre^{phrase}) && (6) \\ &\times Card(Sproeminence^{syllabe}) && (5) \\ &= 900 \end{aligned} \tag{4.13}$$

Tandis que l'Univers $\mathcal{U}_{ARd^{phone}}$ a une cardinalité plus grande puisque cette variable prend en compte le phoneme :

$$\begin{aligned} Card(\mathcal{U}_{Af0^{syllabe}}) &= Card(Ssexe^{phrase}) && (2) \\ &\times Card(Sexpressivite^{phrase}) && (15) \\ &\times Card(Sdegre^{phrase}) && (6) \\ &\times Card(Sphoneme^{phone}) && (38) \\ &= 6840 \end{aligned} \tag{4.14}$$

Dans le cas d'un réseau bayésien, complètement observé, et dont le graphe est connu, la phase d'apprentissage est triviale. En revanche, cette phase est plus ardue s'il y a des données manquantes ou si le graphe n'est pas connu.

Observation incomplète Durant la phase d'apprentissage, il se peut que certaines données ne soient pas disponibles. Par exemple, une analyse acoustique du signal est manquante par morceaux. Une technique généralement utilisée dans ce cas consiste à inférer ces données manquantes à partir de la $GPDF$, puis à réestimer celle-ci en prenant en compte ces données générées, et ceci plusieurs fois : c'est l'algorithme EM (Espérance-Maximisation). Il alterne des étapes d'évaluation de l'espérance (E), où l'on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées, et une étape de maximisation (M), où l'on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E. On utilise ensuite les paramètres trouvés en M comme point de départ d'une nouvelle phase d'évaluation de l'espérance, et l'on itère ainsi. L'algorithme EM permet de trouver le maximum de vraisemblance des paramètres de modèles probabilistes lorsque le modèle dépend de variables latentes non observables. En réalité, ce cas de figure ne s'est pas présenté dans notre problème, puisque nous avons fais attention à ce que toutes les données acoustiques existent pour les contextes observés.

Observation partielle Une nouvelle phrase peut présenter un contexte qui n'a pas été observé durant la phase d'apprentissage si les données ne couvrent pas tout l'Univers $\mathcal{U}$. Dans ce cas, le modèle génératif doit tout de même fournir des paramètres de transformation. Trois solutions ont été envisagées.

Utilisation de l'algorithme EM L'algorithme EM présenté précédemment est souvent utilisé pour les problèmes faisant intervenir des variables cachées (HMM par exemple). Ainsi, ce procédé peut être utilisé pour parer le problème de l'observation partielle en réunissant toutes les variables contextuelles au sein d'une seule et même variable cachée. Cette variable prend alors un nombre d'état inférieur ou égal à la cardinalité de l'Univers $\mathcal{U}$. Malheureusement, cette variable ne permet pas l'interprétation. C'est pour cela que cette solution n'a pas été implémentée.

Contexte le plus proche La seconde solution consiste à prendre en compte les données acoustiques de la phrase à transformer durant la phase d'inférence, pour définir le contexte observé le plus proche. Deux phases d'inférence sont alors nécessaires. La première utilise directement les valeurs des variables acoustiques mesurées sur la phrase neutre (à transformer) afin d'inférer la séquence de contextes la plus probable. L'expression et le degré d'intensité désirés sont alors ajoutés/modifiés à ces contextes. La seconde phase d'inférence permet alors de prédire des données acoustiques. cela permet de déduire des paramètres de transformation pour tous les contextes possibles en procédant par analogie [van Santen 2003]. Cette solution est

décrite par l'algorithme 3.

→**Initialisation**

- modèle crée : $GPDF$ estimée sur les contextes observés : $\mathcal{U}_{observed}$;
- phrase source neutre : [N] (audio, texte et segmentation phonétique);
- expression [E] et degré d'intensité expressive [D] désirés;

→**Analyses**

- analyses symboliques de la source [N] :
  ⇒ création de P unités : phones, syllabes, groupes de souffle et phrase;
- définition des contextes caractérisant la source : $\{C_N(i)\}_{i\in[1:P]}$;
- définition des contextes caractérisant la cible :

$$\forall i \in [1:P], C_E(i) = C_N(i) \left\{ \begin{array}{ll} Sexpressivite^{phrase} & = \text{E} \\ Sdegre^{phrase} & = \text{D} \end{array} \right.$$

- analyses acoustiques de la source : A_{source} ;

→**Inférence des variables acoustiques**

for $i \in [1:P]$ **do**

- vérifier si le contexte a été observé ou non :

if $C_N(i) \in \mathcal{U}_{observed}$ **then**

- Inférence des variables acoustiques $A_N(i)$ correspondant à la source : $P(A_N(i)|C_N(i))$;
- Inférence des variables acoustiques $A_E(i)$ correspondant à la cible : $P(A_E(i)|C_E(i))$;

else

- Inférence des contextes $C_N(i)$ correspondant aux données acoustiques A_{source} ;
- définition des contextes cible correspondant : $C_E(i)$

$$\forall i \in [1:P], C_E(i) = C_N(i) \left\{ \begin{array}{ll} Sexpressivite^{phrase} & = \text{E} \\ Sdegre^{phrase} & = \text{D} \end{array} \right.$$

- Inférence des variables acoustiques $A_E(i)$ correspondant à la cible : $P(A_E(i)|C_E(i))$;

end

end

→**Paramètres de transformation**

- Calcul des fonctions de transformation $T_{N\to E}$ à partir de A_E et A_N ;
- Application des fonctions de transformation à $A_{source} \Rightarrow A_{cible}$;
- Génération des paramètres de transformation par comparaison de A_{source} et A_{cible};
- Lissage des paramètres de transformation;

→**Transformation du signal de parole**

- transposition dynamique ;
- dilation/compression temporelle dynamique ;
- gain dynamique ;
- réassignement spectral dynamique ;
- modification de la qualité vocale dynamique ;

Algorithme 3 : Algorithme de transformation d'une nouvelle phrase, présentant un contexte non observé durant l'apprentissage.

Cet algorithme présente l'intérêt de fournir des paramètres de transformation par analogie, quelque soit le contexte. Pour les contextes non observés, une phase d'inférence permet de trouver dans le modèle le contexte le plus proche, au sens d'une distance basée sur les variables acoustiques. Or, les valeurs des variables acoustiques représentant la phrase à transformer peuvent s'avérer assez lointaines de celles présentes dans la base de données, puisque les locuteurs peuvent être différents. C'est le point faible de la méthode qui réserve son utilisation au seul cas où la phrase à transformer provient d'un locuteur des corpus d'entraînement.

Marginalisation du contexte : La troisième et dernière méthode envisagée, permettant la génération des paramètres de transformation dans le cas de l'observation partielle, repose sur la diminution du contexte lors de l'inférence. Cette diminution est appelée marginalisation du contexte, à cause de sa similitude avec l'opération de marginalisation de la $GPDF$. La marginalisation de la $GPDF$ revient à fixer des valeurs pour certaines variables, et à réestimer la $GPDF$ compte-tenu de cette information ajoutée. C'est d'ailleurs cette opération qui est à la base de l'inférence dans un réseau bayésien. Marginaliser le contexte revient à diminuer celui-ci en enlevant certaines variables symboliques de la description. Si l'on ne possède pas d'exemples acoustiques correspondant au contexte C_1 (voir exemple de la partie 4.4.2), on peut espérer en avoir en réduisant celui-ci au contexte marginalisé C_1^{marg} suivant :

$$C_1^{marg} = \begin{cases} Ssexe^{phrase} & = \text{“homme”} \\ Smodalité^{phrase} & = \text{“affirmation”} \\ Sexpressivite^{phrase} & = \text{“neutre”} \\ Sdegre^{phrase} & = \text{“1”} \\ Sproeminence^{syllabe} & = \text{“AS”} \end{cases} \tag{4.15}$$

Autrement dit, nous avons “simplifié” le contexte en ôtant de celui-ci la variable $Sphoneme^{phone}$. Une hiérarchisation du contexte est nécessaire pour que l'algorithme choisisse quelle variable ôter. Car cette procédure de marginalisation du contexte est répétée de manière itérative jusqu'à ce que le contexte réduit corresponde à un des contextes observés durant l'apprentissage. Cette hiérarchisation est représentée par l'arbre de la figure 4.5. Si le contexte non observé est d'abord marginalisé par rapport à la variable $Sphoneme^{phone}$, c'est que cette information est considérée comme moins importante que les autres.

Cette réduction adaptative de l'information est appliquée ici dans le cas où aucun exemple n'est disponible. Mais d'autres contraintes peuvent être introduites dans cette méthode d'inférence particulière. En effet, si un exemple est suffisant pour obtenir des valeurs acoustiques, il se peut qu'il fournisse des paramètres de transformation peu fiables, car ses données peuvent être très singulières. De manière à rendre plus fiables les paramètres de transformation, on peut choisir un nombre minimum de candidats qui peut correspondre, par exemple, au nombre minimum d'individus pour obtenir des grandeurs statistiques significatives. Cette contrainte

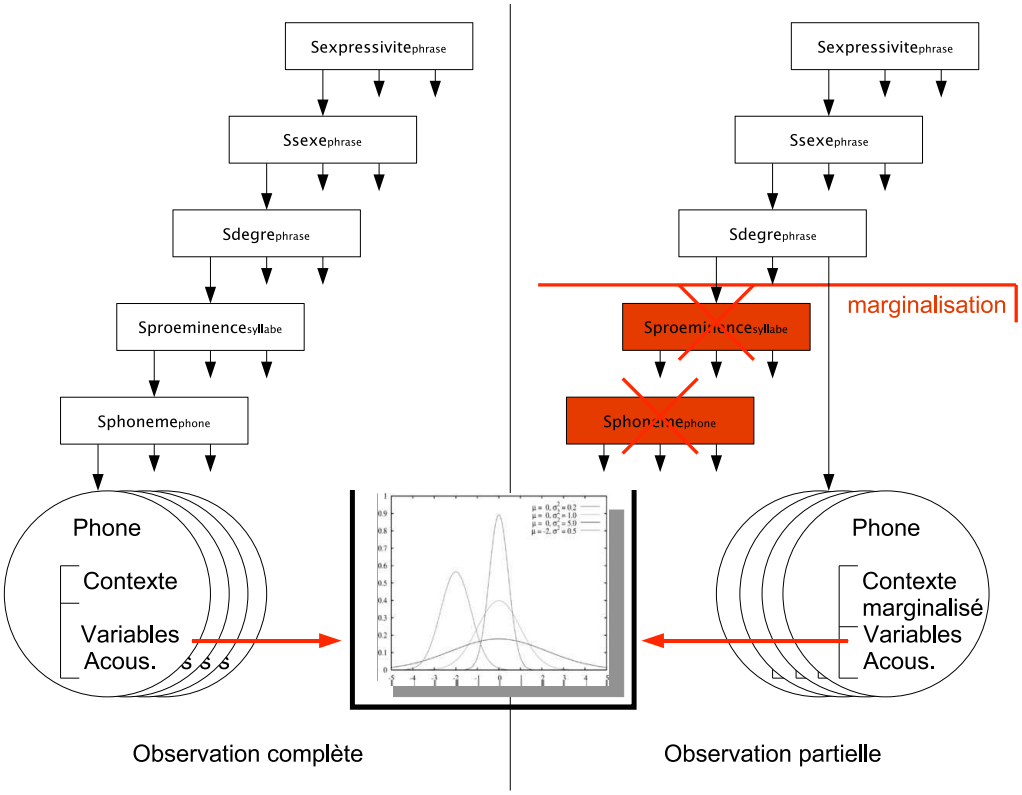

FIG. 4.5: Arbre de décision associé à la méthode d'inférence de marginalisation du contexte. A gauche, cas de l'observation complète du contexte demandé. A droite, cas de l'observation partielle dans lequel le contexte est marginalisé par rapport aux variables $Sproeminence^{syllabe}$ et $Sphoneme^{phone}$, afin de trouver un exemple dans la base de données.

est ajoutée dans l'algorithme qui va marginaliser successivement le contexte, jusqu'à en trouver un réunissant un nombre minimum d'exemples. D'autres contraintes peuvent être utilisées comme des tests sur la variance ou d'autres contraintes arbitraires.

La généralisation permet, paradoxalement, de complexifier le contexte. On peut y ajouter d'autres variables symboliques, comme la nature phonologique des syllabes ou des phonèmes, ou encore la position relative d'un groupe de souffle dans la phrase. Dans tous les cas, l'algorithme précédent trouvera des exemples-candidats qui correspondent en tout ou partie au contexte désiré. Nous avons donc rajouter des variables symboliques pour définir le contexte. Elles sont décrites dans le tableau 4.3

La hiérarchie des variables symboliques au sein du contexte est définie schématiquement dans la figure 4.6. Cette hiérarchie est différente selon le type d'unité choisie. Pratiquement, quatre phases d'inférence se succèdent pour définir les variables acoustiques associées.

variable	unité	card	Description
$Sphono^{phone}$	phone	12	Catégorie phonologique
Scv^{phone}	phone	2	Voyelle ou consonne
$SphonoStruct^{syllabe}$	syllabe	3	V, CV, CVC
$SnbSyl^{gps}$	gpe de souffle	10	Nombre de syllabes

TAB. 4.3: Noms, unités, cardinalités et descriptions des variables ajoutées

4.5 Post-traitements

Une fois que les cibles acoustiques ont été inférées pour chacune des unités, elles sont réunies en deux réalisations possibles de la phrase, une dans le cas neutre, et l'autre dans le cas expressif désiré, grâce au modèle hiérarchique. Le système dispose alors de la réalisation acoustique de la phrase neutre source à transformer, de la réalisation acoustique neutre de cette même phrase inférée par le modèle et de la réalisation acoustique expressive de cette même phrase inférée par le même modèle. Les paramètres de transformation sont alors déduits de la comparaison de ces deux dernières réalisations et appliqués par les algorithmes de transformation du signal sur la phrase neutre source à transformer. Avant d'être appliqués, les paramètres de transformations sont d'abord lissés, grâce à un filtrage médian. D'autres possibilités ont été envisagées mais non implémentées pour cette étape. La prise en compte des variances peut permettre d'obtenir un gabarit dynamique pour chacun des paramètres dans lequel une trajectoire peut être définie par l'utilisation de splines, par exemple. Si cette méthode semble être probante pour la synthèse à partir du texte [Yamagishi 2005, Bailly 2005], elle reste à être mise au point pour la transformation. En effet, comme les paramètres de transformation sont déduits de la comparaison des valeurs neutres et expressives inférées, il reste à définir une méthode de comparaison des variances de ces valeurs possédant des conséquences sur les paramètres de transformation (voir partie 5).

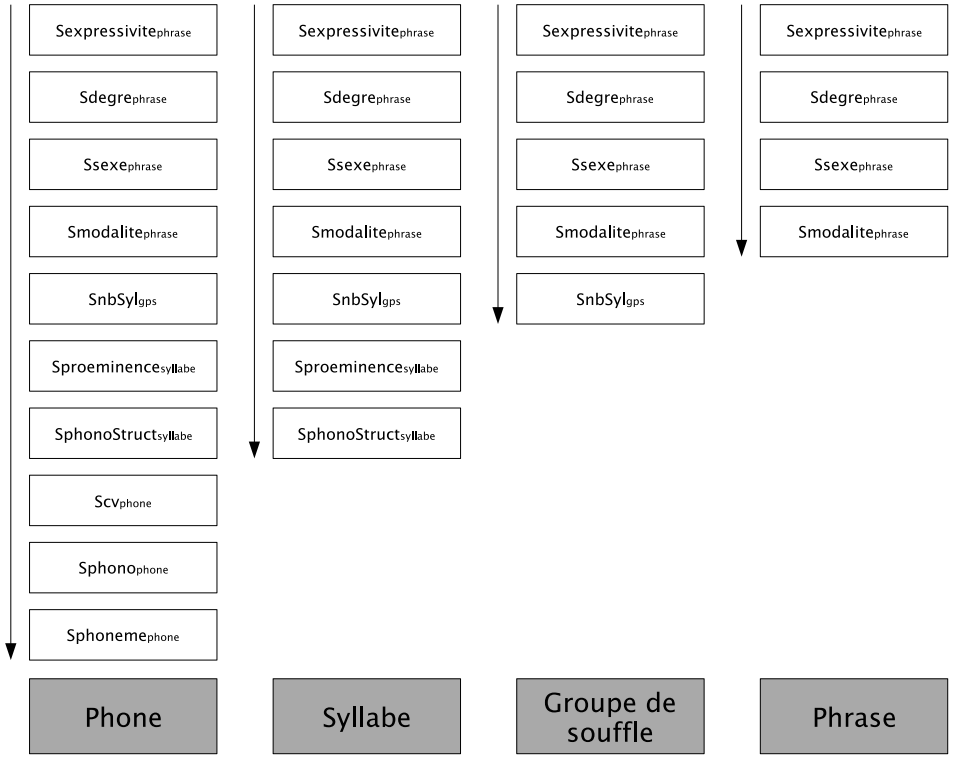

FIG. 4.6: Hiérarchie des variables contextuelles pour les unités de type : phone, syllabe, groupe de souffle et phrase.

4.6 Transformations du signal de parole

Un segment de parole, comme une syllable ou une phrase, peut être manipulé dans l'espace prosodique (à cinq dimensions), grâce à des algorithmes de traitement du signal. La transposition modifie la hauteur de la parole. La dilatation/compression temporelle permet d'en changer le débit. Un gain variable permet de changer l'intensité. Le degré d'articulation peut être modifié grâce à la dilatation/compression non linéaire de l'enveloppe spectrale [Beller 2007b, Beller 2008a]. Enfin, la qualité vocale commence à devenir modifiable par manipulation d'un modèle LF et de bruits modulés de manière synchrone à la fréquence fondamentale [Lu 2002].

4.6.1 Traitement du signal

Toutes ces modifications du signal de parole peuvent être réalisées par plusieurs algorithmes de traitement du signal. Ceux-ci impliquent pour la plupart un modèle de signal de parole. Certains modèles sont plutôt tournés vers les manipulations temporelles, tandis que d'autres sont destinés aux manipulations spectrales. En réalité, la plupart des méthodes existantes à ce jour permettent les deux types de manipulations.

4.6.1.1 Méthodes temporelles

Cela étant, les méthodes temporelles de modification du signal de parole reposent sur la synchronisation et la modélisation du pulse glottique. La méthode PSOLA et ses nombreux dérivés [Peeters 2001], ainsi que les méthodes AR-X dans lesquelles le pulse glottique est souvent modélisé par un modèle LF [Vincent 2007], permettent de transformer une et une seule période de signal. Elle sont donc très précises temporellement mais dépendent fortement de la synchronisation et donc de l'estimation de la fréquence fondamentale. Un algorithme itératif est souvent utilisé pour améliorer la synchronisation, au fur et à mesure de l'estimation des paramètres du modèle. L'ajout de certaines contraintes de continuité entre les paramètres correspondant à deux périodes consécutives assurent souvent une stabilité au modèle.

4.6.1.2 Méthodes spectrales

Les méthodes spectrales de modélisation et de modification du signal de parole s'appuient sur des représentations fréquentielles du type "transformée de Fourier". Les méthodes à vocodeur de phase comme STRAIGHT [Kawahara 2005b, Kawahara 2005a] et SuperVP [Bogaards 2004] permettent de modifier certaines parties du spectre. Les méthodes "sinusoïdes+bruit" [Stylianou 1996, Campedel-Oudot 1998] séparent les parties harmoniques des parties bruitées du spectre et permettent ainsi l'application de traitements différents et appropriés.

4.6.1.3 Modélisation source-filtre

Depuis Fant [Fant 1960], la parole est souvent représentée par un modèle source-filtre. Un des avantages de ce modèle est la réunion des deux représentations précédentes en un modèle spectro-temporel. D'un côté, les méthodes temporelles semblent mieux appropriées pour caractériser la source glottique puisqu'elles permettent, par exemple, de modifier la forme de la dérivée de l'onde glottique période par période, mais aussi la durée d'une période, d'une période à l'autre, tandis que les méthodes fréquentielles utilisent le plus souvent, trois périodes au minimum pour pouvoir estimer le spectre précisément. De l'autre côté, les méthodes spectrales sont mieux adaptées à la modélisation du filtre, qui traduit l'effet du conduit vocal sur la source. Elles permettent notamment une modification plus aisée de paramètres variant relativement lentement, tels que les variations de résonance associées aux mouvements des articulateurs (formants). Aussi, la plupart des méthodes de modification du signal de parole prennent en compte l'hypothèse source-filtre et consistent en des hybridations des deux types de méthode citées précédemment (méthodes LP-PSOLA, AR-LF, HMM-MSA...).

4.6.1.4 Méthode employée

Toutes ces méthodes se distinguent aussi par leur rapidité (temps réel ou temps différé), par leur qualité (dégradation perceptible des caractéristiques de la voix), et par leur complexité. Notre choix est guidé par la qualité de la transformation, par la disponibilité de l'algorithme, et par la possibilité d'agir sur tous ses paramètres. Après avoir utilisé STRAIGHT [Kawahara 2005a], nous nous somme tourné vers une solution locale développée dans le laboratoire d'accueil de la thèse, le vocodeur de phase SuperVP.

4.6.2 Transposition, compression/dilatation temporelle et gain

Le vocodeur de phase utilisé permet la transposition, la compression/dilatation temporelle et l'ajout de gain, avec une excellente qualité puisqu'il est développé pour des applications artistiques. Les paramètres de ces transformations dynamiques (et, aujourd'hui, usuelles) peuvent évoluer dans le temps. Ces trois opérations sont effectuées en même temps sur un même signal de parole, évitant une multiplicité de transformations qui dégraderait la qualité. Enfin, SuperVP traite les transitoires, les parties du spectre voisées et celles non voisées de manière différente, ce qui évite l'effet "vocodeur" caractéristique de cette approche [Roebel 2003].

4.6.3 Dilatation/compression non linéaire de l'enveloppe spectrale

Lorsque l'on effectue une transposition forte (de plus d'une octave), il est nécessaire de modifier conjointement l'enveloppe spectrale, sous peine d'un effet indésirable provenant de l'inadéquation de la hauteur et des caractéristiques du conduit vocal. Si cet effet peut être désiré (voix de "Mickey"...), il n'en va pas de

même pour les transformations expressives, dont l'un des buts collatéraux est la préservation de l'identité du locuteur. Aussi, Un facteur de dilatation/compression linéaire de l'enveloppe spectrale est souvent associé au facteur de transposition (ou déduit). Cette dilatation/compression linéaire de l'enveloppe spectrale permet ainsi d'adapter la taille du conduit vocal à la hauteur résultante de la transposition [Roebel 2005a]. Nous avons enrichi cet algorithme de manière à pouvoir dilater/compresser l'enveloppe spectrale de manière non-linéaire et donc, localement. Ainsi, non seulement la taille du conduit vocal peut être modifiée (changement de locuteur) mais aussi ses configurations (fréquences des formants). Utilisé dans une version statique, ce nouvel algorithme permet ainsi de conférer à un signal de parole un changement de position des articulateurs qui peut simuler le sourire, la nasalisation ou encore une labialisation. Utilisé dans une version dynamique, il permet de modifier le degré d'articulation.

4.6.3.1 Utilisations classiques

Les applications de la dilatation/compression du spectre, ou " Dynamic Frequency Warping" en anglais, sont nombreuses en traitement du signal. La plupart sont éloignées du traitement de la parole comme la transformation bilinéaire, le changement de résolution fréquentielle pour les DSP (Digital Signal Processor : Traitement du Signal Numérique) [Haermae 2000, Haermae 2001], le codage audio [Wabnik 2005] ou la compression multi-bande dynamique [Kates 2005]. Cette technique est aussi employée pour créer des effets sonores [Evangelista 2001a, Evangelista 2000, Evangelista 2001b]. Enfin, en parole, les applications de cette technique permettent la conversion de voix [Umesh 1996, Toda 2001] [Shuang 2006], l'adaptation et la normalisation du locuteur pour la reconnaissance de la parole [Potamianos 1997, Gouvea 1997, Zhan 1997, Gouvea 1999, Lee 1996] et aussi l'interpolation spectrale afin de diminuer les effets de la concaténation [Kain 2007] en synthèse concaténative. Elle n'a pas encore été utilisée pour modifier les configurations du conduit vocal, ainsi que le degré d'articulation.

4.6.3.2 Principe de la dilatation/compression non linéaire de l'enveloppe spectrale

La dilatation/compression non-linéaire de l'enveloppe spectrale est paramétrée par une fonction linéaire par morceaux qui relie deux axes de fréquences F_S et F_T, possédant les mêmes bornes. Cette fonction, appelée FWF pour "Frequency Warping Function" peut évoluer dans le temps : $FWF(f,t)$. A un instant donné t, elle donne la relation non linéaire (ou linéaire par morceaux) entre deux échelles de fréquence : $F_S = FWF(F_T,t)$. Le rééchantillonnage de l'enveloppe spectrale à partir de cette nouvelle échelle fréquentielle permet de compresser/dilater des zones précises du plan temps-fréquence et donc, de modifier le lieu des fréquences de certaines résonances (formants).

4.6.3.3 Algorithme proposé

La première étape est une segmentation du signal de parole $s(t)$ en N trames de 30 ms se chevauchant de 10 ms (ou de manière synchrone au "pitch"). Après fenêtrage, on estime l'enveloppe spectrale en amplitude $S(f)$ et en phase $\phi(f)$ pour chaque trame. Pour l'estimation de l'enveloppe spectrale, la méthode TrueEnvelop [Villavicencio 2006] plus performante que la méthode LPC, et qui, de plus, provient d'un algorithme qui fournit conjointement l'ordre optimal [Roebel 2005a] pour l'estimation, est utilisée. Les concaténations temporelles des enveloppes en amplitude et en phase donnent respectivement $S(f,t)$ et $\phi(f,t)$. $S(f,t)$ et $\phi(f,t)$ constituent l'*envelopogramme* du signal qui permet notamment de visualiser les résonances modelées par le conduit vocal.

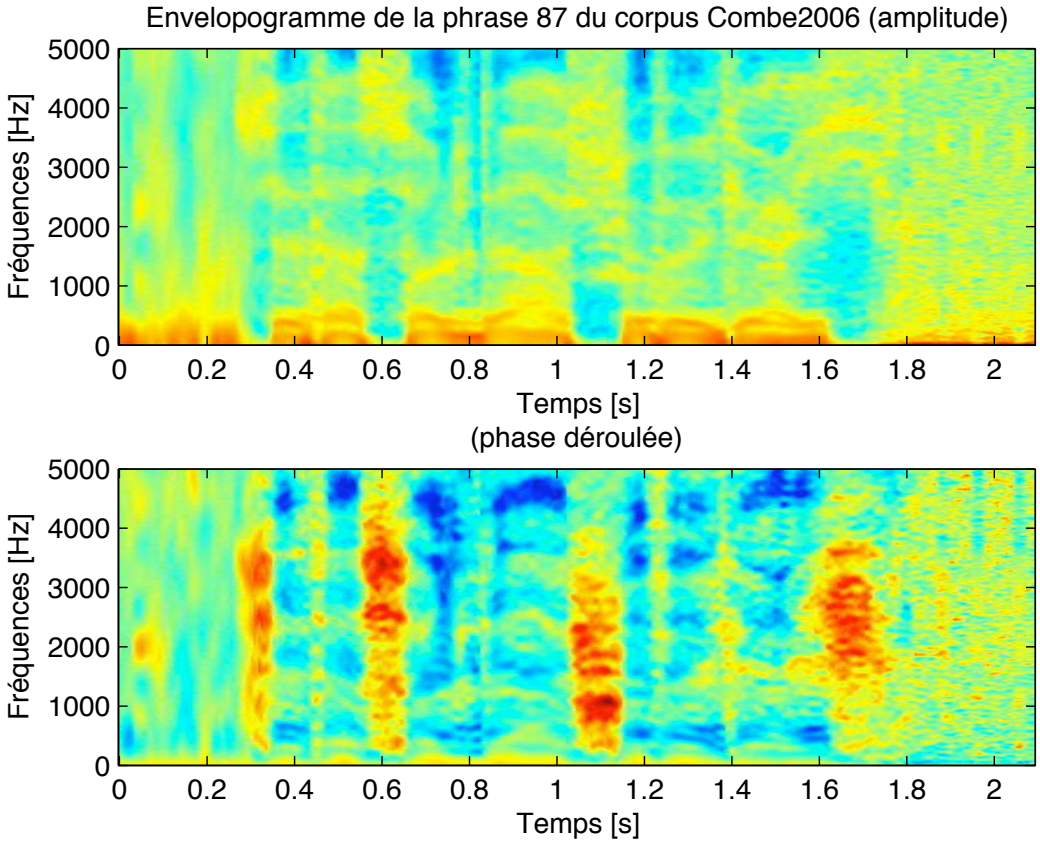

FIG. 4.7: Exemple d'envelopogramme (amplitude en haut et phase en bas). Phrase correspondante à l'exemple donné dans la partie 3.3.

Soit F_S l'axe de fréquence linéaire de la source, signal à transformer. Soit F_T l'axe de fréquence non linéaire de la cible, définie ou non. F_S et F_T sont reliées par FWF :

$$F_T = FWF(F_S, t)$$

Où FWF (Frequency Warping Function) est une fonction de la fréquence variant

dans le temps et dont le résultat est une fréquence. Cette fonction peut être définie de manière arbitraire, comme c'est le cas pour la figure 4.8, ou bien, elle peut être définie par deux ensembles de fréquences correspondants, d'une part, aux lieux des formants de la source et, de l'autre, à ceux de la cible. Ce dernier cas est utilisé pour la conversion de voix vers un locuteur spécifique [Shuang 2006], ainsi que, dans notre cas, pour la modification du degré d'articulation.

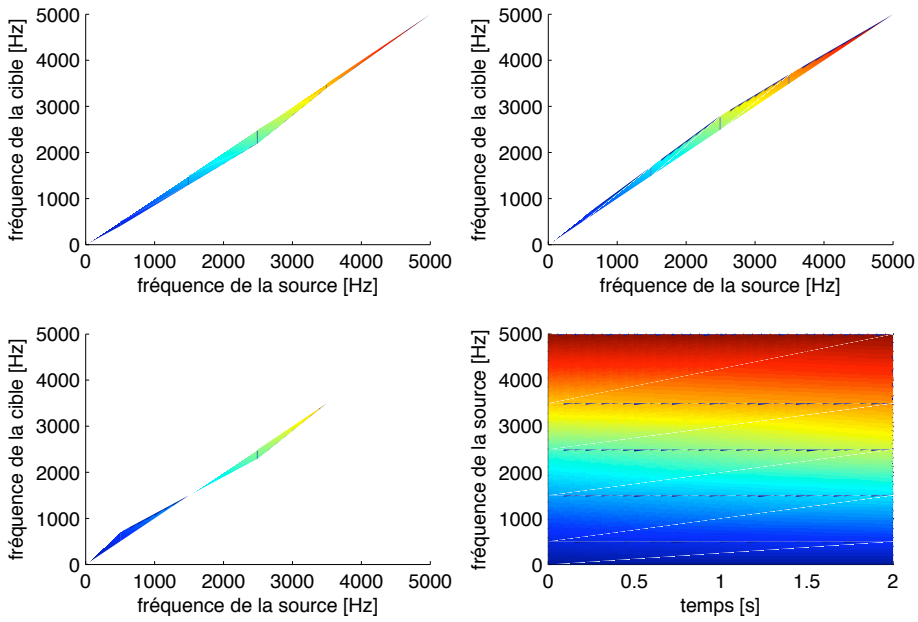

FIG. 4.8: Exemples de fonctions de transfert FWF. En haut à gauche : vers une voix masculine. En haut à droite : vers une voix féminine. En bas à gauche : vers une nasalisation. En bas à droite : évolution dynamique d'une voix masculine à une voix féminine.

Une fois la fonction de transfert déterminée, et l'enveloppe spectrale de la source estimée, on échantillonne non linéairement l'enveloppe spectrale de l'amplitude de la source :

$$\left\{ \begin{array}{lcl} \mathrm{S}_T(F_S,t) & = & \mathrm{S}_S(F_T,t) \\ \mathrm{S}_T(F_S,t) & = & \mathrm{S}_S(FWF(F_S,t),t) \end{array} \right.$$

Ainsi que son enveloppe spectrale de phase déroulée :

$$\left\{ \begin{array}{lcl} \phi_T(F_S,t) & = & \phi_S(F_T,t) \\ \phi_T(F_S,t) & = & \phi_S(FWF(F_S,t),t) \end{array} \right.$$

Puis, on (r)enroule la phase et on reconstruit l'enveloppe cible par concaténation. La différence entre les deux envelopogrammes, correspondant à l'inverse de celui de la source et à celui de la cible permet la définition d'un filtre variant dans le temps. Ce filtre est appliqué, par un vocodeur de phase [Bogaards 2004], à la source de manière à lui conférer les propriétés spectrales de la cible. La figure 4.9 présente une telle opération. Dans cet exemple, la courbe de transformation a été définie par le décalage fréquentiel des formants de la phrase à transformer (source). La fréquence du 1^{er} formant a été abaissée d'un facteur 0.8, la fréquence du 2^{nd} formant a été élevée d'un facteur 1.1 et la fréquence du $3^{ème}$ formant a été élevée d'un facteur 1.2. Cela simule l'apparition d'un zéro dans le spectre, situé entre le 1^{er} et le 2^{nd} formant, ce qui induit la perception d'une plus forte nasalité dans la phrase ainsi transformée.

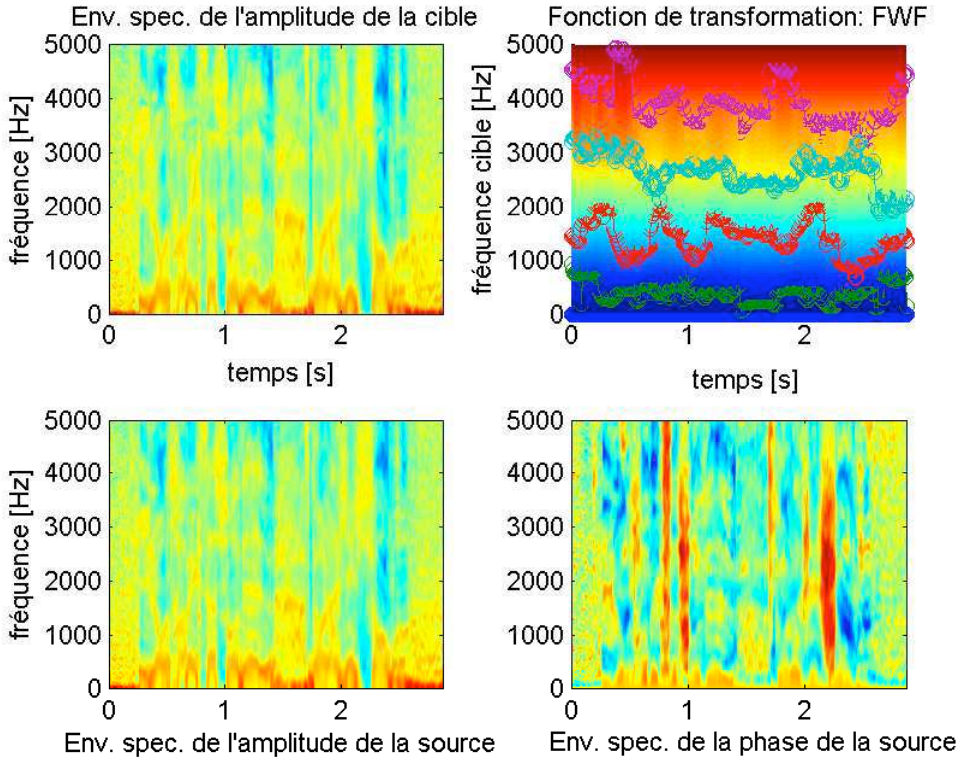

Fig. 4.9: Exemples de transformation du degré d'articulation. En haut à gauche : l'enveloppe spectrale de l'amplitude de la phrase résultante (cible). En haut à droite : la fonction de transformation construite, ici, par déplacement des formants. En bas à gauche : l'enveloppe spectrale de l'amplitude de la phrase à transformer (source). En bas à droite : l'enveloppe spectrale de la phase de la phrase à transformer (source).

4.6.4 Modification de la qualité vocale

Le système Expresso ne modifie que la qualité vocale de la phonation, ni les modes de vibration, ni le degré de voisement. Le procédé utilisé pour modifier la qualité vocale d'une phrase est semblable à celui utilisé, précédemment, pour modifier le degré d'articulation. En effet, la qualité vocale est ici transformée par un filtrage variant dans le temps. A un instant donné, soit Rd_S le coefficient de relaxation estimé sur la source et Rd_T le coefficient de relaxation désiré, ou correspondant à celui d'une cible. Chaque coefficient Rd correspond à un ensemble de de paramètres de modèle LF. Il existe en effet une fonction non bijective entre l'espace de ce coefficient et l'espace des paramètres d'un modèle LF. Or un filtre peut être associé à chaque modèle LF qui représente une forme d'onde glottique élémentaire [Fant 1995, D'Alessandro 1997, Doval 2006, Doval 2000, Henrich 2002, Doval 2003]. La figure 4.10 montre l'effet que produit une variation du coefficient Rd sur la dérivée du spectre de glotte modélisé par un modèle LF. On observe notamment qu'en voix tendue (Rd $\rightarrow$ 0.3), la pente spectrale est plus faible qu'en voix relâchée (Rd $\rightarrow$ 2.7), ce qui traduit une voix plus brillante. On y voit aussi apparaître le formant glottique dont la fréquence diminue au fur et à mesure que la voix se relâche.

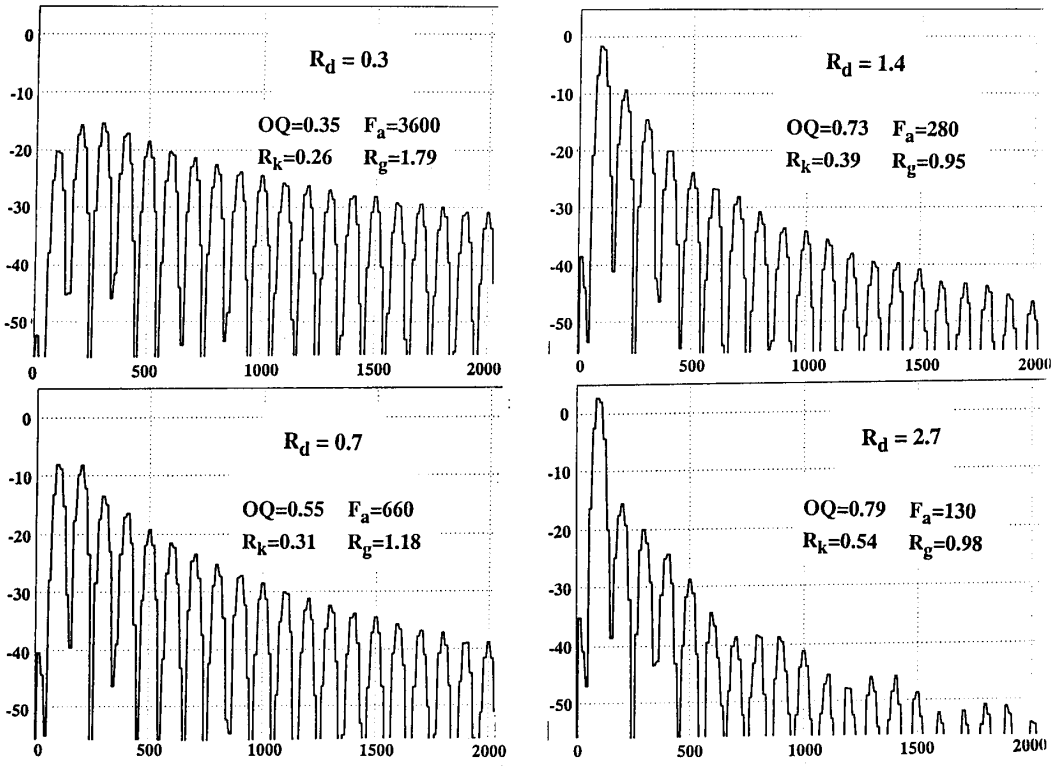

FIG. 4.10: Exemples des effets du coefficient de relaxation Rd sur le spectre de la dérivée du débit glottique modélisé par un modèle LF (f0 = 100 Hz). Image tirée de [Fant 1995].

La technique employée pour modifier la qualité vocale utilise cette représentation spectrale du coefficient de relaxation Rd. En effet, s'il est possible d'associer à chaque valeur de Rd, un filtre, alors la transformation d'une valeur de ce coefficient de relaxation à une autre, peut se faire par filtrage. Le filtre employé résulte de la multiplication de deux filtres : le filtre inverse correspondant au Rd de la source Rd_S et le filtre correspondant au Rd désiré de la cible Rd_T. L'application de ce nouveau filtre, issue de la comparaison des deux, à la source, par un vocodeur de phase [Bogaards 2004], permet de lui conférer, une nouvelle qualité vocale par transformation.

4.7 Evaluation

Dans le but d'évaluer Expresso, le système de transformation de l'expressivité mis au point durant cette thèse, un test perceptif de reconnaissance de l'expression a été mis au point. L'expressivité a été définie, dans le chapitre 2.4.5, comme un niveau d'information disponible dans le message vocal. Ainsi, seule une mesure de la façon dont ce message est perçu peut permettre d'évaluer un système qui se veut capable de modifier ce niveau d'information. C'est pourquoi il a été préféré une mesure subjective, sous la forme d'un test perceptif, à une mesure objective comprenant une évaluation numérique des résultats (par une distance de similarité ou autre). D'ailleurs, la définition d'une métrique objective constituerait en soit une connaissance intéressante qu'il s'agirait plutôt d'introduire dans la constitution du modèle.

4.7.1 Tests perceptifs directs et indirects

Le chapitre 2.4.5 a présenté des méthodes d'acquisition de données émotionnelles dans lesquelles on pouvait distinguer les méthodes directes et les méthodes indirectes. Il existe classiquement cette distinction pour les tests perceptifs [D'Alessandro 2004]. En effet, la mesure de la perception d'un stimuli expressif par un sujet peut se faire avec ou sans connaissance de sa part de l'objet de mesure. Dans un cas, le cas des méthodes indirectes, une tâche est proposée au sujet qui ne se rend pas compte que l'on mesure chez lui sa perception de l'expressivité. Par exemple, on lui demande de retranscrire la phrase qu'il entend ; Il pense alors qu'il s'agit d'un test d'intelligibilité, mais en fait on mesure le temps de réaction entre l'audition de la phrase et la retranscription que l'on met, a posteriori, en relation avec l'expression. Dans l'autre cas des méthodes directes, le sujet sait que l'on mesure sa perception de l'expressivité, qui devient le but explicite de l'expérience. Si le choix du type de méthode (directe ou indirecte) s'est avéré cruciale pour la production des stimuli expressif, il n'en va pas de même pour la mesure de leur perception. En effet, il ne s'agit pas de mesurer l'induction d'une émotion chez le sujet, c'est à dire son état interne provoqué par un stimuli. Mais il s'agit plutôt de mesurer l'agrément d'une population dans la qualification de l'expression de stimuli. C'est pour cela que nous avons choisi une méthode de mesure perceptive directe.

Il existe plusieurs types de tests perceptifs directes. Des stimuli sont présentés à un auditeur qui doit les qualifier : - par verbalisation ouverte, - par catégorisation guidée, - par regroupement (paires, groupes, ...). Nous avons choisi d'utiliser les mêmes termes que ceux donnés aux acteurs pour la production des stimuli. De ce fait, notre choix s'est porté sur une test de catégorisation guidée.

De nombreux facteurs, correspondant pour la plupart aux niveaux d'information attenant à la communication (voir chapitre 2.4.5), peuvent influencer la perception de l'expressivité : L'âge, le sexe, l'origine, la catégorie socio-professionnelle... Si l'étude de l'influence de chacun de ces facteurs sur la perception de l'expressivité peut être passionnante, elle peut s'avérer très complexe tant ils sont nombreux. De

acteur	modèle	nombre de phrases
Combe	corpus (pas de transformation)	15
Roullier	corpus (pas de transformation)	15
Combe	ModelCombe	15
Combe	ModelRoullier	15
Roullier	ModelCombe	15
Roullier	ModelRoullier	15

TAB. 4.4: Simuli utilisés pour le test de reconnaissance de l'expressivité.

manière à marginaliser l'influence du contexte de perception (par analogie avec le contexte d'apparition (voir chapitre 3.2.4), il est alors nécessaire de disposer de la population la plus large possible. Pour cela, une interface web a été choisie pour réaliser ce test.

4.7.2 Test perceptif mis au point

Le test perceptif, élaboré pour évaluer le système Expresso, est un test direct de reconnaissance de l'expressivité, à partir de stimuli audio actés (ceux du corpus) ou transformés (provenant du système expresso). Il est disponible sur le web[2]. Le but de cette expérience est de comparer les scores de reconnaissance de l'expressivité des stimuli actés et ceux des stimuli transformés. Deux modèles, appris sur deux acteurs masculins, ont été évalués : *ModelCombe* et *ModelRoullier*. Une phrase sémantiquement neutre vis à vis de l'expressivité a été tirée du corpus pour l'expérience : "Il ne pourra pas me voir si j'éteins la lampe.". Les 14 versions expressives de cette phrase, exprimées par chacun des acteurs font partie du test, en tant que stimuli de contrôle. A cet ensemble ont été ajoutés les produits de transformations expressives. Les phrases neutres de ces deux acteurs ont été transformées dans toutes les expressions et en utilisant les deux modèles. Les 90 stimuli faisant partie du test sont résumés par le tableau 4.4 suivant :

Au total, le test présente de manière aléatoire 90 stimuli, dont le tiers sont des phrases de contrôle qui n'ont pas été transformées. Le test de reconnaissance repose sur la classification de ces stimuli dans les catégories expressives exprimées par les acteurs. L'interface web du test perceptif est présenté par la figure 4.11.

Chaque stimulus a été étiqueté en moyenne 75.122 fois, ce qui produit des résultats statistiquement significatifs. Les variables pouvant influencer la perception comme l'âge, le sexe ou la catégorie socio-professionnelle du participant n'ont pas été observées. Si ces facteurs peuvent influencer les résultats localement, leur influence disparaît de manière statistique tant le nombre d'individus par stimulus est grand. C'est d'ailleurs pour cela que nous effectué ce test via une interface web simplifiée, légère et préservant l'anonymat. La récolte des résultats s'est effectuée sur deux mois. Seuls les tests présentant plus de dix annotations ont été gardés. Pour

[2]Test perceptif : http://perspection.ircam.fr/beller/expresso_quiz/

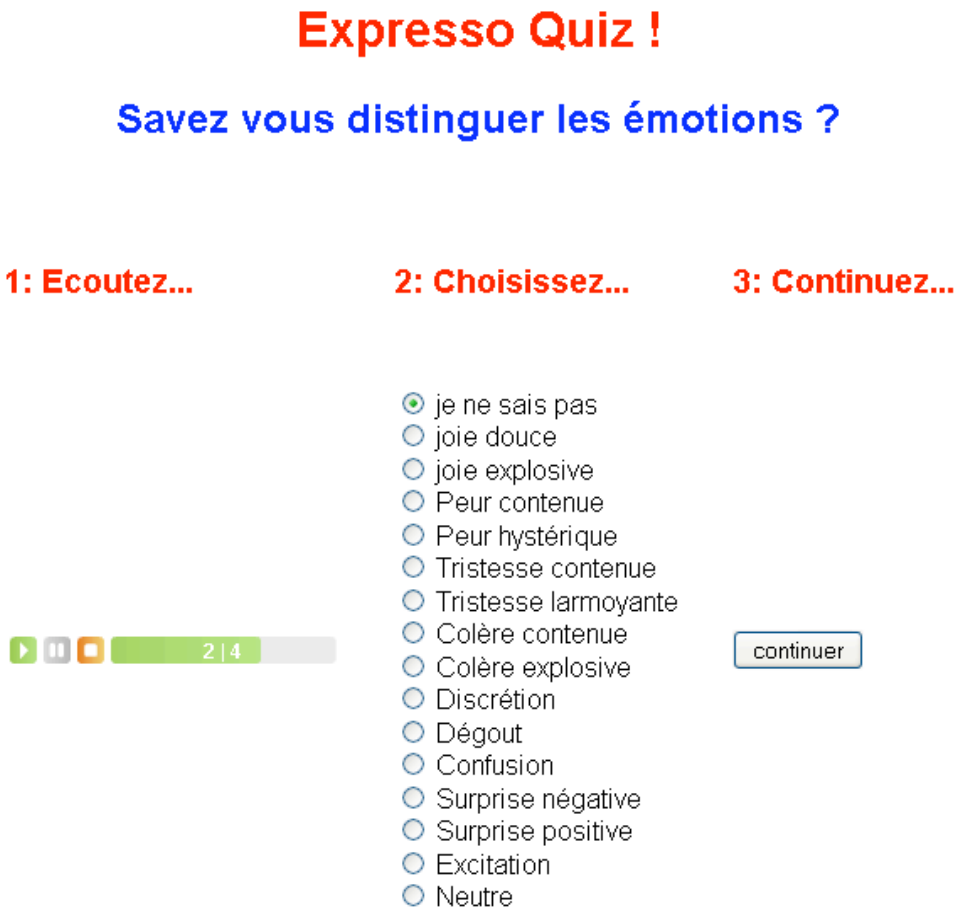

FIG. 4.11: Aperçu de l'interface web du test perceptif.

chacun des ces tests, les sept premières annotations ont été évincées, considérant qu'elles constituaient une phase de découverte et d'entraînement pour les sujets. La catégorie "Je ne sais pas" a été écartée des résultats puisqu'elle est une étiquette "joker" permettant de ne pas se prononcer.

4.7.3 Résultats du test

D'une manière globale, les résultats de reconnaissance sont assez faibles, bien qu'au-dessus du niveau du hasard. Ce fait a été corroboré par certains commentaires de participants qui ont jugé le test intéressant mais difficile, notamment à cause de la profusion des classes (trop de choix possibles). Le tableau 4.5 présente les scores de reconnaissance pour les stimuli actés et pour les stimuli transformés :

Le tableau 4.5 montre que les taux de reconnaissance sont assez faibles, même pour les stimuli actés. Cependant, ils restent tous supérieurs au hasard exceptés ceux des stimuli transformés en surprise positive et en tristesse introvertie. La tris-

expressions	score acteurs	score transfos	hasard
joie intro	61.538	14.218	6.667
joie extra	63.158	40	6.667
peur intro	37.662	8.1395	6.667
peur extra	31.507	10.465	6.667
tristesse intro	57.143	4.5977	6.667
tristesse extra	76.471	55.367	6.667
colère intro	54.229	14.4	6.667
colère extra	40.47	33.333	6.667
discrétion	36.194	12.745	6.667
dégoût	25.806	8.2707	6.667
confusion	84.507	29.565	6.667
surprise négative	22.764	15.556	6.667
surprise positive	30.345	6.0241	6.667
excitation	12.121	43.038	6.667
neutre	65.025	14.716	6.667
Moyenne	46.596	20.696	6.667

Tab. 4.5: Taux de reconnaissance des expressions.

tesse extravertie et la confusion semble être les expressions les mieux reconnues, aussi bien dans le cas des stimuli actés que dans le cas des stimuli transformés. les scores de reconnaissance des phrases actées sont en moyenne deux fois supérieurs à ceux des phrases transformées. Ce résultat n'est pas étonnant puisque les transformations sont issues de modèles appris sur les phrases actées. Toutefois, et de manière surprenante, l'excitation semble mieux reconnue dans le cas transformé, que dans le cas acté. Les taux de reconnaissance sont assez différents d'une expression à l'autre, c'est pourquoi il est nécessaire de regarder plus en détail les sources de confusion. La figure 4.12 présente les matrices de confusion pour les stimuli actés de contrôle et pour les stimuli transformés.

Le figure 4.12 montrent les taux de reconnaissance en pourcentage des expressions pour les stimuli actés et transformés. Le tableau se lit de la manière suivante : Pour une expression désirée (en ordonnée), on lit les pourcentages de votes issus du test, de gauche à droite, relatifs à l'expression perçue. La somme des pourcentages d'une ligne est égale à 100%. Le caractère diagonal de la matrice de confusion des stimuli actés révèlent une assez bonne reconnaissance. Seule l'excitation n'a pas été bien reconnue et a été majoritairement annotée comme joie extravertie (mais de manière non significative). En revanche, le caractère désordonné de la matrice de confusion des stimuli transformés révèlent certaines confusion. Ainsi les émotions primaires, dans leurs versions introverties, ont été moins bien reconnues que dans leurs versions extraverties. La peur extravertie a été plutôt perçue comme de l'excitation, label souvent utilisé aussi pour annoter les stimuli transformés en colère extravertie. La surprise positive a été perçue en tant que surprise négative.

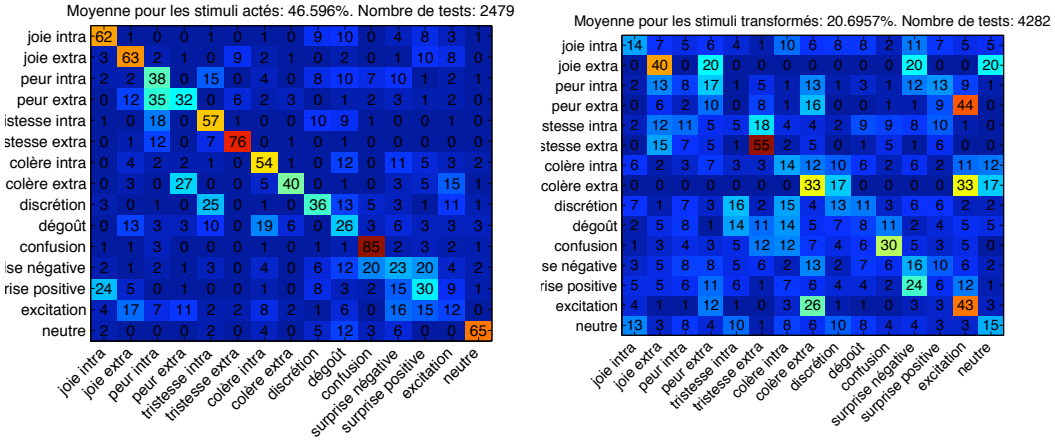

FIG. 4.12: Matrices de confusion pour les stimuli actés (à gauche) et pour les stimuli transformés (à droite). En abscisse, les catégories perçues. En ordonnées, les catégories réelles auxquelles appartiennent les stimuli. La somme des scores de reconnaissance (pourcentages) le long d'une ligne est de 100%.

4.7.4 Interprétations par regroupements

Le nombre important de catégorie a été critiqué par certains participants. Dans la suite, certains regroupements sont effectués pour y palier, a posteriori. Bien sûr, il ne s'agit que d'une interprétation des données.

4.7.4.1 Regroupement de l'introversion et de l'extraversion

Le premier regroupement présenté regroupe les versions introverties et extraverties des émotions primaires, ainsi que les surprises positives et négatives. Le tableau 4.6 présente les taux de reconnaissance de ces catégories regroupées.

expressions	score acteurs	score transfos	hasard
joie	64.074	21.296	10
peur	51.667	23.401	10
tristesse	76.512	32.516	10
colère	48.801	26.718	10
discrétion	36.194	12.745	10
dégoût	25.806	8.2707	10
confusion	84.507	29.565	10
surprise	44.403	27.523	10
excitation	12.121	43.038	10
neutre	65.025	14.716	10
Moyenne	50.911	23.979	10

TAB. 4.6: Taux de reconnaissance des expressions regroupées.

Ce regroupement améliore les taux de reconnaissance des expressions. Seul les stimuli transformés en dégoût présente un taux en dessous du hasard. En moyenne, les stimuli actés sont reconnues deux fois mieux que les stimuli transformés dont le taux de reconnaissance moyen est plus de deux fois celui du hasard (10%). Encore une fois, la matrice de confusion permet d'apprécier le détail de cette évaluation.

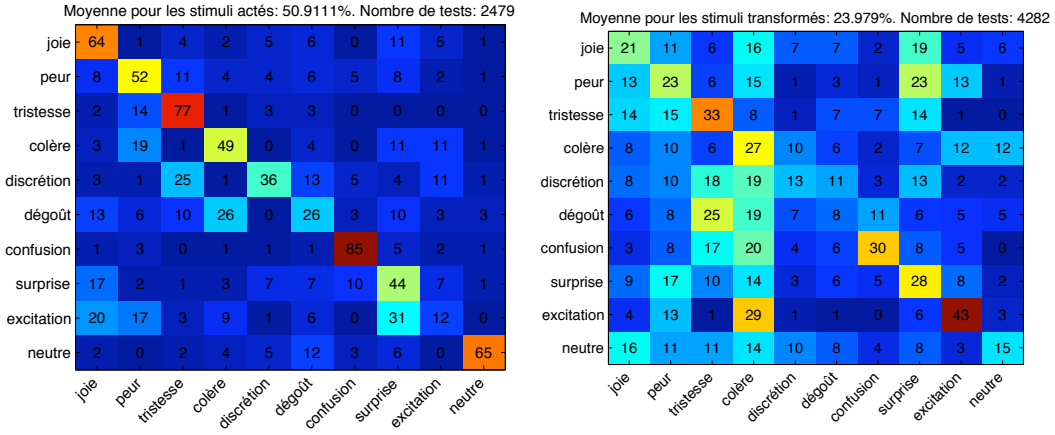

FIG. 4.13: Matrices de confusion pour les stimuli actés (à gauche) et pour les stimuli transformés (à droite). En abscisse, les catégories perçues regroupées. En ordonnées, les catégories réelles regroupées auxquelles appartiennent les stimuli. La somme des scores de reconnaissance (pourcentages) le long d'une ligne est de 100%.

La figure 4.13 présente les matrices de confusion des stimuli actés et des stimuli transformés. En ce qui concerne les stimuli actés, on remarque que l'excitation a été majoritairement perçue en tant que surprise. Pour les stimuli transformés, la peur a été associée à la surprise, la discrétion à la colère, le dégoût à la tristesse, et le neutre à la joie. La colère est une expression qui a été perçue dans presque tous les stimuli (sauf la tristesse) et elle a notamment été souvent employée pour qualifier l'excitation.

4.7.4.2 Comparaison des modèles

Un autre regroupement peut nous permettre d'apprécier lequel des deux modèles utilisés, *ModelCombe* et *ModelRoullier*, a donné de meilleurs résultats. Bien sûr, il ne s'agit pas ici de comparer les performances des deux acteurs mais bien des modèles issus de leurs performances. La figure 4.14 présente les matrices de confusion des stimuli transformés par ces deux modèles.

La figure 4.14 montre que les résultats sont assez différents selon le modèle utilisé pour la transformation. En moyenne, le modèle *Combe2006* semble donner de meilleurs résultats que le modèle *Roullier2006*. Mais en réalité, les résultats sont assez différents selon les expressions. Le modèle *Combe2006* fournit des résultats très disparates. En effet, si la joie extravertie, la tristesse extravertie et l'excitation semblent des transformations performantes, les autres ne possèdent pas vraiment la

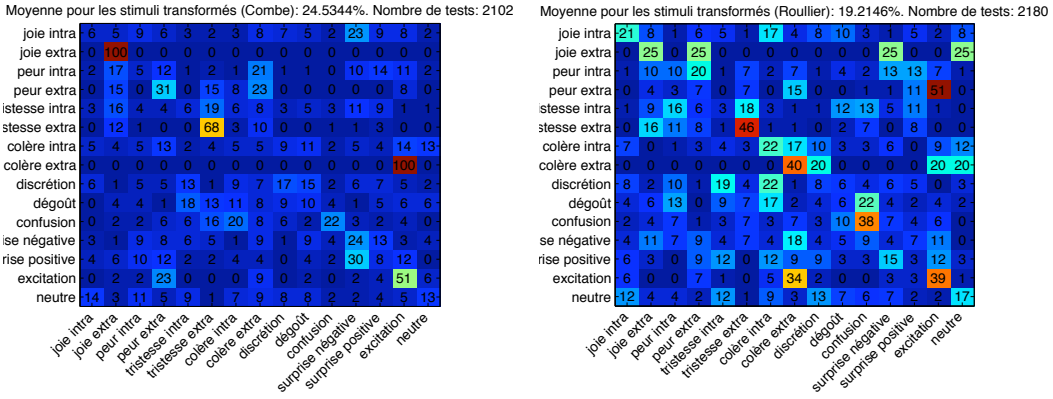

FIG. 4.14: Matrices de confusion pour les stimuli transformés par le modèle *Combe2006* (à gauche) et par le modèle *Roullier2006* (à droite). En abscisse, les catégories perçues. En ordonnées, les catégories désirées appliquées en entrée du système Expresso. La somme des scores de reconnaissance (pourcentages) le long d'une ligne est de 100%.

capacité d'être perçue comme l'expression désirée. Notamment la colère extravertie qui a été perçue à chaque fois comme de l'excitation. En ce qui concerne le modèle *Roullier2006*, Les résultats sont moins tranchés. Ce modèle est performant pour la tristesse extravertie, la colère extravertie et la confusion. La première conclusion de cette comparaison est qu'il semble manifestement que les modèles possèdent des performances dépendants de l'expression. Il semble alors qu'un modèle hybride, réunissant des parties des deux modèles selon leurs taux de reconnaissance par expression pourrait fournir de meilleurs résultats.

La disparité des performances des modèles selon les expressions amène à la comparaison des résultats des deux acteurs vis à vis des stimuli actés et non transformés. Bien entendu, il ne s'agit pas d'une comparaison subjective et définitive de leurs jeux, mais plutôt des taux de reconnaissance des quelques phrases choisies au hasard dans la base de données (1 phrase par expression et par acteur).

La figure 4.15 présentent les matrices de confusion des stimuli actés par les acteurs Combe et Roullier. Le caractère diagonal de ces deux matrices révèlent de bons taux de reconnaissance. Pour l'acteur Combe, la peur extravertie a été perçue comme introvertie, le dégoût a été perçue comme joie extravertie, la discrétion et l'excitation n'ont pas été bien reconnues. Pour l'acteur Roullier, la surprise négative a été perçue comme confusion, la surprise positive a aussi été perçue comme joie introvertie, l'excitation a été perçue comme surprise négative et la peur introvertie n'a pas été bien perçue. Bien sur, ces conclusions sont à relativiser puisqu'elle ne concerne qu'une phrase choisie aléatoirement dans la base de données. Toutefois, ces données de contrôle révèlent certaines lacunes, au sein de ce test, qui engendrent la critique.

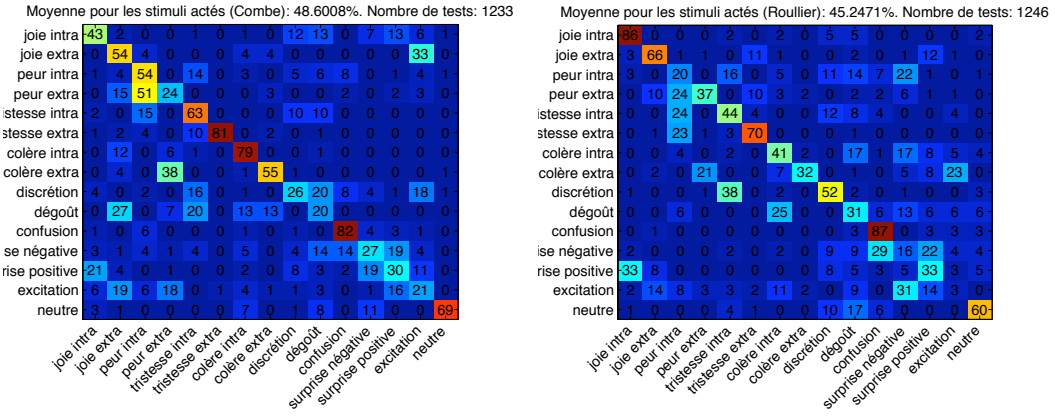

FIG. 4.15: Matrices de confusion pour les stimuli actés par les acteurs Combe (à gauche) et Roullier (à droite). En abscisse, les catégories perçues. En ordonnées, les catégories actées. La somme des scores de reconnaissance (pourcentages) le long d'une ligne est de 100%.

4.7.5 Validité de l'évaluation

Si cette évaluation préliminaire nous permet une mesure de l'expressivité perçue, qui peut ressembler à un gage d'objectivité, elle possède certains problèmes. Le premier est le nombre trop grand de classes expressives présentées qui a pu dérouter un participant non habitué à la tâche. Une solution consisterait à diminuer le nombre de classes. Une autre résiderait dans la mise en contexte des stimuli. Un second problème réside dans l'incomplétude de cette évaluation puisqu'il ne s'agit que de l'évaluation de certains stimuli, 90 au total, soit une infime partie de ce que les acteurs et le système Expresso sont capables de produire. Pour les phrases actées, il n'existe qu'un stimuli par acteur et par expression. Pour les phrases transformées, il n'existe qu'un stimuli par trio {phrase neutre - modèle utilisé - expression}. Ce faible nombre de stimuli a pour conséquences de possiblement engendrer des biais dans l'évaluation puisqu'on ne mesure que les résultats d'une occurence. Un exemple de tels biais provient de la différence entre le taux de reconnaissance des stimuli actés neutres et celui des stimuli transformés neutres, bien que ces stimuli possèdent tous la même expression et n'ont pas été transformés. En effet, la transformation expressive du cas neutre vers le cas neutre est nulle puisque le modèle utilise la comparaison des deux expressions pour produire des paramètres de transformation. C'est à dire que les stimuli transformés neutres correspondent en réalité aux stimuli actés neutres qui ont servi de base pour les autres transformations. Par conséquent, les taux de reconnaissance des stimuli actés neutres devraient être les mêmes que ceux des stimuli transformés neutres. Cela montre qu'il serait nécessaire de poursuivre cette évaluation avec un nombre plus important de stimuli pour tirer des conclusions objectives sur le jeu des acteurs ou sur les performances du système.

4.8 Discussions

L'approche bayésienne soulève de nombreuses questions concernant la nature des variables (discrètes ou continues), ainsi que sur l'indépendance des variables entre elles. Si le choix de la nature et de l'indépendance des variables permet une simplification du calcul de la $GPDF$, quelques cas prêtent tout de même à discussion.

4.8.1 Interdépendances des variables symboliques

Les phonéticiens qui ont corrigé la segmentation phonétique de la base de données, ont observé que pour certaines expressions, des phonèmes attendus étaient absents, ajoutés ou différents de ceux fournis par la segmentation automatique (un "/E/" ouvert peut sonner comme un "/œ/", par exemple). Même si le même texte a été prononcé, les fréquences d'apparition des phones pour chaque expression sont différentes (voir partie 3.4). Aussi, il parait nécessaire d'ajouter une dépendance entre les variables $Sphono^{phone}$ et $Sexpressivite^{phrase}$. Estimer $P(Sphono^{phone}|Sexpressivite^{phrase})$ en ajoutant une flèche dans le graphe : $Sexpressivite^{phrase} \rightarrow Sphono^{phone}$.

Une interdépendance contextuelle similaire est observable pour le niveau de proéminence : $Sexpressivite^{phrase} \rightarrow Sproeminence^{syllabe}$. Par exemple dans le cas de la colère extravertie, presque chaque syllabe est perçue comme proéminente, car elles sont séparées par des césures. Toutefois, comme le montre les analyses effectuées dans la partie 3.4, ces dépendances semblent faibles et les variables $Sexpressivite^{phrase}$, $Sphono^{phone}$ et $Sproeminence^{syllabe}$ sont considérées comme indépendantes.

4.8.2 Interdépendances des variables acoustiques

Un second type d'interdépendance important existe entre les variables acoustiques. Il a été montré, par exemple, que la moyenne de la fréquence fondamentale est fortement corrélée à la moyenne de l'intensité perçue (voir partie 3.5. Ainsi des relations comme $Af0^{syllabe}_{model_{moyenne}} \rightarrow Aint^{syllabe}_{model_{moyenne}}$ peuvent être ajoutées au modèle. Le réseau bayésien modélise alors aussi les corrélations possibles entre les dimensions prosodiques. Les distributions LCG des paramètres acoustiques sont alors linéairement dépendantes selon une relation donnée ou estimée W_i :

$$P(Af0^{syllabe}_{model_{moyenne}}|S = C_i, Aint^{syllabe}_{model_{moyenne}} = aint) = \mathcal{N}(\mu_i + W_i \times aint, \sigma_i) \quad (4.16)$$

Malheureusement, ces dépendances n'ont pas pu être ajoutées au modèle pour des raisons computationnelles.

4.8.3 Dépendance entre deux contextes successifs

Les chaînes de Markov cachées sont largement répandues en parole et partagent le même formalisme que les réseaux bayésiens (modèles graphiques). Il a été montré

que la connaissance des probabilités de transition entre les unités augmente les performances des systèmes de reconnaissance automatique [Gauvain 1996], ainsi que la qualité des modèles prosodiques [Bailly 2005]. Si bien qu'une augmentation de la qualité de la prédiction semble possible par la connexion de deux contextes successifs : $S(i-1) \rightarrow S(i)$. Les phénomènes de coarticulation peuvent ainsi être pris en compte par un réseau bayésien dynamique. Cependant, cette liaison n'est pas sans contrepartie puisqu'elle multiplie la cardinalité de l'Univers U par lui même, puisque chaque contexte devient dépendant du contexte précédent. Or la taille de nos corpus d'apprentissage ne nous permet pas de couvrir un tel Univers. C'est pourquoi, cette dépendance n'a pas été envisagée.

4.8.4 Variable expressivité : discrète ou continue ?

La nature de certaines variables peut être discutée : par exemple, la nature de l'expressivité de la phrase $Sexpressivite^{phrase}$ peut être discutée car il existe plusieurs représentations de l'expressivité dont certaines sont catégorielles (discrète) et d'autres sont dimensionnelles (continues) (voir partie 2.2.5).

Il en va de même pour l'intensité de l'expression. Beaucoup de représentations considèrent l'intensité expressive comme un paramètre variant continuement. Cependant, la variable $Sdegree^{phrase}$ a été considérée, jusqu'ici, comme une variable discrète car elle correspond, dans les corpus enregistrés, à l'index de la phrase dans sa répétition (répétition de la même phrase avec une intensité croissante, voir chapitre 2.6.5). Changer la nature des variables $Sdegree^{phrase}$ et $Sexpressivite^{phrase}$ conduit à des distributions gaussiennes linéairement conditionnelles (LCG) de paramètres acoustiques, telles que dans l'équation 4.16.

4.8.5 Synthèse de sons paralinguistiques

Pour la synthèse de parole expressive, il est possible d'insérer des sons paralinguistiques tels que des respirations et des fillers (rires, pleurs...). Une tentative a été menée dans la synthèse semi-paramétrique de rires propres à un locuteur donné, ne possédant qu'une seule phrase de ce locuteur (voir l'annexe D). Grâce à l'examen d'une base de données de rires, nous avons établi un modèle statistique. Le système est alors capable de synthétiser un rire à partir d'un seule phrase donnée. La voyelle la plus serrée (du point de vue de la qualité vocale), ainsi qu'une consonne sont concaténées, dupliquées puis transformées selon des patrons produits par le modèle appris. Le résultat est un rire (ou plusieurs, car plusieurs solutions sont proposées), qui conserve l'identité du locuteur. Si les résultats sont variables selon la phrase présentée, il sont encourageants tant la perception de l'expressivité de la phrase peut s'en trouver modifiée. A titre d'exemple, la tristesse peut être perçue seulement par un pleur et la peur, seulement par un cri, sans aucun autre mot verbal. Plus finement, une expérience perceptive informelle montre que le simple ajout local d'une respiration au milieu d'une phrase neutre peut changer l'expression perçue de toute la phrase (en l'occurence, l'expression perçue était la peur). Le pouvoir

expressif des mots non verbaux est tel, qu'il devient nécessaire aux synthétiseurs de parole de les générer.

4.9 Conclusion

Ce chapitre a présenté le système Expresso de transformation de l'expressivité. Une première partie a mis en évidence les différents plusieurs paradigmes de génération des paramètres de transformation. Ils peuvent être définis de manière heuristique, ou bien de manière adaptative au signal à transformer, appelé source. Dans ce cas, différentes "adaptativités" sont décrites selon qu'elles prennent en compte des informations de type symbolique, des informations de type acoustique ou bien la réunion de ces deux types d'information. Le système proposé implique un modèle qui réagit en fonction des données symboliques et des données acoustiques caractérisant la phrase à transformer. Cela permet notamment de changer l'expression d'une phrase sans en altérer les autres niveaux d'information (identité du locuteur...), par l'observation des différences entre cas neutre et cas expressifs d'une base de données.

Puis le modèle génératif utilisé pour la transformation de l'expressivité d'une nouvelle phrase a été présenté. Un modèle génératif statistique est appris sur une base de données de parole expressive multi-locuteur. Les paramètres des transformations acoustiques varient dans le temps et sont dépendants des contextes symboliques extraits du texte et d'une définition de l'état du locuteur. Il a été montré comment un réseau bayésien réalise le passage entre un modèle à base de règles et un modèle guidé par les données. L'un des points cruciaux réside dans la capacité de généralisation du modèle à des contextes non observés. Pour cela nous avons réalisé un algorithme d'inférence qui réduit le contexte jusqu'à trouver un nombre suffisant d'exemples pour estimer les paramètres de transformation.

Une fois les paramètres de transformation générés, ils sont appliqués par des algorithmes de traitement du signal. Les opérations de transposition, de dilatation/compression temporelle, de gain, de dilatation/compression de l'enveloppe spectrale et de changement de la qualité vocale permettent de modifier les cinq dimensions de la prosodie que sont respectivement l'intonation, le débit de parole, l'intensité, le degré d'articulation et la qualité vocale.

Enfin, ce chapitre a présenté la mise en place d'un test perceptif pour l'évaluation du système Expresso. Conformément à notre définition de l'expressivité, ce test perceptif de mesure directe repose sur la catégorisation de l'expressivité de stimuli actés et transformés. Une interface web a été implémentée pour permettre une diffusion large du test et ainsi récolter une population importante de participants. Les résultats du test sont encourageants bien que les taux de reconnaissance soient assez faibles. Malheureusement, cette évaluation reste partielle, car elle présente trop peu de stimuli, et problématique, car elle présente trop de classes expressives. Toutefois, elle permet de préparer le terrain d'évaluations futures, et de pressentir certains comportements du système. En moyenne, les performances d'Expresso sont de moitié celles des acteurs. De plus, les matrices de confusion ont permis de mettre en évidence qu'un modèle hybride, basés sur différents acteurs selon l'expression, pourraient donner de meilleurs résultats.

Chapitre 5

Conclusion générale

Cette thèse s'inscrit dans les recherches actuelles sur les émotions et les réactions émotionnelles, sur la modélisation et la transformation de la parole, ainsi que sur l'interprétation musicale. Pour transformer ces réactions, il est tout d'abord nécessaire de définir la notion d'expressivité, d'un point de vue théorique. Puis, d'un point de vue pratique, il faut modifier le signal de parole, de manière à ce que la perception de l'expressivité en soit changée. Cette apparente simplicité cache de nombreux verrous théoriques et techniques (notamment liés à la nature idiosyncrasique des émotions, et à la variabilité, à la richesse et à la complexité du phénomène de la parole), que cette thèse a tenté de mettre en évidence, tout en essayant d'y apporter quelques éléments de réponse. D'un point de vue pratique, un programme expérimental a été crée, permettant de conférer à n'importe quelle phrase, en français, enregistrée ou synthétisée, une expression désirée avec un certain degré d'intensité. Des exemples sonores sont disponibles à l'adresse suivante : http://recherche.ircam.fr/equipes/analyse-synthese/beller.

Un état de l'art des théories sur le contrôle des émotions et une revue des techniques expérimentales d'acquisition des données émotionnelles ont montré la nécessité de séparer l'émotion de sa réaction émotionnelle usuellement associée. En guise de réponse, une définition originale de l'expressivité a été donnée. Cette définition a été mise à l'épreuve dans le contexte de la performance artistique. L'étude conjointe des interprétations verbales et musicales a permis la constitution de six hypothèses pour l'observation expérimentale et scientifique de l'expressivité de l'interprétation qui sont présentées dans la partie suivante. Une méthode de représentation hybride catégorico-dimensionnelle de l'expressivité a été proposée. Elle a été employée, avec les hypothèses précédentes, pour la réalisation de trois corpus expressifs, dont l'un est constitué d'interprétations musicales (violon). Les deux corpus de parole expressive ont été exploités par un système original de gestion de corpus (voir l'annexe A). Puis, un nouveau modèle de la parole a été appliqué aux corpus dans le but de fournir des analyses symboliques et acoustiques de l'influence de l'expressivité sur l'interprétation verbale. Ces résultats, associés à des algorithmes de traitement du signal, dont un permettant la modification du degré d'articulation, ont permis l'établissement d'un modèle bayésien génératif pour la transformation de l'expressivité de la parole. Ces différentes contributions sont explicitées ci-dessous.

D'un point de vue théorique, cette thèse a proposé une définition de l'expressivité, une définition de l'expression neutre, un nouveau mode de représentation de l'expressivité, ainsi qu'un ensemble de catégories expressives communes à la parole et à la musique, dans la partie perspective. Elle a situé l'expressivité parmi le recensement des niveaux d'information disponibles dans l'interprétation qui peut être vu comme un modèle de la performance artistique. Elle a proposé un modèle original de la parole et de ses constituants, ainsi qu'un nouveau modèle prosodique hiérarchique.

D'un point de vue expérimental, cette thèse a fourni un protocole pour l'acquisition de données expressives interprétées. Colatéralement, elle a rendu disponible trois corpus pour l'observation de l'expressivité. Elle a produit une nouvelle mesure statistique du degré d'articulation ainsi que plusieurs résultats d'analyses concer-

nant l'influence de l'expressivité sur la parole.

D'un point de vue technique, elle a proposé un algorithme de traitement du signal permettant la modification du degré d'articulation. Elle a présenté un système de gestion de corpus novateur qui est, d'ores et déjà, utilisé par d'autres applications du traitement automatique de la parole, nécessitant la manipulation de corpus (comme la synthèse à partir du texte, par exemple). Elle a montré l'établissement d'un réseau bayésien en tant que modèle génératif de paramètres de transformation dépendants du contexte.

D'un point de vue technologique, un système expérimental de transformation, de haute qualité, de l'expressivité d'une phrase neutre, en français, synthétique ou enregistrée, a été produit. De même, une interface web a été constitué de manière à évaluer ses performances, sur la base d'un test perceptifs.

Enfin et surtout, d'un point de vue prospectif, cette thèse propose différentes pistes de recherche pour l'avenir, tant sur les plans théorique, expérimental, technique, que technologique. Parmi celles-ci, la confrontation des manifestations de l'expressivité dans les interprétations verbales et musicales semble être une voie prometteuse, comme le montre la partie suivante.

La suite de cette thèse propose une partie prospective dédiée à la comparaison des manifestations de l'expressivité dans la parole et dans l'interprétation musicale. S'il existe des catégories expressives communes à la parole et à la musique, est-ce qu'elles se manifestent de la même manière à travers ces deux moyens d'expression ?

Chapitre 6

Perspective : Expressivité de l'interprétation musicale

Sommaire

6.1 Résumé du chapitre

Cette partie, plus prospective, permet d'examiner ce qu'il se passe dans le cas de l'interprétation musicale, et dans le cas de l'interprétation en général. L'idée centrale est que la parole et la musique partage un pouvoir expressif. La première question abordée concernant ce pouvoir expressif est : Est-ce qu'il est commun ? Une revue des catégories émotionnelles que l'on trouve dans la musique permet de penser qu'il existe certaines expressions communes à la parole et à la musique, bien que la majorité d'entre elles sont différentes, compte tenu des différents rôles que tiennent ces deux médiums de communication. La seconde question abordée est : Peut-on observer l'expressivité de l'interprétation musicale ? Une analogie est alors construite entre la parole et la musique, sur la base de l'interprétation d'un support. L'observation des différents niveaux d'information de la parole et de l'interprétation musicale permet de bâtir un ensemble d'hypothèses pour l'observation de l'expressivité dans l'interprétation. Ces hypothèses, qui ont été respectées pour la constitution du corpus de parole expressif IrcamCorpusExpressivity, permettent la constitution d'un corpus d'interprétations musicales expressives, semblable aux corpus de parole. Ce corpus est présenté car il a été enregistré durant cette thèse. Malheureusement son exploitation n'est pas décrite par ce manuscrit, bien que des pistes pour le faire soient données. Parmi ces pistes, un axe de recherche est évoqué, à partir de l'écoute comparée des corpus de parole et d'interprétations musicales : Par analogie, l'expressivité de la parole est en partie révélée par la prosodie, ce qui laisse à penser que l'expressivité de l'interprétation musicale puisse aussi être, en partie, révélée par ce que l'on appelle : une *prosodie instrumentale*. Cette perspective est développée dans cette thèse, pour indiquer nos axes futurs de recherche, et parce qu'elle peut contribuer à l'amélioration de notre modèle génératif. S'il existe des gestes acoustiques de l'expressivité, communs à la parole et à l'interprétation musicale, alors leur confrontation peut nous aider à mieux les cerner, dans chacun de leurs contextes respectifs. Ce qui nous amène à la dernière question : L'expressivité se révèle t'elle dans la parole et dans l'interprétation musicale, par des gestes acoustiques communs ?

6.2 Emotions verbales et musicales

La musique et la parole partagent de nombreuses caractéristiques :

- Elles sont basées sur la production, la transmission et la perception d'un signal acoustique.
- Leurs productions et leurs perceptions partagent de nombreux processus neuro-cognitifs [LeDoux 2005, Patel 2008].
- Elles impliquent au moins deux acteurs (humain ou non) : un *interprète* qui produit et un *auditeur* qui perçoit.
- Elles sont généralement considérées comme des languages en tant qu'elles sont deux moyens de communiquer certaines informations (qui ne sont pas forcément les mêmes) entre ces deux parties.
- Elles peuvent reposer sur des structures, des grammaires, des règles ou plus généralement des codes communs aux deux parties (interprète et auditeur).
- Ces codes rendent la parole et la musique, en partie transcriptibles, sous les formes respectives d'un texte et d'une partition.
- Ces parties transcrites peuvent être interprétées par des individus (acteurs ou instrumentistes).

Bien que ces deux moyens de communication possèdent chacun leurs domaines d'utilisation (parfois commun, comme c'est le cas dans certaines pièces de musique contemporaine[1]), ils partagent le terrain de l'expression des émotions [Meyer 1956, Hevner 1936, Patel 2008].

Les différentes représentations exposées dans la partie 2.2 ont été tirées de données émotionnelles provenant de plusieurs moyens d'expression (parole, expressions faciale et gestuelle, musicale...). De manière générale, il est admis que la parole peut être le véhicule de toutes les émotions. Notamment car elle porte en son sein, le verbe, qui peut être utilisé pour décrire explicitement la catégorie linguistique émotionnelle dans laquelle le locuteur se trouve. Cependant, certains phénomènes affectifs sont plus fréquents en parole, comme les attitudes ou les préférences, tandis que d'autres semblent plus fréquents en musique comme les émotions esthétiques. Est-ce que la musique et la parole occupent le terrain des émotions de manière identique ? Autrement dit, (et avant de se lancer dans la comparaison de l'expressivité de la parole et de celle de l'interprétation musicale), est-ce que les catégories émotionnelles exprimées par la parole et par la musique, sont les mêmes ? Est-ce que les dimensions des représentations dimensionnelles des émotions musicales et verbales, sont semblables ?

Nous allons tenter d'apporter des éléments de réponse à ces questions par un recensement chronologique des dimensions et des termes émotionnels utilisés en musique. Longtemps implicites, une dénomination commune s'est établie dès l'époque classique, afin de spécifier l'émotion d'un passage musical lors de sa composition. De nos jours, et depuis l'avènement de la psycho-physique, différents domaines d'étude scientifique se sont constitués autour des émotions dans la musique. Que ce soit

[1] pièce utilisant le *Sprechgesang*, par exemple

pour la classification ou l'étude de l'induction des émotions dans/par la musique, chacun de ces domaines a vu émerger son champ sémantique des émotions dans la musique.

6.2.1 Caractères dans la musique classique occidentale

Dans la musique classique occidentale, le "caractère" désigne la façon d'interpréter une pièce musicale, indépendamment des indications concernant le rythme et l'intonation. De même que le compositeur indique tempo et nuances en italien, il va, dès le XVIIIe siècle, indiquer, sur la partition, le caractère du morceau par des termes empruntés à la même langue. Il convient cependant de soigneusement distinguer les termes indiquant tempo, intensité et caractère, car le solfège leurs a attribué des fonctions spécifiques, distinctes du sens originel qu'ils ont dans la langue italienne[2].

Les termes indiquant le caractère, ont pour fonction d'aider l'interprète à trouver « sa propre expression musicale ». Cette dernière ne doit pas être confondue avec la sentimentalité, procédé dont maints interprètes du XIXe siècle ont usé, et qui consiste à mettre en avant divers sentiments (sincères ou factices). Pourtant certains de ces termes semblent clairement évoquer des émotions. Le tableau 6.1 réunit ces termes, leurs traductions (de l'auteur) et tente de les relier aux phénomènes affectifs proposés par Scherer (voir partie 1).

La figure 6.1 recense les différents phénomènes affectifs associés aux termes italiens. Elle montre que les émotions utilitaires sont peu utilisées dans cette nomenclature musicale. Un recensement statistique similaire, sur les caractères réellement utilisés par les compositeurs, permettraient certainement de conforter cette idée (voir chapitre 5).

Le recensement de ces caractères nous permet de mettre en évidence que les émotions musicales ne se réduisent pas aux émotions utilitaires. Il semble en effet qu'elles soient plutôt de l'ordre des émotions esthétiques.

6.2.2 Emotions musicales contemporaines

En 1936, Hevner donne une représentation géométrique des émotions musicales [Hevner 1936]. Huit ensembles sont construits par Hevner. Ces ensembles consistent en des vecteurs lexicaux qui permettent, sans nommer l'ensemble, de produire une description complexe de l'émotion. Deux émotions utilitaires seulement, "sad" (tristesse) et "happy" (joyeux) apparaissent dans cette représentation. De plus, Hevner

[2]On remarque que l'utilisation des termes italiens pour indiquer tempo, intensité et caractère, a pour principal inconvénient de provoquer des confusions dans les esprits. Par exemple, le terme piano - qui en italien, signifie doucement - fait référence à l'intensité en notation musicale, et non pas au tempo, comme on pourrait le supposer. De même, le terme allegro - qui en italien, signifie allègre, gai - équivaut à "rapide" en notation musicale. Il y a donc dans ce dernier cas, une ambiguïté entre le tempo et le caractère. Il est vrai également qu'un caractère donné, appelle à un certain type de tempo et pas à un autre - par exemple, "tranquillo" sous-entend un tempo plutôt lent, alors que "furioso" sous-entend un tempo plus rapide, etc...

Terme italien	Signification	Phénomène affectif
affettuoso	affectueux	position relationnelle
agitato	agité	attitude / mouvement
amabile	aimable	position relationnelle
amoroso	amoureux	position relationnelle
appassionato	passionné	émotion esthétique
ardito	hardi	attitude
brillante	brillant	attitude
cantabile	chantant	émotion esthétique
capriccioso	capricieux	disposition affective
comodo	commode, aisé	attitude / humeur
con allegrezza	avec allégresse	mouvement
con anima	avec âme	émotion esthétique
con bravura	avec bravoure	émotion esthétique
con brio	avec brio (entrain, vivacité)	attitude / mouvement
con delicatezza	avec délicatesse	mouvement
con dolore	avec douleur	émotion utilitaire
con espressione	avec expression	émotion esthétique
con fuoco	avec flamme	émotion esthétique
con grazia	avec grâce	émotion esthétique
con gusto	avec goût	émotion esthétique
con moto	avec mouvement	mouvement
con spirito	avec esprit	émotion esthétique
con tenerezza	avec tendresse	position relationnelle
delicato	délicat	mouvement
disperato	désespéré	humeur
dolce	doux	attitude / humeur
doloroso	douloureux	attitude / humeur
drammatico	dramatique	émotion esthétique
energico	énergique	attitude / mouvement
espressivo	expressif	position relationnelle
furioso	furieux	émotion utilitaire
giocoso	joyeux	émotion utilitaire
grazioso	gracieux	émotion esthétique
lagrimoso	éploré	humeur
leggero	léger	mouvement
maestoso	majestueux	émotion esthétique
malinconico	mélancolique	attitude /humeur
mesto	triste	émotion utilitaire
nobile	noble	position relationnelle
patetico	pathétique	humeur
pomposo	pompeux	position relationnelle
religioso	religieux	émotion esthétique
risoluto	résolu	émotion esthétique
rustico	rustique	émotion esthétique
scherzando	en badinant	mouvement
semplice	simple	émotion esthétique
teneramente	tendrement	position relationnelle
tranquillo	tranquille	humeur
tristamente	tristement	émotion utilitaire

Tab. 6.1: Liste des termes musicaux italiens les plus utilisés, traduits (par l'auteur) et reliés aux phénomènes affectifs de Scherer [Scherer 2005].

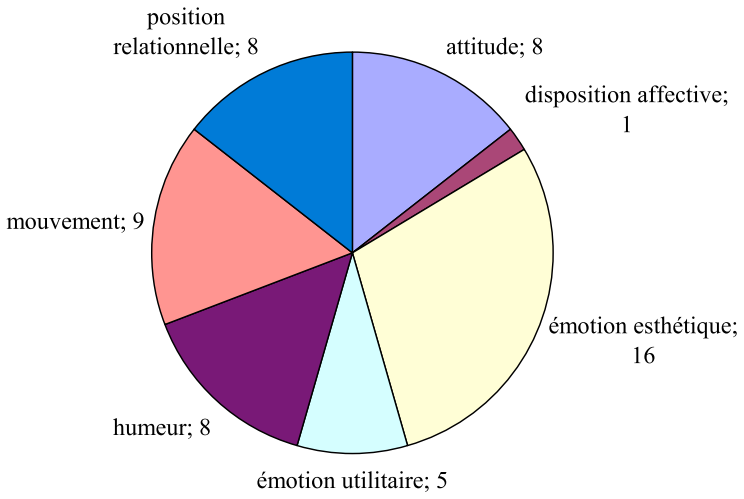

FIG. 6.1: Caractères italiens du tableau 6.1, réunis par phénomènes affectifs.

n'utilise pas les caractères italiens.

En effet, depuis le début du XXe siècle, l'utilisation des termes italiens devient de plus en plus rare. Certains compositeurs leurs substituent directement des catégories émotionnelles, des attitudes ou des comportements psychologiques. Dans la sequenza III pour voix de femme soliste (1965), Luciano Berio indique sur la partition les phénomènes affectifs suivants (traduits par l'auteur) : "distant and dreamy" (distant et rêveur), "nervous" (nerveux), wistful (mélancolique), ecstatic (extasié), faintly (faible), apprehensive (inquiet), tender (tendre), languorous (langoureux), noble (noble), joyful (joyeux), subsiding (soumis), frantic (frénétique), whining (geignant), gasping (haletant), urgent (urgent), serene (serein), desperate (désespéré), anxious (inquiet), calm (calme), witty (plein d'esprit)... De plus, il donne quelques indications paralinguistiques : laughter (rire), whispering (chuchotement)... Là encore, Luciano Berio n'utilise pas les termes relatifs aux émotions utilitaires, mais plutôt des termes relatifs à des émotions esthétiques ou à des situations psychologiques plus complexes. Cela montre encore une fois que les phénomènes affectifs présents en musique ne se réduisent pas aux émotions utilitaires, plus fréquentes dans des situations verbales.

6.2.3 Emotions musicales d'aujourd'hui

La caractérisation des émotions dans la musique, à partir du signal acoustique [Pohle 2005], permet d'accroître notre connaissance sur ce phénomène et de réaliser de nombreuses applications. La détection et la classification des émotions, participe aujourd'hui à une vaste campagne d'indexation de la musique [Lu 2006, Feng 2003, Li 2003]. Ces tâches sont souvent effectuées de manière automatique grâce à des algorithmes d'apprentissage [Mandel 2006, Yang 2006]. Grâce à des modèles psycho-

FIG. 6.2: Représentations géométriques des émotions musicales d'Hevner [Hevner 1936].

acoustiques de notre perception des émotions dans la musique, ils s'appliquent à quantifier l'émotion de pièces ou de performances musicales. Cette caractérisation permet, entre autres, de donner un retour à de jeunes instrumentistes et participe à leur éducation musicale [Juslin 2006]. Un autre domaine de recherche vise à observer le phénomène d'induction des émotions par la musique [Kreutz 2008, Juslin 2004]. L'observation de ces phénomènes d'induction peut conduire à l'établissement de véritables thérapies des émotions. La musicothérapie tente de soigner des patients (souvent atteints de maladies touchant la gestion des émotions) par la musique. Pour cela, les thérapeutes ont besoin de stimuli pouvant rendre joyeux, triste, effrayé ou encore apaisé [Vieillard 2007].

De nos jours, donc, des groupes d'application se forment autour de plusieurs champs lexicaux des émotions. Certaines distinguent les humeurs (moods) des émotions (emotions) [Pohle 2005]. Par exemple, une campagne MIREX récente a défini 5 ensembles d'humeurs pour classer la musique "pop" [Hu 2006, Hu 2007]. Ce classement est relatif à la catégorisation AMMC[3]. Voici ces ensembles et leurs traductions (de l'auteur) :

- passionate (passionné), rousing (réveillant), confident (confiant), boisterous (violent), rowdy (chahuteur)
- rollicking (gaieté), cheerful (gai), fun (amusant), sweet (doux), amiable/good natured (aimable/bon)

[3]AMMC : Audio Mood and Music Classification : http://www.music-ir.org/mirex/2008/index.php/Audio_Music_Mood_Classification

- literate (littéraire), poignant (intense), wistful (mélancolique), bittersweet (aigre-doux), autumnal (automnal), brooding (rêveur)
- humorous (plein d'humour), silly (idiot), quirky (débrouillard), whimsical (étrange), witty (plein d'esprit), wry (désabusé)
- aggressive (agressif), fiery (ardent), tense/anxious (tendu/inquiet), intense (intense), volatile (volatil), visceral (viscéral)

Ces ensembles d'humeurs, parfois métaphoriques ("volatil"), ressemblent aux vecteurs lexicaux utilisés par Hevner pour décrire des émotions complexes [Hevner 1936]. Parmi ces ensembles d'humeurs, ne figure aucune émotion utilitaire.

Il semble donc, au vue de ce recensement des termes utilisés pour nommer l'émotion dans la musique, que les émotions musicales soient plutôt de l'ordre des émotions esthétiques. Toutefois, des émotions utilitaires apparaissent aussi dans ce vaste champ lexical et semblent pouvoir appartenir au monde musical. D'ailleurs, les représentations dimensionnelles des émotions verbales et des émotions musicales semblent recouvrir des espaces communs. Dans une expérience récente [Vines 2005], des participants évaluaient des enregistrements de clarinette, dont seule variait l'expression de l'interprétation, sur des échelles perceptives reliées à des attributs émotionnels. Une analyse factorielle a révélé deux dimensions émotionnelles importantes et indépendantes : La valence (positive/négative) et l'activation (passif/actif). Ces deux axes, associés à l'axe de l'intensité, constituent l'espace le plus utilisé pour représenter les émotions, quelque en soit leurs moyens d'expression.

6.2.4 Emotions communes à la parole et à la musique

En première conclusion, la complexité des émotions musicales a amené certains à l'utilisation de plusieurs termes, pour décrire une émotion. Ces vecteurs lexicaux rappellent ceux employés par certains annotateurs qui ont étiqueté l'émotion d'un corpus de parole [Devillers 2003a, Devillers 2003b] (voir partie 2.2). Ce parallèle montre que les situations émotionnelles sont souvent complexes et ne peuvent se résumer à un seul mot. Or, l'annotation et l'étude des émotions verbales est souvent réduite aux émotions utilitaires. La complexité de l'annotation des émotions musicales suscite et engage donc, à une complexification de l'annotation des émotions verbales. C'est pourquoi nous avons choisi une représentation des expressions verbales, sous la forme de vecteurs lexicaux dimensionnels. En effet, la partie suivante, qui montre la constitution des corpus, présente les expressions sous la forme d'un vecteur contenant, une catégorie, une intensité et un degré d'activation ("joie extravertie intense", par exemple).

En seconde conclusion, il apparaît que les émotions musicales sont plutôt de l'ordre des émotions esthétiques que des émotions utilitaires. Toutefois, l'utilisation de représentations communes montre que certaines émotions appartiennent à la fois au monde de la parole et au monde de la musique. Les émotions utilitaires font partie de ce noyau commun, qui peut être perçu dans la parole et dans la musique. Nous allons donc tenter de comparer l'expressivité de la parole et celle de la musique, vis à vis de certaines émotions utilitaires. Pour cela, la constitution d'un

corpus d'interprétations musicales, répondant au même protocole expérimental que pour les corpus de parole, est nécessaire. Mais tout d'abord, il est nécessaire de situer l'expressivité de l'interprétation musicale, tout comme nous l'avons fait pour la parole.

6.3 L'interprétation

Lors d'une interprétation musicale, vocale ou gestuelle (danse), un public peut ressentir des états internes qui ont été provoqués par une interprétation particulièrement réussie d'une oeuvre. S'il ne ressent aucune émotion, il peut tout de même reconnaître, comprendre et qualifier certaines marques ou indices, lui permettant d'inférer l'état interne de l'interprète, ou d'inférer un état interne que l'oeuvre prête à l'interprète (état interne d'un personnage, ou d'un passage musical). Ainsi dans le théâtre ou le cinéma, un spectateur peut être ému d'une représentation, et/ou il peut comprendre les états internes des personnages, grâce à l'interprétation des comédiens, et grâce à la mise en situation des personnages et de leurs relations, opérée par le scénario (au sens large) de la pièce. Durant un concert, un spectateur peut être ému par une interprétation et/ou il peut imaginer l'état interne d'un instrumentiste et à travers ce dernier, celui de l'oeuvre, ou même celui du compositeur lui-même. Laissant de côté l'induction d'un état interne chez le public par une interprétation musicale ou vocale, nous allons tenter de mettre en évidence comment l'expressivité de l'interprétation peut permettre au public de comprendre, non seulement l'état interne de l'interprète, mais aussi celui du compositeur, du "parolier", du scénariste ou encore du chorégraphe.

6.3.1 Acteurs de la performance

Une performance musicale ou vocale, non improvisée, implique, au minimum, trois acteurs. Dans le cas de la performance musicale, un compositeur, un instrumentiste et un auditeur. Dans le cas de la performance verbale, un "parolier", un acteur et un public. Dans le cas d'un ballet, un chorégraphe, un danseur et un spectateur. Nous désignerons le *créateur-support*, l'*interprète* et l'*auditeur-spectateur* pour désigner respectivement ces différents acteurs, et ce, afin de réunir la performance musicale et la performance verbale dans un cadre général, celui de la *performance artistique* (voir figure 6.3). Ces acteurs sont mis en relation durant la performance. la notion d'acteur employée ici est abstraite et dépersonnalisée pour trois raisons ; La première est qu'une seule et même personne physique peut tenir le rôle de plusieurs acteurs. C'est le cas d'un créateur qui interprète ses propres créations. C'est aussi le cas, bien sûr, de l'interprète qui est en même temps auditeur de son interprétation. Enfin, c'est le cas du créateur qui devient auditeur d'une interprétation de sa création. La deuxième raison est qu'un acteur peut comprendre plusieurs personnes physiques. Le rôle d'interprète peut être tenu par plusieurs acteurs, par un orchestre ou une compagnie de danse, à l'instar du rôle de l'auditeur tenu par un public qui est le plus souvent composé de plus d'un seul individu. Enfin, la troisième raison est que certains de ces acteurs peuvent être remplacés par des objets non vivants, au sens biologique du terme. Le créateur peut être une création, une composition, une partition, un texte... Et l'interprète peut être un enregistrement d'une interprétation, voire l'ordinateur exécutant un programme créé par le créateur. Dans cette étude, seul l'auditeur doit être une personne vivante.

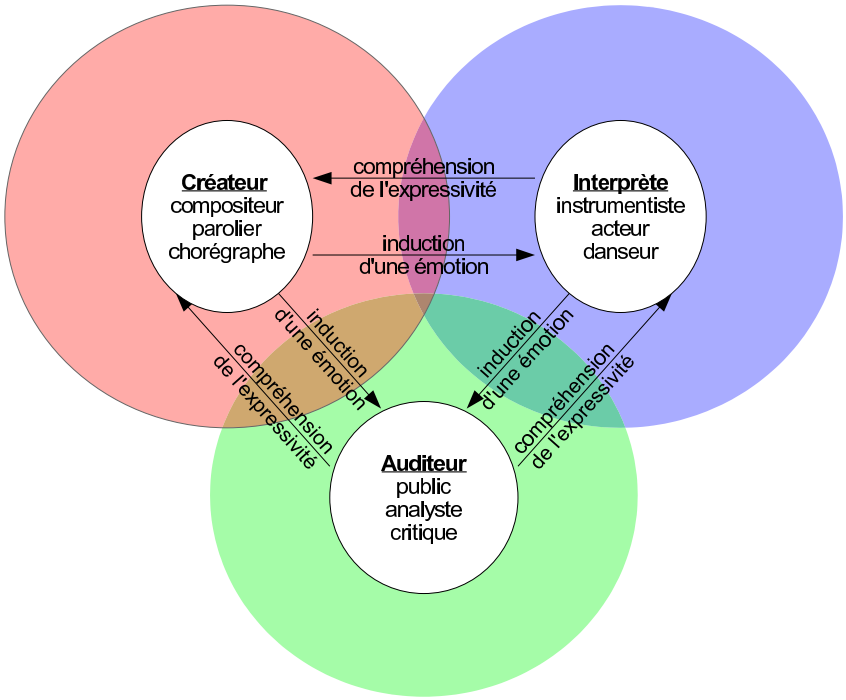

FIG. 6.3: Représentation schématique des trois acteurs impliqués dans une performance, de leurs contextes respectifs, et de leurs relations vis à vis de l'expressivité.

6.3.2 Contextes des acteurs

Chacun des acteurs possèdent aussi de nombreux rapports avec son contexte respectif (représenté par des cercles de couleur sur la figure 6.3). Le recouvrement des cercles signifie que ces contextes peuvent aussi être partagés. Le contexte du créateur est bien sur lié à ses créations et aux multiples relations existantes avec d'autres créations (références, citations, emprunts, adhésion à un courant esthétique...). Il est bien sûr aussi lié à son expérience directe (techniques d'écriture, connaissances spécifiques) et à son expérience indirecte qui dépasse le champ de la création et que l'on appelle ici, l'expérience au sens large. Le contexte de l'interprète est constitué de son environnement extérieur direct (salle de concert, théâtre, auditorium), des caractéristiques de l'événement (concours, concert privé ou public, répétition, enregistrement...), et bien sûr, de son expérience directe (apprentissage, technique) et de son expérience indirecte (expérience au sens large). Ce contexte est partiellement partagé par l'auditeur qui se situe dans le même environnement extérieur et qui peut connaître des caractéristiques relatives à l'événement, tout en possédant un autre point de vue. Enfin, l'expérience de l'auditeur[4] est souvent différente de celle de l'interprète. Le contexte de l'auditeur recouvre une partie du contexte du créateur si l'auditeur possède des connaissances a priori sur la création.

[4]l'expérience de l'auditeur est souvent difficile à déterminer car celui-ci est généralement un expert - de la parole et/ou de l'interprétation musicale - au sens ou il a été exposé largement à ces stimuli, mais il lui est difficile de verbaliser cette connaissance.

6.4 Expressivité de la performance

L'expressivité est partagée et se situe dans l'interaction de ces trois acteurs, lors d'une performance. Des états internes peuvent être induits parmi chacun des acteurs, sous l'action des autres. Il est crucial de différentier l'induction d'un état interne, en soi, de la perception d'un état interne, chez autrui, qui n'est possible que grâce à l'expressivité de celui-ci. De plus, l'expression perçue peut provenir d'une interaction entre acteurs ou peut être multiple et simultanée[5]. Si l'on cherche une représentation chronologique de la communication de l'expressivité à travers l'interprétation, l'on peut partir du travail solitaire du créateur.

6.4.1 Expressivité d'une création

Le créateur crée une oeuvre qui contient une certaine expressivité. Or, cette oeuvre ne possède pas d'état interne propre, car elle n'est pas vivante au sens biologique du terme. Dissocier l'état interne de son éventuelle perception par autrui, c'est à dire de l'expression associée, permet en effet à une oeuvre "morte" de posséder un pouvoir dit "expressif". Cette dissociation s'applique à tout objet d'art, capable de nous faire comprendre un état interne, alors qu'il en est dénué, par sa faculté de nous le faire percevoir par son expressivité (par des signes extérieurs). Cela le rend, de plus, capable de nous émouvoir. En effet, la simple lecture de la création par l'interprète peut plonger celui-ci dans un état interne différent (flèche "induction de l'émotion" allant du créateur à l'interprète, sur la figure 6.3). Sans être ému par la création, l'interprète peut comprendre l'expressivité de l'oeuvre, et commencer, d'ores et déjà, à effectuer des choix quant à son interprétation (flèche "compréhension de l'expressivité" allant de l'interprète au créateur, sur la figure 6.3).

6.4.2 Expressivité perçue par l'auditeur

Lors de la performance, l'auditeur assiste à l'interprétation de la création. A son tour, il peut être ému par la création, grâce à son interprétation. Il peut aussi être ému de l'interprétation en elle-même, notamment s'il connaît déjà la création[6]. Enfin, il peut ressentir un état interne induit par la combinaison de la création et de l'interprétation. Il en va de même pour l'expressivité perçue. En effet, l'auditeur peut percevoir à la fois, l'expressivité de la création et l'expressivité de l'interprétation. Ces relations multiples sont représentées sur la figure 6.3), par les flèches reliées à l'auditeur.

[5]Par exemple, lors d'une représentation théâtrale, le public peut comprendre l'état interne d'un personnage par sa mise en situation dans la pièce, tout en percevant un certain trac chez le comédien. cf. le Paradoxe du comédien

[6]Ce qui peut être le cas du créateur, lui-même, ému de voir sa création se réaliser, par exemple.

6.4.3 Expressivité de la performance

Ces multiples relations de nature expressive, entre ces trois acteurs, rendent complexe la communication de l'expressivité en situation de performance. De plus, l'induction d'états internes, reliés ou non, ainsi que la nature subjective de l'expérience perceptive, compliquent l'étude scientifique des processus de communication de l'expressivité. En effet, l'expressivité d'une création artistique est très difficile à définir, puisque celle-ci n'est pas forcément reliée à un état interne que nous sommes capables d'expérimenter, en tant qu'être vivant, au sens biologique du terme. Cependant, l'expressivité de l'interprétation est quant à elle, plus simplement observable, car reproductible, comparable et mesurable, puisqu'elle se résout à des manifestations externes.

6.4.4 Expressivité de l'interprétation

En effet, l'expressivité de l'interprète devient mesurable, comparable et reproductible, dès lors que l'auditeur est en connaissance de l'état interne que l'interprète exprime. Par exemple, si l'auditeur connaît la création et donc son expression inhérente, il peut se concentrer sur l'expressivité de l'interprète, c'est dire sur la manière dont celui-ci va produire des signes extérieurs liés à ses états internes ou liés directement à l'expression de la création. Si l'on suppose que l'interprète n'est pas dans un état interne particulier (voir partie 1.3.2.2), l'expression que l'on perçoit de son interprétation est alors directement liée à celle de la création. L'interprète exprime alors l'expression inhérente à la création et représentative de ses états internes prêtés, puisque ceux-ci n'existent pas. Enfin, si l'auditeur est en accord sur ces prétendus états internes de la création, alors il peut se focaliser sur la réalisation de ceux-ci par l'interprète, et percevoir l'expression de la création par le biais de celle de l'interprète. C'est sous ces quelques hypothèses que l'expressivité peut alors être objectivement analysée et devenir un objet d'étude de la science, puisqu'elle devient alors reproductible. Puisqu'elles sont le fondement de cette étude et qu'elles vont nous permettre à la fois de constituer un corpus et de réaliser des mesures, nous les résumons dans la partie suivante.

6.4.5 Hypothèses d'étude de l'expressivité de l'interprétation

La première hypothèse permet d'évincer la difficulté résidant dans la définition de l'expressivité d'une création. Elle peut être résolue si le créateur désigne explicitement les états internes présents dans sa création (états internes des personnages d'une pièce, ou états internes de passages musicaux[7]). Elle peut aussi être réalisée par un accord préalable à la performance, entre l'interprète et l'auditeur, sur l'expressivité de la création.

[7]Luciano Bério, dans la Sequenza Tre, appose à la portée des termes désignant des émotions, afin de guider l'expressivité de l'interprétation (voir partie 2.2 pour le détail).

Hyp N°1 : L'auditeur et l'interprète sont en accord sur l'expressivité de la création.

De manière schématique, cette hypothèse permet de rayer les flèches "compréhension de l'expressivité" allant de l'interprète au créateur et allant de l'auditeur au créateur, sur la figure 6.3. Ainsi il ne reste plus que celle allant de l'auditeur à l'interprète (voir figure 6.4.

La seconde hypothèse permet d'annuler les effets de l'induction d'un état interne chez l'interprète par la création, état interne qui peut influencer l'expression perçue de l'interprètation. Cette hypothèse est réalisable en plaçant un interprète qui connaît bien la création, dans un contexte habituel non stressant (répétition ou contexte de laboratoire, par exemple).

Hyp N°2 : L'interprète est dans l'état interne neutre lors de son interprétation.

Par rapport à la figure 6.3, cette hypothèse permet d'annihiler l'influence de la flèche "induction d'une émotion" allant du créateur à l'interprète, ainsi qu'une flèche, non représentée, allant du contexte de l'interprète sur lui-même (trac dû au public, stress d'un concours...). Toutefois et parce que l'état interne neutre est invérifiable, cette hypothèse est plutôt préventive et se traduit dans l'expérience, par une demande explicite à l'interprète, de rechercher cet état interne.

De la même manière, il est important que l'auditeur souhaitant focaliser son attention sur l'expression de l'interprétation, soit aussi dans un état interne neutre, vis à vis de la création, et vis à vis de l'interprétation. Cela est possible si l'auditeur connaît la création préalablement à la performance, et s'il est lui aussi dans des dispositions visant à réduire l'apparition d'état interne dû au contexte[8]

Hyp N°3 : L'auditeur est dans l'état interne neutre lors de l'interprétation.

Cette dernière hypothèse permet enfin d'écarter tout biais dans la perception de l'expressivité de l'interprète, chez l'auditeur. Schématiquement, elle revient à barrer les flèches "induction d'une émotion" dans la figure 6.3 (voir figure 6.4). A l'instar de l'hypothèse N°2 et parce que l'état interne neutre est invérifiable, cette hypothèse est plutôt préventive et se traduit dans l'expérience, par une demande explicite à l'auditeur, de tendre vers un état interne neutre.

Ces trois hypothèses respectées au sein d'un même protocole expérimental permettent la mesure, la reproductibilité et la comparaison des démonstrations externes relatives à l'expressivité. Comme le montre la figure 6.4, le respect de ces hypothèses permet d'éviter l'influence de l'induction d'une émotion sur la perception de l'expressivité, à la fois au niveau de l'interprète et au niveau de l'auditeur, et permet le consensus des deux parties en ce qui concerne la compréhension de l'expression de la création. Il ne reste plus que la compréhension de l'expressivité de l'interprète par l'auditeur, ce qui va être notre objet d'étude par la suite.

[8]Le contexte du laboratoire est-il vraiment une contexte permettant la neutralité de l'état interne des sujets ?

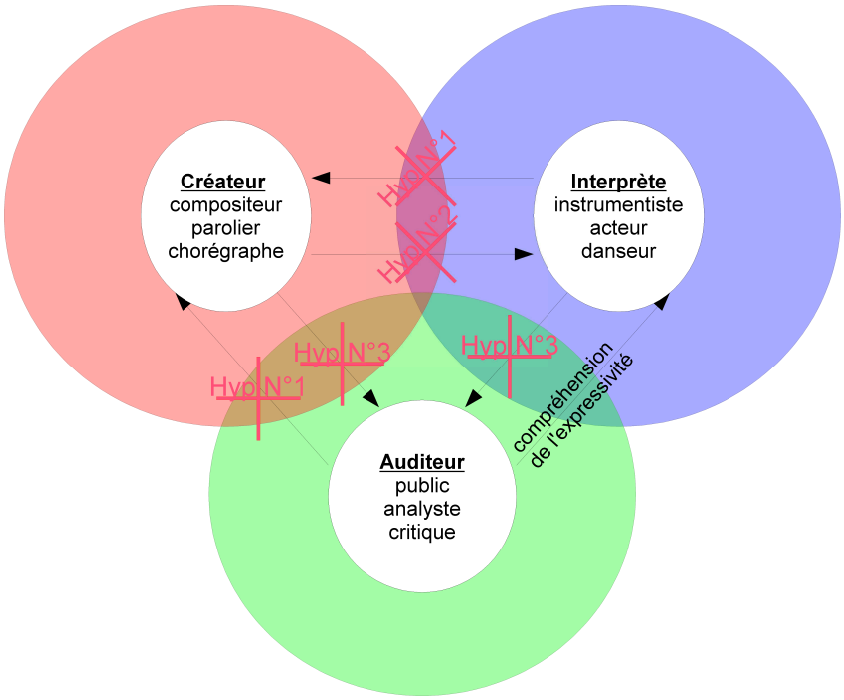

FIG. 6.4: Figure 6.3 sur laquelle des relations relatives à l'expressivité ont été évitées par le respect des hypothèses.

6.5 L'expressivité de l'interprétation musicale

Nous nous plaçons à nouveau dans le contexte de la performance, mais cette fois-ci, musicale. La performance musicale, non improvisée, implique une oeuvre, un interprète et un auditeur. Lors d'une interprétation musicale non improvisée, une oeuvre est jouée par un ou plusieurs instrumentistes pour le public. Nous nous plaçons dans le cadre réduit du jeu solo, c'est à dire de l'interprétation par un seul instrumentiste. Dans ce cas, plusieurs niveaux d'information sont accessibles, à l'instar de la parole.

6.5.1 Le modèle GERMS

Le modèle GERMS [Juslin 2003] réunit ces différents niveaux d'information qui sont rendus accessibles par le jeu d'un instrumentiste, en situation d'interprétation musicale. Ils sont relatifs au support, i.e à la partition (G), aux relations entre les contraintes mécaniques de l'instrument et les contraintes bio-anatomiques de l'instrumentiste (R et M), aux choix stylistiques (S) et à l'expressivité (E).

6.5.1.1 G : Les éléments relatifs au support

Ils permettent de clarifier la structure de l'oeuvre. Pour certaines œuvres appartenant à des courants musicaux dont les règles de composition sont connues, comme c'est le cas pour le style classique, par exemple, les règles d'interprétation sont définies à l'avance. Parmi ces règles, le rubato final (décélération) peut servir à signifier la fin d'un groupe. La vitesse de décélération peut servir à exposer des relations hiérarchiques entre groupes. D'autres règles concernent la dynamique, le timing et l'articulation [Gabrielson 1996].

6.5.1.2 E : L'expressivité

En ce qui concerne l'expressivité, une analogie avec la parole peut être faite. En effet, un instrumentiste interprète une partition écrite, déjà porteuse de plusieurs sens musicaux. Parmi ces sens musicaux figure l'émotion [Meyer 1956]. Ainsi, certaines émotions peuvent être comprises directement par la lecture de la partition. L'exemple le plus célèbre concerne la modalité mineure/majeure souvent attribuée à la valence (positive/négative) des émotions. D'autres exemples existent : Luck [Luck 2008] relie "l'activité" à une forte densité de notes et à un tatum (pulse) clairement défini, le caractère "agréable" à une densité faible de notes et à une harmonie claire, et la "force", à une forte vélocité et à une faible densité de note. L'instrumentiste peut alors actualiser ce sens par son jeu, mais peut aussi, à l'instar de l'acteur, conférer une information expressive originale à sa performance, nouvelle ou même contradictoire par rapport à celle sous-jacente à la partition [Bresin 2003].

C'est pourquoi l'expressivité d'une interprétation musicale peut être indépendante de son support (de la partition) et constitue le second niveau d'information présent dans une interprétation musicale. L'expression d'un état interne,

celui de l'œuvre et/ou celui de l'instrumentiste est fondamentale à la musique. C'est cet aspect de l'interprétation qui nous intéresse.

6.5.1.3 R : La part de l'aléa

La part de l'aléa reflète les multiples variations acoustiques incontrôlées de l'interprétation. Malgré de multiples répétitions, un instrumentiste ne réalise rigoureusement jamais la même interprétation. Il reste toujours de petites variations dues aux fluctuations du système moteur humain. A contrario, la maîtrise de la technique de l'interprète, qui peut se voir en partie comme la recherche de la minimisation de la part de l'aléa, contribue à l'identité de l'instrumentiste. Des études comparatives de plusieurs performances permettent de mettre en évidence les variations entre instrumentistes [Palmer 1989].

6.5.1.4 M : Le mouvement

Un autre élément affectant l'interprétation musicale concerne le mouvement. Friberg et Sundberg [Friberg 1999] ont montré que les ritardanti finaux suivent une fonction mathématique similaire à celle suivie par les coureurs (joggers) lors d'une phase de décélération. Shove et Repp [Shove 1995] ont proposé qu'une interprétation esthétiquement plaisante possède une microstructure satisfaisant les contraintes biomécaniques basiques. Il existe donc deux motivations pour des mouvements naturels. D'une part, ils sont recherchés pour leur aspect esthétique. D'autre part, ils reflètent l'interaction entre les contraintes anatomiques de l'instrumentiste et le principe moteur de production du son de son instrument.

6.5.1.5 S : Le style

Le dernier élément influant sur l'interprétation musicale provient du choix stylistique de l'instrumentiste. Ce choix peut éventuellement aller à l'encontre des règles génératives (G), créant un jeu de tension et de détentes vis à vis d'une attente stylistique. Le style employé par un instrumentiste dépend fortement de la partition et de l'époque à laquelle celle-ci a été écrite. Les styles de jeu sont des notions définies par des communautés (appelées aussi des "écoles"). Des règles et des normes conditionnent la production et la perception du style d'une interprétation. L'acquisition d'un style peut prendre du temps à l'instrumentiste qui doit en apprendre les règles, et maîtriser la technique sous-jacente. Certains styles reposent sur le choix de rendre le jeu plus expressif ou non. Même si le style peut contraindre l'expression d'une interprétation, celle-ci reste observable tant que l'on connaît le style employé.

6.5.2 Les niveaux d'information

Dans le cadre de l'interprétation musicale solo non improvisée, la partition est livrée au public par un instrumentiste. De manière à percevoir l'émotion musicale,

l'auditeur va chercher dans l'interprétation, les traces acoustiques relatives à l'expressivité. Consciemment ou non, il va devoir trier parmi les différentes informations présentes dans l'interprétation musicale :

- l'identité de l'instrumentiste, relative à sa technicité et à sa maîtrise de l'aléa
- le style employé
- la structure de la pièce, la pièce elle-même
- le respect du mouvement naturel
- l'expressivité

6.6 Expressivité de l'interprétation

La performance donnée par l'interprète comporte d'autres informations que l'expressivité. Tout d'abord, l'identité de l'interprète est déclinée. Puis son adhérence/répulsion face à certaines normes, à travers le style. Des relations avec le message qu'il transmet (le support) et la manière dont il les transmet apparaissent aussi. Toutes ces informations peuvent, bien entendu, influencer notre perception de l'expressivité. Pour n'observer que l'influence de l'expressivité d'une interprétation, il est nécessaire de démêler les influences de chacun de ces niveaux d'information. En effet, ceux ci sont présents dans l'interprétation en des effets qui pourraient être malencontreusement attribués à l'expressivité, si nous n'avons pas connaissance de ces niveaux d'information.

La confrontation des manifestations acoustiques de l'expressivité de l'interprétation musicale et de l'interprétation verbale permet de mettre en évidence les différents niveaux d'information propres à ces moyens de communication. L'analogie favorise la proposition d'une définition de l'expressivité qui est complémentaire à la définition donnée précédemment (voir partie 1).

En effet, d'un côté, les niveaux d'information accessibles dans la parole sont :

- l'identité du locuteur, porté par les caractéristiques de sa voix
- le style de parole emprunté
- le message sémantique, porté par les mots et la syntaxe
- l'aspect pragmatique de la parole en contexte
- l'expressivité.

De l'autre côté, les niveaux d'information présents dans l'interprétation musicale sont :

- l'identité de l'instrumentiste, relative à sa technicité et à sa maîtrise de l'aléas
- le style employé
- la pièce et sa structure
- le respect du mouvement naturel
- l'expressivité.

La mise en parallèle de ces niveaux d'information liés à la parole et à l'interprétation musicale permet de définir les niveaux d'information liés, plus généralement, à l'interprétation :

- l'identité (technique de l'instrumentiste ou caractéristique de la voix du locuteur)
- le style (de jeu ou de parole)
- le support (dont la structure doit être explicitée ; texte ou partition)
- le mouvement (tour de parole, mouvement physique)
- l'expressivité.

Dans le but de dresser ce parallèle de manière plus précise, nous réunissons pour chacun de ces niveaux d'information, des éléments qui remplissent des rôles similaires. De plus et dans le but de compléter nos hypothèses pour l'étude de l'expressivité de l'interprétation, nous donnons quelques contraintes supplémentaires pour l'établissement du protocole expérimental.

6.6.1 Le support

Le support représente le texte ou la partition qui va être interprété dans le contexte d'une performance non improvisée.

Une différence notable entre les deux types de support réside dans le sens qu'on leur attribue. Le sens musical et le sens sémantique diffèrent bien que certaines ressemblances (comme l'utilisation de la citation, par exemple) subsistent. Cela pose des problèmes quant à la définition d'un support musical d'expression neutre. En effet, s'il paraît possible de définir un texte sémantiquement neutre vis à vis de l'expressivité, cela semble, en revanche, difficile de trouver une phrase musicale dénuée d'expressivité. Pour enregistrer les performances d'une même phrase musicale avec des expressions différentes, il faut définir une phrase musicale d'expressivité neutre, dont la partition offre un sens musical pour chacune des expressions.

Hyp N°4 : Le support de l'expérience doit être d'expression neutre. C'est à dire que celui-ci doit accueillir un sens (sémantique ou musical) pour chacune des expressions.

6.6.2 L'identité

Afin de marginaliser les variations dues aux changements de l'identité de l'interprète, une solution consiste à impliquer, au sein de la même expérience, plusieurs interprètes (locuteurs ou instrumentistes).

Hyp N°5 : L'expérience doit faire intervenir plusieurs interprètes afin de marginaliser la variation due à l'identité.

6.6.3 Le style

Parce que le style amène les interprètes à adhérer à ou a se dissocier de certaines règles, qui peuvent avoir des conséquences sur l'expressivité de l'interprétation, il est important de connaître le style employé par l'interprète.

Hyp N°6 : Le style employé par l'interprète, ainsi que ces conséquences sur l'interprétation, doivent être connus et définis avant l'expérience.

6.6.4 L'aspect pragmatique

Si la notion de mouvement naturel dans l'interprétation musicale ou dans la danse est évidente, elle semble plus difficile à considérer dans le cas de la parole. Or, la confrontation des niveaux d'information place les aspects pragmatiques en tant que candidat potentiel à la représentation du mouvement de la parole. Si plusieurs expressions du language populaire, comme "passer la parole" renforce l'analogie, il n'en reste pas moins à démontrer que le niveau pragmatique engendre des effets qui suivent les règles élémentaires du mouvement. Comme si l'attention portée à la parole était une balle que les différents protagonistes se renverraient grâce au niveau pragmatique.

6.6.5 L'expressivité

L'expressivité, enfin, est l'élément commun, par excellence, à la parole et à l'interprétation musicale [Meyer 1956]. Le locuteur ou l'instrumentiste va modifier son interprétation afin de conférer à sa performance une certaine expression. Celle-ci peut être voulue par le support ou bien adjointe par l'interprète. Ses états internes pourront aussi influencer l'expressivité avec laquelle il va jouer. Les hypothèses de la partie 6.4.5 doivent être respectées dans le cas de la parole, tout comme dans le cas de l'interprétation musicale, si l'on souhaite faire de l'expressivité un objet d'étude scientifique.

6.6.6 Conclusion

Cette mise en parallèle permet donc de définir l'expressivité d'une interprétation, comme un des niveaux d'information disponibles dans une performance. Afin de l'étudier, il est alors nécessaire de séparer ces effets de ceux découlant des autres niveaux d'information. Dans le cadre d'une expérience, trois hypothèses supplémentaires à celles décrites dans la partie 6.4.5 doivent être respectées de manière à connaître les influences du support, de l'identité et du style sur l'interprétation. La ressemblance entre ces deux moyens d'expression que sont la musique et la parole, suscite une possible généralisation à toute forme de performance artistique, possédant un support écrit (non improvisé) et qui est interprétée par un individu. L'expressivité dans la danse est analysable si l'on connaît la chorégraphie, la personnalité du danseur (sa technique), son style ainsi que les règles régissant le "naturel du mouvement".

6.7 Corpus d'interprétations musicales

Dans cette partie, la constitution d'un corpus musical est présenté. Le protocole expérimental tente de donner satisfaction aux six hypothèses exposées précédemment tout en composant avec certaines contraintes techniques ou expérimentales. Ces six hypothèses sont réunies ici.

Hyp N°1 : L'auditeur et l'interprète sont en accord sur l'expressivité du support.

Hyp N°2 : L'interprète est dans l'état interne neutre lors de son interprétation.

Hyp N°3 : L'auditeur est dans l'état interne neutre lors de l'interprétation.

Hyp N°4 : Le support de l'expérience doit être neutre vis à vis de l'expressivité. C'est à dire que celui-ci doit posséder un sens (sémantique ou musical) pour chacune des expressions.

Hyp N°5 : L'expérience doit faire intervenir plusieurs interprètes afin de marginaliser la variation due à l'identité.

Hyp N°6 : Le style employé par l'interprète, ainsi que ces conséquences sur l'interprétation, doivent être connus et définis avant l'expérience.

Les corpus de parole expressive réalisés, présentés dans la partie 2 satisfont à ces contraintes. Le corpus d'interprétations musicales est présenté, même s'il a été enregistré en fin de thèse, car il sera utilisé dans des études ultérieures.

6.7.1 Particularités de l'interprétation musicale

Une particularité de l'interprétation musicale relève de la variété des instruments utilisés, qui peuvent parfois être très éloignés, de part leur mode de production, de la parole, comme c'est le cas pour les instruments percussifs. Par exemple, le piano ne possède pas de variation dynamique de la hauteur, alors que celle-ci est fondamentale à la parole. Cependant, un pianiste joue des phrases, alors que ses doigts n'ont pas besoin de respirer. Ces phrases sont entrecoupées de pauses, parfois appelées "respirations", qui ont une importance capitale pour l'expressivité [Bresin 2000]. Ainsi, il semble que l'expressivité du jeu d'un pianiste contienne des gestes acoustiques similaires à ceux employés par la voix.

Si ces gestes acoustiques sont réalisables de différentes manières, selon la source acoustique employée (comme le montre l'exemple du piano), ils semblent résulter en des effets similaires à ce que pourrait produire la parole. La respiration, qui est commune à tous les instrumentistes et à tous les acteurs, semble être un candidat d'étude idéal pour la définition de gestes acoustiques expressifs. Pour qu'une étude comparative des gestes acoustiques expressifs dans la parole et dans l'interprétation musicale puisse être menée, il est nécessaire de réunir deux corpus d'étude dont les contraintes sont identiques. Du côté de la parole, le corpus est présenté dans le

chapitre 2.6.5. Les respirations y ont été notamment annotées (voir partie 3.4). Du côté de l'interprétation musicale, nous avons sollicité des violonistes, dans le but de créer une corpus multi-culturel pour l'étude de l'expressivité de l'interprétation musicale. Pour cette étude pilote, le violon a été choisi car il produit des sons dont la perception est proche de celle de la voix.

6.7.2 Support : Partition utilisée

La difficulté majeure, dans la constitution de ce corpus, consiste à définir une phrase musicale "sémantiquement neutre", c'est à dire une phrase musicale dépourvue d'expressivité, ou encore, une phrase musicale qui possède un sens musical, quelque soit l'expression avec laquelle elle est jouée. Dans un premier temps, le support musical neutre choisi (la partition), a été défini par une gamme chromatique ascendante, puis descendante. Cependant, le choix d'une gamme s'est vu critiqué car celle-ci peut musicalement signifier l'exercice, l'effort et les émotions associées. De plus, le contour mélodique peut influencer la perception de l'expressivité. C'est pourquoi la phrase musicale neutre choisie, dans un second temps, est constituée de la succession de 8 notes identiques dont seule peut varier le registre (l'octave).

6.7.3 Identité et style

Dans l'expérience choisie, nous avons convoqué 3 instrumentistes (violonistes) de manière à marginaliser la variation due à l'identité. De plus, nous avons opté pour 3 violonistes professionnels, de formation musicale classique.

6.7.4 expressions

Si le chapitre 2.2.5 montre que les émotions verbales et les émotions musicales ne sont pas forcément les mêmes, il montre aussi que certaines catégories (comme les émotions utilitaires) peuvent être exprimées par ces deux moyens d'expression. Les expressions retenues pour ce corpus sont les suivantes.

- *Neutre* : " L'état d'être émotionnellement neutre, ni émotion positive, ni émotion négative. "
- *Joie* : " L'état d'un agréable contentement de l'esprit, qui résulte du succès ou de l'accomplissement de ce que l'on considère être bon. "
- *Colère* : " Un sentiment de mécontentement réveillé par la blessure réelle ou imaginée et d'habitude accompagné par le désir d'exercer des représailles. "
- *Tristesse* : " La condition ou la qualité d'être triste. "
- *Peur/Nervosité* : " Un sentiment d'agitation et d'anxiété causée par la présence ou l'imminence d'un danger. "
- Humour/Amusement : " La qualité de quelque chose qui la rend ridicule ou amusante. "
- Désir : " Un désir fort, particulièrement pour quelque chose d'inaccessible. "
- Solennité : " Un sentiment solennel et digne. "
- Spiritualité : " Sensibilité ou attachement à des valeurs religieuses. "

– Tendresse/Affection : " Un tendre sentiment envers une personne (amical ou amoureux). "
– Apaisement : " La qualité ou l'état d'être paisible ; liberté vis-à-vis des perturbations ou de l'agitation. "

Les expressions, en italique, correspondent à des émotions utilitaires, aussi présentes dans la parole. Cette partie du corpus pourra donc servir à la comparaison. Il est à noter que, dans ce cas, nous avons choisi de représenter les expressions par une catégorie, munie d'une description sémantique.

6.7.5 Contenu du corpus

Le corpus se compose des enregistrements répondant aux trois tâches suivantes :

1. Exprimer les 10 expressions, en choisissant une pièce de musique pour chacune d'entre elles (c'est-à-dire 10 pièces différentes). La pièce que vous choisissez, pour une émotion, doit faciliter l'expression de cette émotion. (Au total, 10 interprétations)
2. Choisir une pièce de musique qui vous semble relativement neutre du point de vue émotionnel, et jouez-la de manière à lui conférer chacune des 10 émotions. Vous devez aussi jouez cette même pièce de manière à ce qu'elle sonne émotionnellement neutre. Notez qu'il faut jouer rigoureusement la même mélodie (les mêmes notes) pour chacune des émotions. (Au total, 11 interprétations)
3. Exprimez chacune des 10 émotions (+ neutre) en jouant 8 fois la même note. Vous êtes libre de changer tous les aspects de votre performance dans le but de lui conférer l'expression désirée, avec la seule restriction de n'utiliser qu'une seule note par émotion. Par exemple, si vous choisissez un "Do", alors chacune des 8 notes jouées doit être un "Do" (Bien que vous puissiez jouer des "Do" dans différents registres). (Au total, 11 interprétations)

Ces instructions, destinées aux instrumentistes, ont permis la création d'un corpus original d'interprétations musicales. Le choix d'expressions communes à la parole et à la musique, va permettre une confrontation acoustique des deux corpus présentés précédemment[9]. Cette confrontation semble d'ores et déjà prometteuse, compte tenu d'exemples sonores déjà produits. Malheureusement, elle ne peut être relatée dans ce manuscrit puisque les enregistrements se font de manière simultanée à sa rédaction. Toutefois quelques pistes pour la comparaison sont données dans la partie suivante.

[9] De plus, dans le cadre d'une étude multi-culturelle de l'expressivité de l'interprétation musicale, ce protocole va être mis en œuvre dans d'autres parties du globe (Japon, Inde et Suède), avec des instruments locaux qui ressemblent au violon (cordes frottées), respectivement le kokyu, le sarangi et le violon.

6.8 Comparaison de l'expressivité des interprétations verbale et musicale

Sans poser la question de l'œuf et de la poule : "Est-ce la musique qui a précédé la parole ou l'inverse ?" ou "Y a t'il une imitation, de mêmes origines, un langage préverbal commun ?", on peut se demander si l'interprète à qui l'on demande de jouer une phrase musicale "sémantiquement neutre" avec son instrument (par exemple, une gamme chromatique), de manière expressive, n'imite pas les effets de la parole expressive. En effet, les exemples sonores disponibles à l'adresse suivante : `http://recherche.ircam.fr/equipes/analyse-synthese/beller`, produits par un violoncelliste, semblent conforter cette idée, tant la proximité acoustique avec des phrases dites par un acteur semble corrélée à l'expressivité. La similitude de ces gestes acoustiques dans la parole et dans l'interprétation musicale conforte l'idée d'un champ de gestes acoustiques expressifs commun.

6.8.1 Prosodie instrumentale

Comme le montre le modèle de la parole utilisé (voir partie 3.2), ainsi que les différentes analyses effectuées sur le corpus de parole expressif (voir partie 3.5), la prosodie ou la manière de parler est un élément capital pour l'analyse de l'expressivité dans l'interprétation verbale. Par analogie, la *prosodie instrumentale* semble être un bon modèle pour définir des informations acoustiques relatives à l'expressivité. Cependant, la prosodie instrumentale est propre à chaque instrument. Pour le violon, tout comme pour la voix, elle possède cinq dimensions :

- l'intonation : écart relatif de la hauteur par rapport aux notes de la partition
- l'intensité : écart relatif de l'intensité par rapport aux nuances de la partition
- la vitesse d'exécution : écart temporel relatif instantané par rapport au tempo, aux durées et aux onsets de la partition
- le degré d'articulation : caractérisant l'articulation entre les notes [Rasamimanana 2008, Rasamimanana 2007]
- la qualité de la source : caractérisant les effets de la pression de l'archet sur la corde

Par analogie avec la parole, la prosodie instrumentale peut-être décrite comme une variation de paramètres acoustiques par rapport à ce qui est attendu, c'est à dire la réalisation de la partition. Une étude comparative des rapports entre la prosodie et l'expressivité dans l'interprétation musicale et dans l'interprétation verbale semble donc être propice à la découverte de gestes acoustiques de l'expressivité. C'est cette étude que nous allons entreprendre dans la suite, en débutant par la comparaison des analyses du corpus de parole expressive IrcamCorpusExpressivity et des analyses du corpus d'interprétations musicales, jouées par le violon. En ce qui concerne les prosodies instrumentales d'autres instruments, la prosodie instrumentale reste à être définie selon les possibles variations acoustiques que les instrumentistes sont capables de produire. Le choix du violon permet de faciliter cette approche puisqu'il est généralement admis que les cordes frottées produisent des

sons, très proches de ceux produits par la voix, d'un point de vue de la perception. La proposition centrale de la thèse présente l'expressivité comme une modulation de l'expression neutre. L'expression neutre contient toute l'information structurelle, qui dans le cas de l'interprétation musicale, peut se déduire de la partition. La modulation relative à l'expressivité devient alors identifiable par la comparaison de paires neutre-expressive. C'est ce paradigme qui sera utilisé pour identifier les variations relatives à l'expressivité dans l'interprétation musicale et pour les comparer avec celles observées dans la parole.

6.9 Conclusion

Cette partie prospective a permis d'examiner ce qu'il se passe dans le cas de l'interprétation musicale, et dans le cas de l'interprétation en général. L'idée centrale est que la parole et la musique partage un pouvoir expressif. La première question abordée concernant ce pouvoir expressif a été : Est-ce qu'il est commun ? Une revue des catégories émotionnelles que l'on trouve dans la musique a permis de montrer qu'il existe certaines expressions communes à la parole et à la musique, bien que la majorité d'entre elles soient différentes, compte tenu des différents rôles que tiennent ces deux médiums de communication. Parmi ces expressions communes figurent les émotions utilitaires, parfois aussi exprimées par la musique. La seconde question abordée a été : Peut-on observer l'expressivité de l'interprétation musicale ? Une analogie a alors été construite entre la parole et la musique, sur la base de l'interprétation d'un support. L'observation des différents niveaux d'information de la parole et de l'interprétation musicale a permis de bâtir un ensemble d'hypothèses pour l'observation de l'expressivité dans l'interprétation. Ces hypothèses, qui ont été respectées pour la constitution du corpus de parole expressif IrcamCorpusExpressivity, ont permis la constitution d'un corpus d'interprétations musicales expressives, semblable aux corpus de parole. Ce corpus a été présenté car il a été enregistré durant cette thèse. Malheureusement son exploitation n'a pas été relatée par ce manuscrit, bien que des pistes aient été données. Parmi ces pistes, un axe de recherche a été évoqué, à partir de l'écoute comparée des corpus de parole et d'interprétations musicales : Par analogie, l'expressivité de la parole est en partie révélée par la prosodie, ce qui laisse à penser que l'expressivité de l'interprétation musicale puisse être aussi, en partie, révélée par ce que l'on appelle, une prosodie instrumentale. Cette perspective a été développée, pour indiquer nos axes futurs de recherche, et parce qu'elle peut contribuer à l'amélioration de notre modèle génératif. S'il existe des gestes acoustiques de l'expressivité, communs à la parole et à l'interprétation musicale, alors leur confrontation peut nous aider à mieux les cerner, dans chacun de leurs contextes respectifs. Ce qui nous amène à la dernière question : L'expressivité se révèle t'elle dans la parole et dans l'interprétation musicale, par des gestes acoustiques communs ?

Annexe A

Annexe : ICT, plateforme pour la gestion des corpus

Sommaire

A.1 Résumé du chapitre

Il existe un éventail d'outils pour la création, l'accès et la synchronisation des données d'un corpus de parole, mais ils sont rarement intégrés dans une seule et même plateforme. Dans cette partie, nous proposons IrcamCorpusTools, une plateforme ouverte et facilement extensible pour la création, l'analyse et l'exploitation de corpus de parole. Elle permet notamment la synchronisation d'informations provenant de différentes sources ainsi que la gestion de nombreux formats. Sa capacité à prendre en compte des relations hiérarchiques et séquentielles permet l'analyse contextuelle de variables acoustiques en fonction de variables linguistiques. Elle est déjà employée pour la synthèse de la parole par sélection d'unités, les analyses prosodique et phonétique contextuelles, pour exploiter divers corpus de parole en français et d'autres langues, et bien sur, pour la modélisation de l'expressivité. Cette plateforme a été réalisée durant la thèse avec l'aide de C. Veaux, G. Degottex et N. Obin. Une vue plus détaillée de la plateforme est donnée dans l'article de la revue TAL [Beller 2009b].

A.2 Introduction

Les méthodes à base de corpus sont désormais très largement répandues en traitement de la parole et en traitement du langage pour le développement de modèles théoriques et d'applications technologiques. Que ce soit pour vérifier des heuristiques, découvrir des tendances ou modéliser des données, l'introduction de traitements calculatoires et/ou statistiques basés sur les données des corpus a multiplié les possibilités et permis des avancées considérables dans les technologies de la parole et du langage. La traduction automatique, la lexicométrie et l'inférence de règles grammaticales en sont des exemples en traitement automatique des langues (TAL). La reconnaissance [Zarader 1997] et la synthèse de parole [Bailly 1992, Dutoit 1997] en sont d'autres pour le traitement automatique de la parole (TAP). De plus en plus, les besoins et les questions des deux communautés TAL et TAP se rapprochent comme le montre la récente fusion du traitement de l'oral avec celle du langage naturel. De même l'utilisation de corpus annotés prosodiquement tel que le corpus LeaP[1] intéresse aussi bien la recherche en linguistique que celle en traitement de la parole. Toutefois, cette complémentarité n'est possible que par la mise en commun des corpus. C'est pourquoi les questions de représentation et de gestion des données des corpus sont centrales.

Les corpus oraux sont constitués de deux types principaux de ressources, les signaux temporels et les annotations. Les signaux temporels sont les enregistrements audio, vidéo et/ou physiologiques, ainsi que toutes les transformations s'y rapportant (fréquence fondamentale, spectrogramme, ...). Les annotations sont la transcription textuelle ainsi que toutes les notations ajoutées manuellement ou automatiquement (transcription phonétique, catégories grammaticales, structure du discours, ...). Les différents niveaux d'annotations possèdent généralement des relations hiérarchiques et/ou séquentielles et sont synchronisés avec l'axe temporel. Les outils de gestion des corpus recouvrent tout un ensemble de fonctionnalités allant de la création et de la synchronisation des ressources, aux requêtes (pouvant porter autant sur les annotations que sur les signaux temporels), en passant par le stockage et l'accès aux données. La plupart des systèmes de gestion de corpus existants ont été développés pour des corpus spécifiques et sont difficilement adaptables et extensibles [Oostdijk 2000]. Des efforts ont été faits pour faciliter l'échange de données par la conversion de formats [Gut 2004] ou pour dégager une représentation formelle pouvant servir d'interface commune entre les divers outils et les données [Bird 2000].

Cette notion d'interface entre les méthodes et les données est à la base de la plateforme IrcamCorpusTools présentée dans cette partie. Cette plateforme utilise l'environnement de programmation Matlab/Octave afin d'être facilement extensible. Elle permet notamment la synchronisation d'informations provenant de différentes sources (vidéo, audio, symbolique, ...) ainsi que la gestion de nombreux formats (XML, SDIF, AVI, WAV, ...). Elle est munie d'un langage de requête pre-

[1] Learning Prosody Project : http://leap.lili.uni-bielefeld.de

nant en compte les relations hiérarchiques multiples, les relations séquentielles et les contraintes acoustiques. Elle permet ainsi l'analyse contextuelle de variables acoustiques (prosodie, enveloppe spectrale, ...) en fonction de variables linguistiques (mots, groupe de sens, syntaxe, ...). Elle est déjà employée pour la synthèse de la parole par sélection d'unités, les analyses prosodique et phonétique contextuelles, la modélisation de l'expressivité et pour exploiter divers corpus de parole en français et d'autres langues.

Dans un premier temps, cette partie présente les problématiques que doivent résoudre les systèmes de gestion de corpus de parole en donnant quelques exemples de plateformes existantes. Dans un second temps, il décrit comment IrcamCorpusTools apporte des solutions originales à ces problématiques. La connaissance de Matlab/Octave peut aider à la compréhension de notre choix pour cet environnement de programmation, mais n'est en aucun cas requise pour la compréhension de cette partie, qui propose d'ailleurs au lecteur, une courte présentation générale de l'environnement.

A.3 Systèmes de gestion et de création de corpus de parole

Depuis l'essor de la linguistique de corpus [Chafe 1992], de nombreux corpus annotés ont été exploités par le TAL, dont des corpus oraux comme ceux recensés par LDC[2]. La nécessité de traiter une grande quantité de métadonnées linguistiques est inhérente aux problèmes posés par le TAL. Aussi, de nombreux systèmes de gestion de larges corpus sont aujourd'hui disponibles pour cette communauté [Cunningham 2002]. Dans le domaine du TAP, le corpus TIMIT fut le premier corpus annoté à être largement diffusé. Une tendance actuelle est à l'utilisation de corpus multimodaux avec l'intégration de données visuelles, ce qui accroît encore la diversité des formats à gérer. Permettre à une communauté de chercheurs de partager et d'exploiter de tels corpus ne pose pas simplement la question de la gestion des formats, mais aussi celles de la représentation des données, du partage des outils de génération, d'accès et d'exploitation, et du langage de requêtes associé.

A.3.1 Modèle de représentation des données

Un modèle de représentation des données doit pouvoir capturer les caractéristiques importantes de celles-ci et les rendre facilement accessibles aux méthodes les traitant. Ce modèle constitue en fait une hypothèse sous-jacente sur la nature des données et sur leur structure. Il doit donc être aussi général que possible afin de pouvoir représenter différents types de structures phonologiques et permettre une grande variété de requêtes sur ses structures.

Les modèles principalement utilisés en TAL sont des structures hiérarchiques

[2]Linguistic Data Consortium : http://www.ldc.upenn.edu/Catalog/

comme celles du Penn Treebank[3] qui peuvent être alignés temporellement dans le cas des corpus oraux. Certains systèmes comme Festival [Taylor 2001] ou EMU [Cassidy 2001] vont au delà de ces modèles en arbre unique et supportent des hiérarchies multiples, c'est-à-dire qu'un élément peut avoir des parents dans deux hiérarchies distinctes sans que ces éléments parents soient reliés entre eux. Ces représentations sont particulièrement adaptées pour les requêtes multi-niveaux sur les données du corpus. D'autres approches telles que Bird et Lieberman [Bird 2001] ou Müller [Müller 2005] se concentrent sur des représentations des données qui facilitent la manipulation et le partage des corpus multi-niveaux. Il s'agit généralement de représentations temporelles des données qui explicitent uniquement la séquence des événements, les relations hiérarchiques étant représentées implicitement par la relation d'inclusion entre les marques temporelles. Enfin, Gut et collaborateurs [Gut 2004] exposent une méthode et des spécifications minimales permettant de convertir entre elles les différentes représentations des données utilisées par les corpus.

A.3.2 Partage des données

Afin de pouvoir partager les corpus, comme dans le cas du projet PFC[4] [Durand 2005], des efforts de standardisation ont été entrepris à différents niveaux. Un premier niveau de standardisation consiste à établir des conventions sur les formats de fichiers et les méta-données décrivant leur contenu. Ainsi, le format XML[5] s'est de plus en plus imposé comme le format d'échange des annotations. Cette solution permet la compréhension des données par tous les utilisateurs, tout en leur permettant de créer de nouveaux types de données selon leurs besoins. Un second niveau consiste à standardiser le processus de génération des données elles-mêmes. Cela conduit par exemple à des recommandations comme celles de la Text Encoding Initiative[6] pour les annotations des corpus oraux. Certains projets, tel CHILDES [MacWhinney 2000] pour l'analyse des situations de dialogues chez l'enfant, proposent à la fois des normes de transcription et les outils conçus pour analyser les fichiers transcrits selon ces normes.

A.3.3 Partage des outils

Des efforts ont également été entrepris pour créer des outils libres adaptés aux annotations des ressources audio et/ou vidéo des corpus comme Transcriber[7] [Barras 1998] ou ELAN du projet DOBES[8]. Vis-à-vis des outils pour l'annotation, des outils de visualisation et d'analyse acoustique sont disponibles et largement uti-

[3]Penn Treebank : http://www.cis.upenn.edu/treebank/home.html/
[4]PFC : Phonologie du Français Contemporain : http://www.projet-pfc.net/
[5]XML : eXtensible Markup Language : http://www.w3.org/XML/
[6]Text Encoding Initiative : http://www.tei-c.org/
[7]Transcriber : http://trans.sourceforge.net/en/presentation.php
[8]DOBES : documentation sur les langues rares : http://www.mpi.nl/DOBES/

lisés, comme WaveSurfer[9] [Sjölander 2000] ou Praat[10] [Boersma 2001]. Ces logiciels permettent l'analyse, la visualisation/annotation, la transformation et la synthèse de la parole. Ils sont programmables sous la forme de scripts pour Praat et sous la forme de « plugins » pour WaveSurfer. Malheureusement, le choix de TCL/TK[11] pour ces logiciels n'est pas répandu dans les communautés du traitement du signal, de la modélisation statistique, du calcul numérique ou de la gestion de base de données. Le choix d'un format propriétaire pour les données, dans le cas de Praat, réduit considérablement les possibilités de partage de ces données qui nécessitent une étape de conversion. Cela amène ces plateformes dédiées à la phonétique à incorporer quelques méthodes statistiques et des machines d'apprentissage, bien que leur langage de programmation ne soit pas adéquat aux calculs numériques. D'ailleurs, bien qu'affichant des annotations, ces logiciels ne sont pas munis de langage de requêtes, ni de systèmes de gestion de base de données.

A.3.4 Langage de requête

Pour être exploitable par une large communauté d'utilisateurs, un corpus doit être muni d'un langage de requête qui soit à la fois simple et suffisamment expressif pour formuler des requêtes variées. Une liste minimale de requêtes existe pour tout système de gestion des corpus [Lai 2004]. L'outil de requête doit aussi offrir une bonne « extensibilité », c'est-à-dire pouvoir traiter de larges corpus en un temps raisonnable. On peut distinguer deux grandes familles de systèmes utilisés pour stocker et rechercher de l'information structurée, les bases de données et les langages de balisages de textes comme le XML. Des exemples de systèmes de requête basés sur XML sont le Nite XML [Gut 2004] ou la version initiale d'EMU [Cassidy 2001]. Selon ces approches, les relations hiérarchiques multiples entre les données sont stockées dans une série de fichiers XML mutuellement liés. Les langages de requête comme XSLT/XPath sont naturellement adaptés à la formulation des contraintes d'ordre hiérarchiques mais la syntaxe des requêtes se complique lorsqu'il s'agit d'exprimer des contraintes séquentielles. Un effort de simplification de ces requêtes est proposé par Gut et collaborateurs [Gut 2004] avec le langage NXT Search. Cependant, les systèmes basés sur le XML offrent une « extensibilité » limitée car ils nécessitent une recherche linéaire dans le système de fichiers [Cassidy 2001]. A l'inverse, les systèmes de base de données sont capables de stocker de très grandes quantités d'information et d'effectuer des requêtes rapides sur celles-ci. Il a été montré que les requêtes sur les hiérarchies multiples peuvent être traduites en langage SQL [Cassidy 2001]. Cependant, le modèle relationnel étant par nature moins adapté à la représentation des contraintes hiérarchiques et séquentielles que le XML, une requête donnée en XML se traduit de manière beaucoup plus complexe en SQL. Si des langages intermédiaires plus simples comme LQL ont été proposés [Nakov 2005], les requêtes les plus complexes ne sont pas toujours formulables selon

[9] WaveSurfer : http://www.speech.kth.se/wavesurfer/
[10] Praat : http://www.fon.hum.uva.nl/praat/
[11] TCL/TK : http://www.tcl.tk/

cette approche.

A.3.5 Exploitation des données

Une fonctionnalité essentielle des plateformes de gestion de corpus est la possibilité d'interfacer les données (éventuellement après filtrage par des requêtes) avec des outils de modélisation. Ainsi, alors que certains environnements de développement linguistique permettent de construire, de tester et de gérer des descriptions formalisées [Bilhaut 2006], d'autres se sont tournés vers les traitements statistiques [Cassidy 2001]. L'apprentissage automatique pour les tâches de classification, de régression et d'estimation de densités de probabilités est aujourd'hui largement employé. Qu'elles soient déterministes ou probabilistes, ces méthodes nécessitent des accès directs aux données et à leurs descriptions. C'est pourquoi certains systèmes de gestion de corpus tentent de faciliter la communication entre leurs données et les machines d'apprentissage et d'inférence de règles comme c'est le cas pour le projet EMU et le projet R[12].

A.4 Vers une plateforme complète

Comme nous venons de le voir, si certains outils comme Praat apportent des solutions partielles permettant l'exploitation des corpus, peu de systèmes proposent une solution complète allant de la génération des données jusqu'aux requêtes sur celles-ci. Lorsque de tels systèmes existent, ils ont été le plus souvent été conçus au départ pour une application spécifique comme la synthèse de parole [Taylor 2001] ou l'observation de pathologies comme c'est le cas pour le projet CSL (Computerized Speech Lab). Cela comporte des limitations intrinsèques sur le type de données, sur leur représentation et donc sur leur capacité à être partagées. Ainsi le chercheur à la frontière du TAL et du TAP est pour le moment contraint d'utiliser une batterie d'outils dédiés et basés sur plusieurs langages de programmation, l'obligeant à effectuer de nombreuses conversions de formats et interdisant toute automatisation complète d'un processus.

A.4.1 Environnement Matlab/Octave

Matlab est un environnement de programmation produit par MathWorks[13]. Octave[14] est une solution *open-source* qui vise les mêmes fonctionnalités et conserve une syntaxe de programmation identique. Matlab tout comme Octave fournit un langage et un environnement de programmation qui permettent d'automatiser des calculs numériques, d'afficher des résultats sous la forme de graphiques, de visualiser des données multimodales (audio, images, vidéo, ...), et de réaliser des interfaces utilisateurs sur mesure.

[12] R Project : http://www.r-project.org/
[13] MathWorks : http://www.mathworks.com
[14] Octave : http://www.gnu.org/software/octave

Le développement d'algorithmes et de logiciels prototypes en Matlab/Octave est en général beaucoup plus aisé que dans des langages compilés comme C/C++ ou Java. Ceci est dû à plusieurs propriétés :

- le langage est interprété,
- le calcul matriciel/vectoriel facilite le traitement des corpus,
- il dispose d'une très grande quantité de librairies (*toolboxes*) pour le traitement des données, l'apprentissage et l'optimisation,
- il dispose de multiples primitives et d'outils graphiques interactifs,
- de nombreux programmes libres sont disponibles sur le web[15],

Cet environnement est propice à la recherche où la rapidité de programmation est plus importante que la rapidité d'exécution.

Des scripts permettent la mise en série des commandes, et par conséquent, l'élaboration de séquences complexes reproductibles. Des interfaces graphiques sont facilement programmables et exploitables par des utilisateurs ne désirant utiliser qu'une partie des commandes disponibles. Des fonctions optimisées en C/C++ peuvent remplacer des fonctions Matlab pour accélérer les calculs. Enfin, une commande « system » permet d'envoyer directement des commandes au système d'exploitation, ce qui permet de lancer d'autres programmes depuis Matlab/Octave. Matlab/Octave est un environnement multiplateforme (Mac OS, Windows et les systèmes UNIX dont Linux) devenu extrêmement puissant et répandu, utilisé par un grand nombre de laboratoires de recherche.

Par le pouvoir expressif de son langage, par la profusion des librairies déjà disponibles, par le calcul matriciel et par une popularité forte au sein de la communauté scientifique, Matlab/Octave s'est naturellement imposé comme l'environnement idéal pour accueillir la plateforme IrcamCorpusTools.

A.5 La plateforme IrcamCorpusTools

Pour répondre aux besoins spécifiques de la parole, de son traitement et de l'analyse de corpus, la plateforme IrcamCorpusTools a été développée et est utilisée dans une grande variété d'applications. Elle s'inscrit à l'intersection de deux domaines de recherches complémentaires : la recherche linguistique et le développement de technologies vocales. Nous la présentons dans cette section en commençant par une vue générale du système et de son architecture. Puis nous présentons deux spécificités de la plateforme : son langage de requête qui prend simultanément en compte des contraintes d'ordre linguistique et des contraintes sur les signaux ; et le principe d'auto-description des données et des outils, qui permet de répondre aux problématiques de gestion et de création de corpus abordées dans la partie A.3.

[15] http://www.mathworks.com/matlabcentral/fileexchange/

A.5.1 Architecture de la plateforme

Afin de répondre à différentes demandes de recherche et de développement industriel, l'architecture d'origine [Beller 2006a] s'est naturellement orientée vers une solution extensible, modulaire et partagée par plusieurs utilisateurs/développeurs [Veaux 2008]. Cette mutualisation des outils et des données implique une certaine modularité tout en maintenant des contraintes de standardisation qui assurent la cohérence du système. La solution choisie repose sur le principe d'auto-description des données et des outils permettant de définir une interface commune entre ces objets. Une vue générale de l'architecture de ICT est offerte par la figure A.1, elle fait apparaître la couche d'interface que nous introduisons entre les données et les outils, et qui est constituée par notre environnement Matlab/Octave. Cette architecture à trois niveaux est semblable à celle proposée pour le système ATLAS [Bird 2001], elle permet à différentes applications externes ou internes de manipuler et d'échanger entre elles des informations sur les données du corpus.

Les différents éléments composant IrcamCorpusTools sont des instances (objets) de classes qui forment le cœur de la plateforme. Ces classes sont représentées dans la figure A.2. Elles sont décrites par la suite et se dénomment :

- la classe *descripteur* : classe dont les instances sont des données auto-décrites,
- la classe *unité* : classe dont les instances sont des unités reliant les données entre elles,
- la classe *analyseur* : classe dont les instances sont des analyseurs, c'est-à-dire des outils de génération, de conversion ou de manipulation des données,
- la classe *fichier* : Classe dont les instances sont des pointeurs vers un système de fichiers,
- la classe *corpus* : Classe mère regroupant un ensemble de descripteurs, d'unités et de fichiers.

Chaque classe possède un ensemble de méthodes qui constituent le langage d'interface de la plateforme. Ce langage (en anglais) a été minimisé afin de faciliter l'abord du système pour un nouvel utilisateur, et dans le but de le rendre le plus expressif possible. Nous reprenons, à présent, chacune des classes en détail.

A.5.2 Descripteurs

L'activité de la parole est intrinsèquement multimodale. La coexistence du texte, de la voix et de gestes (faciaux, articulatoires, ...) génère une forte hétérogénéité des données relatives à la parole. Le système doit être capable de gérer ces données de différentes natures. Voici les types de données gérées par IrcamCorpusTools.

A.5.2.1 Informations de type signal

Les signaux correspondent soit aux enregistrements provenant d'un microphone ou d'autres instruments de mesure (EGG, fMRI, ultrasons, ...), soit à des résultats d'analyse de ces enregistrements. Ils peuvent être unidimensionnels ou multidimensionnels. Parmi les signaux les plus courants, figurent ceux relatifs à la prosodie

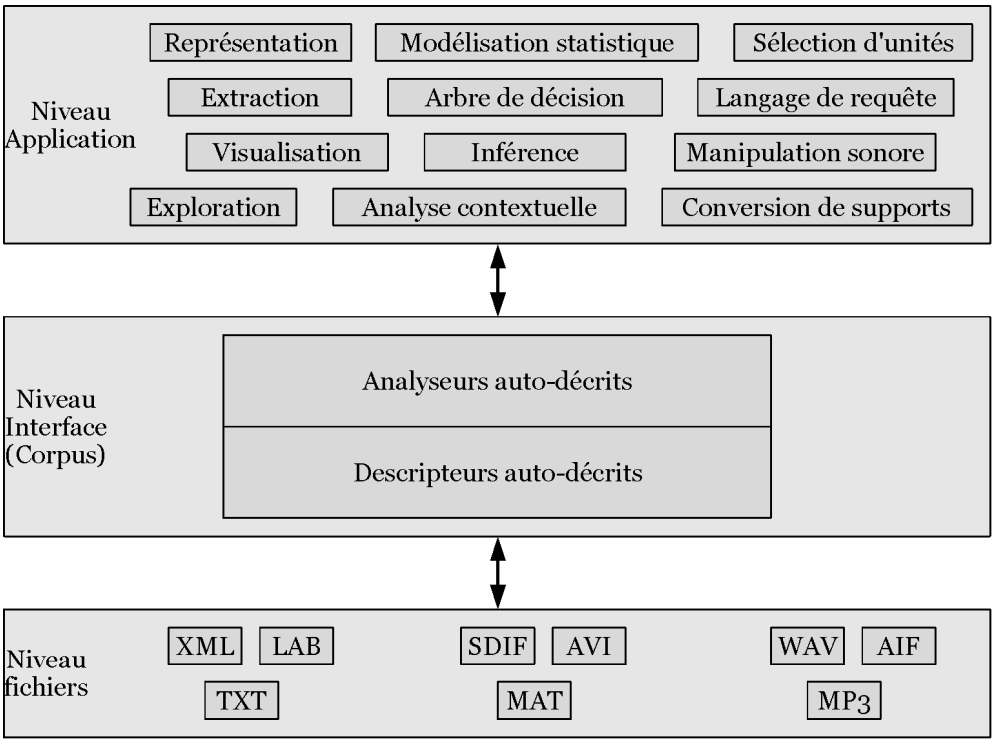

FIG. A.1: Vue d'ensemble de la plateforme IrcamCorpusTools.

comme la fréquence fondamentale f_0, l'énergie, le débit de parole, le degré d'articulation mesuré à partir des formants (fréquence, amplitude, largeur de bande), et la qualité vocale (coefficient de relaxation, modèle LF, mesure du voisement), mais aussi ceux relatifs à l'enveloppe spectrale donnés par différents estimateurs (FFT, MFCC, TrueEnvelope, LPC), et représentables sous la forme de coefficients auto-régressifs (AR), de paires de lignes spectrales (LSF), de pôles, ou d'aires de sections du conduit vocal (LAR), ... Enfin cette liste non exhaustive peut être augmentée de signaux issus d'autres modalités comme c'est le cas par exemple pour la mesure de l'aire glottique par caméra ultra-rapide.

A.5.2.2 Informations de type métadonnée

Ces informations peuvent, par exemple, servir à spécifier un contexte d'enregistrement (lieu, date, locuteur, consigne donnée, expressivité, genre de discours, ...). Elles comprennent les transcriptions textuelles *a priori* (parole lue) ou *a posteriori* (parole spontanée). Elles permettent de définir n'importe quelle information sous la forme de mots/symboles ou de séquence de mots.

A.5.2.3 Informations de type annotation

Comme les informations de type métadonnée, elles sont de nature textuelle. Mais elles possèdent, en plus, un temps de début et un temps de fin, permettant d'attribuer une information de type linguistique à une portion de signal. Cette sorte de donnée est cruciale pour une plateforme de gestion de corpus de parole, puisqu'elle permet le lien entre les signaux et les catégories linguistiques, entre la physique (flux de parole continu) et le symbolique (unités de sens discrètes). Elles sont donc les pierres angulaires à la jonction entre le TAP et le TAL. Elles constituent souvent des dictionnaires clos comme c'est le cas pour les phonèmes d'une langue ou pour d'autres étiquettes phonologiques (onset, nucleus, coda, ...). Parmi ces informations, les segmentations phonétiques sont les plus courantes. Les annotations syntaxiques, de phénomènes prosodiques ou de mots sont autant d'étiquettes qui peuvent être placées manuellement et/ou automatiquement. Elles définissent alors des segments, aussi appelés *unités* dont la durée est variable : senone, semiphone, phone, diphone, triphone, syllabe, groupe accentuel, mot, groupe prosodique, phrase, paragraphe, le discours, ...

A.5.2.4 Informations de type statistique

Sur l'horizon temporel de chacune des unités, les signaux continus peuvent être modélisés par des valeurs statistiques. Ces valeurs, décrivant le comportement d'un signal sur cette unité, sont appelées *valeurs caractéristiques* : moyennes arithmétique et géométrique, variance, intervalle de variation, maximum, minimum, moment d'ordre N, valeur médiane, centre de gravité, pente, courbure, ...

A.5.3 Unités

Les unités sont les objets permettant de relier les données entre elles. Elles sont définies pour chaque niveau d'annotation et regroupent les données symboliques ou acoustiques sur la base de la segmentation temporelle associée à ce niveau d'annotation. Les unités sont reliées entre elles par des relations de type séquentiel et/ou hiérarchique. Les relations hiérarchiques sont représentées sous la forme d'arbres (« phrase→mots→syllabes→phones », par exemple) dont les nœuds correspondent chacun à une unité. Afin de représenter des relations hiérarchiques multiples, une liste d'arbres est utilisée à la manière de Festival [Taylor 2001]. Par exemple, les unités du niveau « phone » sont dans une relation de parenté avec celles du niveau « syllabe » et avec celles du niveau « mot » ; en revanche, les syllabes et les mots n'ont pas de relation de parenté entre eux. Ces arbres permettent de propager les marques temporelles au sein d'une hiérarchie d'unités à partir d'un seul niveau d'annotation synchronisé avec le signal de parole (typiquement le niveau d'annotation issu de la segmentation phonétique). Inversement, à partir d'annotations indépendamment alignées, on peut construire les différentes hiérarchies entre unités, en se basant sur l'intersection des marques temporelles. Cela permet notamment de maintenir la cohérence des diverses données relatives aux unités, tout en

autorisant des interventions manuelles à tous les niveaux. A l'inverse des relations hiérarchiques, les relations séquentielles entre unités ne sont définies qu'au sein d'un même niveau d'annotation.

A.5.4 Fichiers

Nous avons choisi de stocker les différents *descripteurs* indépendamment les uns des autres afin de faciliter la mise à jour et l'échange des données du corpus [Müller 2005]. Ces fichiers reposent sur plusieurs supports dont les formats les plus répandus sont :

- LAB, XML, ASCII, TextGRID, pour les données de type *métadonnée* et *annotation*
- SDIF, AVI, WAV, AIFF, AU, MP3, MIDI, pour les données de type *signal*
- MAT (Matlab), pour les données de type *relation* et *statistique*

En revanche, les *unités* et leurs relations sont stockées dans un fichier unique. Une fonction permet de reconstruire les unités et leur relations lorsqu'un *descripteur* (symbolique ou acoustique) a été modifié.

A.5.5 Analyseurs

Les analyseurs regroupent toutes les méthodes de génération ou de conversion des données. On peut les enchaîner si on veut par exemple obtenir la moyenne de la fréquence fondamentale sur le groupe prosodique avoisinant une syllabe. Certaines de ces méthodes sont dites *internes* : elles sont implémentées dans l'environnement Matlab/Octave car elles nécessitent de jongler avec les différents types de données. D'autres sont dites *externes* : elles utilisent des logiciels qui ne sont pas implémentés dans l'environnement Matlab/Octave (code exécutable, script, ...) mais qui peuvent être exécutés par appel depuis IrcamCorpusTools. Grâce à l'interface du système de fichiers, les données générées par un tel logiciel sont automatiquement rendues accessibles au sein de notre environnement. D'un point de vue utilisateur, le caractère interne/externe ne fait aucune différence. Dans l'exemple cité précédemment, l'utilisateur peut remplacer un estimateur interne de la fréquence fondamentale, par exemple par celui de Praat, de WaveSurfer ou de SuperVP [Bogaards 2004], sans avoir à changer d'environnement.

A.5.6 Corpus

Un corpus peut être représenté comme un ensemble d'énoncés. Chacun de ces énoncés est un ensemble d'analyses. Chacune de ces analyses comportent un ou plusieurs descripteurs. Par exemple, l'analyse « audio » comporte le descripteur « forme d'onde » qui n'est autre que le signal acoustique de la phrase enregistrée. Ces analyses sont donc synchronisées au niveau de la phrase dans un corpus. Mais une synchronisation plus fine existe aussi grâce à l'ajout d'unités décrites par l'analyse « segmentation ». Les objets « Corpus » sont des interfaces avec le système de

fichiers. Lorsqu'un analyseur est appliqué à un corpus, celui-ci fait appel à des fichiers d'entrée et de sortie. Le corpus stocke toute création/suppression d'un fichier, auquel il adjoint les paramètres de configuration de l'analyseur employé, ainsi que des objets descripteurs. L'objet Corpus est lui-même stocké dans un fichier XML à la racine du système de fichier, ce qui permet à plusieurs personnes d'ajouter ou de supprimer des données dans un corpus sans que cela n'entraîne de conflit. En effet, l'objet Corpus conserve au fur et à mesure l'historique des opérations effectuées sur un corpus et lui confère donc un accès multi-utilisateur.

A.5.7 Langage de requête

Certains outils de requête XML (Xpath, Xquery, NXT search) présentent une syntaxe complexe. Dans IrcamCorpusTools, nous privilégions le pouvoir expressif du langage de requête. Une requête élémentaire est ainsi constituée :

1. du niveau dans laquelle on effectue la recherche d'unité
2. d'une relation séquentielle par rapport à l'unité recherchée
3. d'une relation hiérarchique par rapport à l'unité recherchée
4. d'une condition à tester sur les données numériques associées aux unités

Ces requêtes sont rapides car elles ne s'appliquent qu'aux données préalablement stockées en mémoire vive. De plus, elles peuvent être composées afin de faire des recherches complexes prenant en compte l'interaction entre les multiples niveaux d'unités.

A.5.8 Principe d'auto-description

Le pouvoir expressif du langage de requêtes provient de la possibilité de mélanger des contraintes sur des données de types différents. Cela est rendu possible par le principe d'auto-description sur lequel repose IrcamCorpusTools. Chaque instance d'une classe (corpus, fichier, analyseur ou descripteur) est accompagnée de métadonnées décrivant son type, sa provenance, comment y accéder et comment la représenter. Cela permet une compréhension et une exploitation immédiate de tous les objets par tous les utilisateurs, mais aussi par le système lui-même. A l'instar du caractère interne/externe des analyseurs, l'hétérogénéité des données est invisible à l'utilisateur qui ne possède qu'un seul lexique restreint de commandes avec lesquelles il peut rapidement se familiariser. Aucune donnée ne se « perd », car l'objet corpus garde une trace des différentes opérations réalisées sur lui et donc, des différentes analyses ayant généré ses données. Cela permet notamment de conserver un historique de l'accès aux données. En effet, on peut toujours accéder à d'anciennes informations, même si la méthode d'accès à celles-ci a changé entre-temps. Enfin, n'importe quel utilisateur peut comprendre les données des autres et utiliser leurs analyseurs sur ses corpus, sans avoir à changer d'environnement. En résumé, le principe d'auto-description d'IrcamCorpusTools lui assure la pérennité des données, lui fournit un langage de requête expressif et lui confère la possibilité de mutualiser

les données, les fichiers, les corpus et surtout, les analyseurs. La mise en commun des outils est un facteur déterminant pour le développement des recherches en TAL et en TAP, car leur complexité s'accroît rapidement.

A.5.9 Exemple d'utilisation

Nous donnons à présent un exemple montrant comment IrcamCorpusTools peut être utilisé. La figure A.2 donne un exemple schématique d'une instance particulière de quelques objets permettant d'accéder à la moyenne de la fréquence fondamentale correspondant au phone /a/ de la 678ème phrase d'un corpus appelé Ferdinand2007.

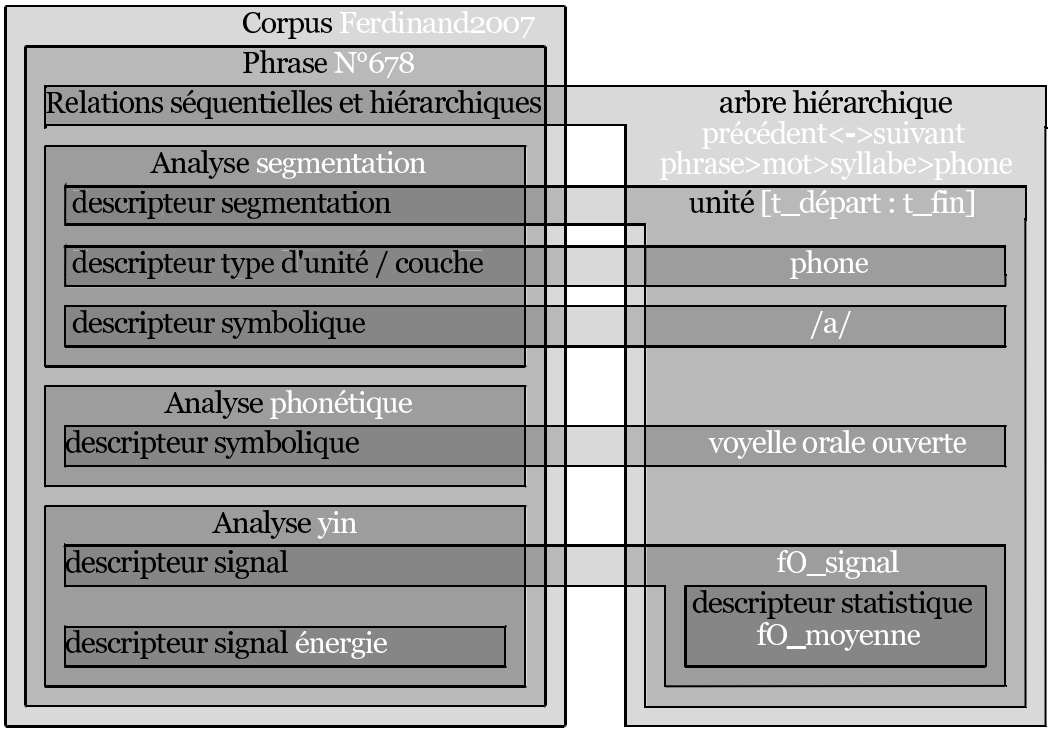

FIG. A.2: Exemple d'utilisation : une instance particulière.

Cette phrase est décrite par trois analyses : une segmentation en phones, une analyse phonétique, et une analyse acoustique effectuée par l'algorithme « yin » [Cheveigné 2002] qui fournit deux descripteurs continus « f_0 » (fréquence fondamentale) et « énergie » (énergie à court terme) évoluant le long de la phrase. L'analyse « segmentation » permet de définir les unités de la phrase. Une unité est décrite par son temps de début (« t_départ ») et son temps de fin (« t_fin »), son appartenance au niveau « phone » et son étiquette correspondante « /a/ ». Sa catégorie phonétique donnée par l'analyse « phonétique » est ici « voyelle orale ouverte ». L'horizon temporel couvert par cette unité permet d'accéder aux portions de signaux correspondantes (« f_0_signal » pour la fréquence fondamentale). De plus, il permet la mesure statistique de cette portion de fréquence fondamentale

« f_0_moyenne » sur l'unité (valeur caractéristique).

Cette opération se réalise simplement dans notre langage par les quelques lignes suivantes :
Tout d'abord, on charge en mémoire, un corpus :

```
>> corpus = loadcorpus("Ferdinand2007");
```

Puis on instancie les objets descripteurs de la fréquence fondamentale et du phone (par la même syntaxe alors que ce sont des types différents) :

```
>> f0 = loadfeatures(corpus, 678, "f0");
>> phones = loadfeatures(corpus, 678, "phone");
```

Enfin, on chaîne la segmentation et le calcul de la moyenne :

```
>> f0_seg = segment(f0, phones);
>> f0_mean = mean(f0_seg);
```

A cette étape, la variable résultante « f0_mean » est un tableau de plusieurs objets contenant chacun, la moyenne de la fréquence fondamentale d'un des phones de la segmentation. Afin d'accéder à la moyenne d'intérêt, c'est-à-dire à celle correspondant au phone /a/, il nous faut filtrer ces objets à partir de considérations linguistiques. Nous avons donc besoin du langage de requête.

```
>> phone_a = getunits(corpus,678,"phone",{"phoneme","is","a"});
```

Cette requête donne accès à l'ensemble des phones /a/, présents dans la phrase. La même requête sans le numéro de la phrase donne accès à tous les phones /a/ du corpus.

```
>> phones_a = getunits(corpus, "phone", {"label","is","a"});
```

A présent, nous allons enrichir cette requête de manière à étudier un phénomène de coarticulation : une voyelle /a/ précédée de la plosive /p/ et suivie de la fricative /f/. Pour cela, nous réduisons le faisceau des unités sélectionnées à partir de contraintes contextuelles, ce qui se réalise simplement en spécifiant les contraintes en chaîne dans la requête.

```
>> phones_paf = getunits(corpus, "phone",
                {"label","is","a"},
                {"prev_phone_label","is","p"},
                {"next_phone_label","is","f"});
```

Pour observer si l'effet de coarticulation partage des propriétés acoustiques similaires lorsque les contextes présentent des similarités phonétiques, nous allons maintenant élargir le contexte gauche aux plosives et le contexte droit aux fricatives. Cette requête peut se formuler de deux manières :

1) En spécifiant manuellement des listes de phonèmes :

```
>> phones_PaF = getunits(corpus, "phone", {"label","is","a"},
                {"prev_phone_label","is",{"p","t","k"}},
                {"next_phone_label","is",{"f","s","S"}});
```

2) Ou alors en faisant une requête directement sur la classe phonétique des phonèmes considérés :

```
>> phones_PaF = getunits(corpus, "phone", {"label","is","a"},
                {"prev_phone_class","is","P"},
                {"next_phone_class","is","F"});
```

Par ailleurs, la représentation des données dans IrcamCorpusTools permet la description des observations sur plusieurs niveaux. Nous pouvons donc accéder, pour un niveau donné (ici, le phonème), à des informations d'un niveau parent (par exemple : la syllabe). Ce faisant, nous rajoutons à présent sur la requête précédente, le désir d'observer un phonème /a/ de même contexte mais faisant partie d'une syllabe proéminente.

```
>> phones_PaF_P = getunits(corpus, "phone", {"label","is","a"},
                  {"prev_phone_class","is","P"},
                  {"next_phone_class","is","F"},
                  {"syllable_phone_proeminence","is","P"});
```

Enfin, une particularité de notre langage, qui le rend propice aux applications TAP, est l'introduction de contraintes sur les données de type signal dans la composition des requêtes. Ainsi, l'adjonction de la contrainte {"f0_mean","$>$","200"} dans la requête précédente, permet d'écarter les unités dont la fréquence fondamentale est inférieure à 200 Hz :

```
>> phones_PaF_f0 = getunits(corpus, "phone", {"label","is","a"},
                   {"prev_phone_class","is","P"},
                   {"next_phone_class","is","F"},
                   {"f0_mean",">","200"});
```

A.6 Conclusion

Dans cette partie, nous avons présenté IrcamCorpusTools, une plateforme extensible pour la création, la gestion et l'exploitation des corpus de parole. Elle permet facilement d'interfacer des données hétérogènes avec des analyseurs internes ou externes, en utilisant le principe d'auto-description des données et des analyseurs. En outre, l'auto-description des données garantit leur pérennité, favorise l'introduction de nouveaux types et leur confère une plus grande visibilité. De même, l'auto-description des analyseurs assure l'extensibilité de la plateforme ainsi que sa modularité et la mutualisation des corpus. La plateforme IrcamCorpusTools est capable de gérer les relations hiérarchiques multiples et séquentielles entre des unités. Un langage de requêtes simple et expressif donne un accès immédiat aux données de ces unités. Ces fonctionnalités appliquées à différents corpus de parole (parole contrôlée, parole spontanée pour des études de la prosodie de l'expressivité) intéressent directement les recherches à la frontière entre le Traitement Automatique des Langues et le Traitement Automatique de la Parole. L'intégration de ces exploitations énumérées au sein d'une même plateforme, illustre les avantages de l'inter-opérabilité. Ceci doit être interprété comme un encouragement au partage

des outils entre les communautés TAL et TAP. Cette plateforme IRCAM permet la gestion et l'usage des corpus expressifs pour la transformation de l'expressivité dépendante du contexte.

Annexe B

Annexe : Sur le débit de parole

Sommaire

B.1 Mesure du débit intra-syllabique

Une mesure du débit dynamique a été présentée dans la partie 3.2.4. La définition de cette mesure repose sur l'unité temporelle de type syllabe. Même si on peut interpoler ces valeurs de manière à obtenir une variation continue, et par conséquent une évolution au sein de la syllabe, il ne s'agit pas d'une mesure du débit intra-syllabique. Or il semble qu'une variation de débit peut être perçue au sein d'une seule syllabe (par exemple pour les mots outils mono-syllabiques comme "oui"/"non"). C'est pourquoi nous proposons maintenant une mesure du débit intra-syllabique. Cette mesure repose sur l'alignement dynamique temporel de syllabes possédant la même séquence phonétique. L'alignement dynamique temporel ("Dynamic Time Warping") est une méthode permettant d'aligner les trames temporelles de deux séquences. Un algorithme de Viterbi décode le chemin le plus probable sur une matrice de distance construite à partir des MFCC des deux signaux. La figure B.1 présente deux phrases et leur alignement temporel. La mesure proposée repose sur un ensemble de syllabes dont le contexte phonétique est commun (ce qui demande une grande base de donnée, qui peut être réduite en considérant le contexte phonologique au lieu du contexte phonétique). Pour chaque classe de syllabes, un profil de déroulement temporel peut être obtenu en moyennant les chemins provenant des multiples alignements par paires des syllabes. Ensuite, le déroulement temporel de chacune d'entre elles peut être modélisé comme la déviation de son chemin par rapport à ce chemin moyen (au sens statistique du terme).

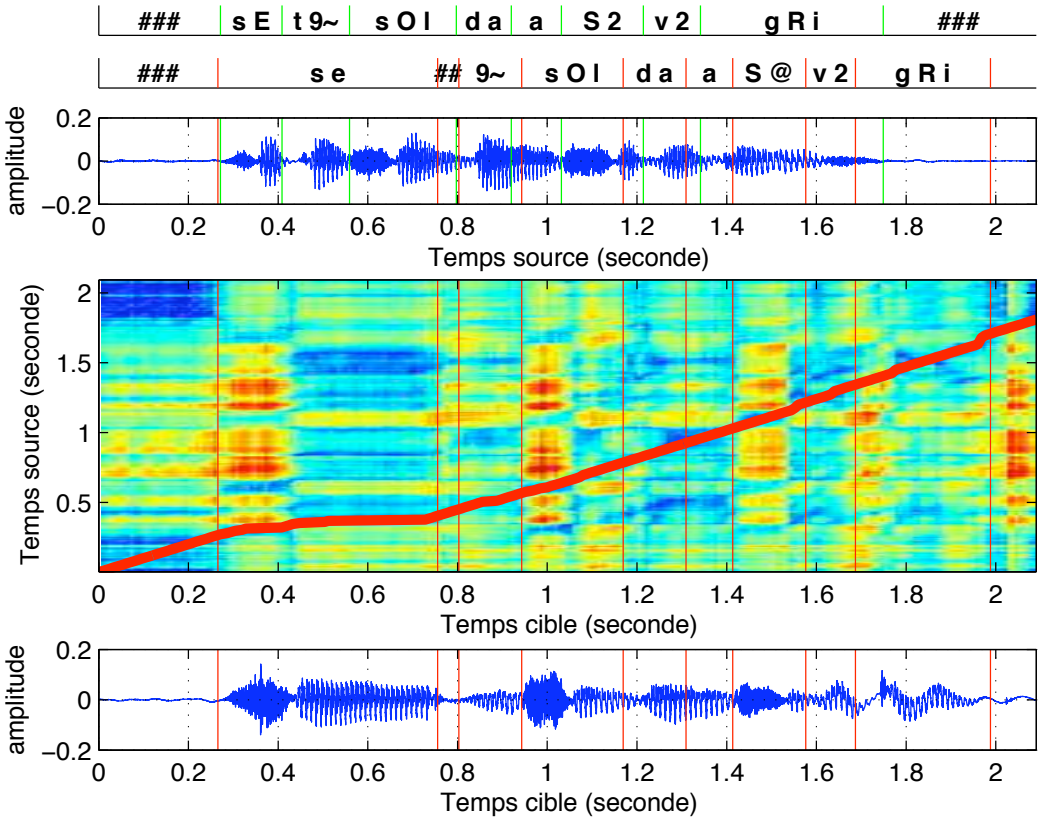

FIG. B.1: Exemple de visualisation du débit syllabique.

B.2 Métricité de la parole

Une étude originale consiste en l'analyse de la *métricité* de la parole. L'Eurhythmie qui vient du grec "Eurhythmos", qui signifie "bon rythme" a pour étude le rythme de la parole. Nous avons cherché à mesurer l'isochronie de la parole, appelée métricité. Compte tenu de l'importance de la syllabe dans la perception du rythme, nous avons construit des patrons rythmiques pour chaque phrase, à partir de leurs segmentations syllabiques et de l'annotation du degré de proéminence. Puis nous avons convolué chacun des patrons avec plusieurs trains d'impulsions isochrones possédant des tempi croissants. A chaque tempo, le meilleur délai initial entre le train d'impulsions et le patron rythmique est aussi trouvé par multiples convolutions. La figure B.2 présente les patrons rythmiques de deux phrases (tristesse introvertie et colère extravertie), ainsi que les meilleurs trains d'impulsions possédant une convulotion maximale, dans chaque cas. Ce score de convolution est représenté en dessous selon les tempi croissants. Le train d'impulsions isochrone possédant la plus forte convolution est retenu pour la mesure de la métricité de la phrase.

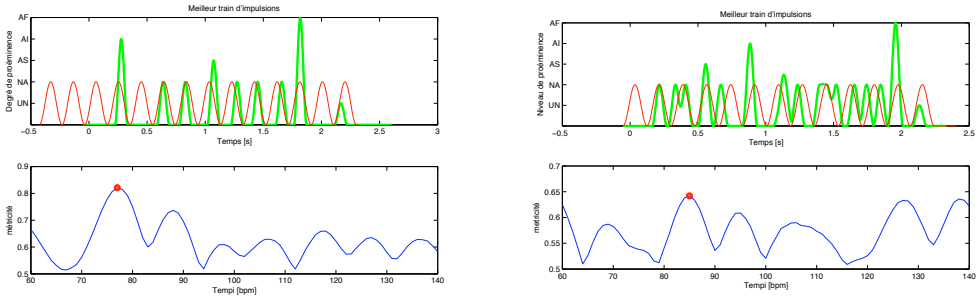

FIG. B.2: En haut, meilleur train d'impulsions isochrone permettant de modéliser le patron rythmique dérivant de l'annotation de la proéminence des syllabes. En bas, représentation de la métricité en fonction du tempo du train d'impulsions isochrone. A gauche et à droite, phrases du corpus Combe2006 possédant respectivement la plus forte et la plus faible métricité, d'expressions tristesse introvertie et colère extravertie.

Une fois la métricité de chaque phrase estimée, on verifie que les tempi correspondants sont bien corrélés au débit de parole. Puis on estime les distributions de la métricité pour chaque expression. La figure B.3 présente ces résultats pour la réunion des corpus Combe2006 et Roullier2006. Les surprises et le dégoût semblent posséder des rythme plus isochrones que les autres expressions, tandis que l'excitation et la colère extravertie semblent moins régulières dans leurs débits.

Ces résultats sont purement spéculatifs car avant d'être exposée à la parole expressive, cette mesure nécessite une application contrôlée à la parole et une validation par confrontation à des données humaines de "tapping". Enfin, si ils ne paraissent pas très significatifs, c'est aussi parce que les décélérations initiales et

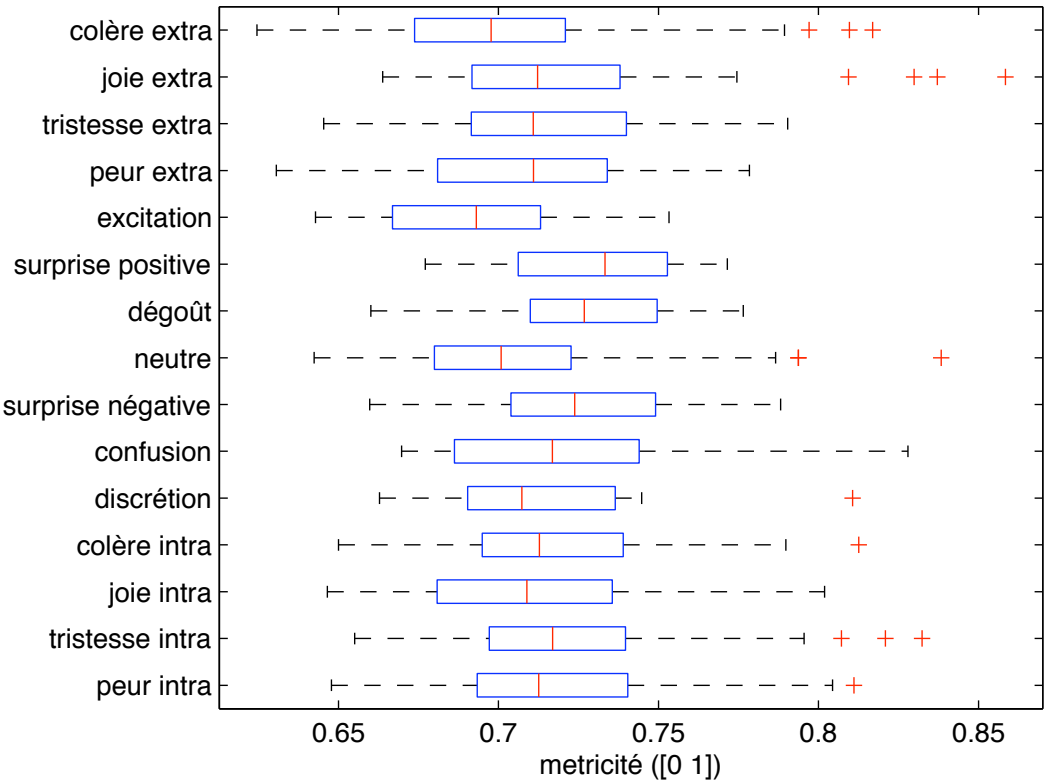

FIG. B.3: Acteurs masculins. "Boxplots" de la métricité des phrases, classées par expression.

accélérations finales ne sont pas prises en compte. Ainsi, si cette mesure ouvre des pistes de recherche, elle demande plus de considérations pour être utile pour cette étude (voir chapitre 5).

Annexe C

Annexe : Traitement de signaux EGG

Sommaire

De manière à mesurer la qualité vocale, des signaux Electro Glotto Graphique (EGG) ont été utilisé. Dans un premier temps, ces signaux sont pré-traités. Ensuite, un algorithme permet de déduire de ces signaux EGG, les instants de fermeture GCI et d'ouverture de la glotte GOI, ainsi que la période fondamentale T0 et le quotient ouvert Oq. Enfin, un post-processing de ces signaux EGG permet de les synchroniser avec l'audio.

C.1 Pré-processing des signaux EGG

Le signal dEGG peut être associé à une image du signal d'excitation glottique dans la phase de fermeture [Henrich 2001]. La dEGG est obtenue par simple dérivation temporelle du signal EGG :

$$dEGG(n) = \frac{EGG(n+1) - EGG(n)}{Te} \tag{C.1}$$

Ce type de dérivation suffit puisque le spectre du signal EGG possède une fréquence de coupure nettement inférieure à la moitié de la fréquence d'échantillonnage (44.1KHz). Nul n'est donc besoin d'employer une transformation bilinéaire ou de filtrer le signal EGG au préalable.

C.2 Marquage des signaux EGG

Le signal EGG et sa dérivée dEGG permettent d'accéder aux instants de fermeture (GCI pour Glottal Closing Instant) et aux instants d'ouverture (GOI pour Glottal Opening Instant) de la glotte. Le rapport entre la durée d'ouverture de la glotte et la période fondamentale (durée entre deux GCI) correspond au quotient ouvert (Oq pour Open Quotient).

C.2.1 Détection de l'activité glottique

Plusieurs algorithmes de détection de l'activité glottique reposent sur l'observation de la dérivée de l'EGG. En effet, la dEGG présente aux instants de fermeture de la glotte, un pic proéminent facilement détectable par seuillage [Gendrot 2004]. Dans le cadre de la base de données emoDB, la variété intra et inter locuteurs ne permet pas cette approche car le RSB (Rapport Signal sur Bruit) des enregistrements EGG n'est pas constant. La plupart du temps, le bruit est blanc et donc de distribution gaussienne $\mathcal{N}(0, \sigma_k)$. La distribution du signal dEGG n'est pas gaussienne et de moyenne non nulle à cause d'une dissymétrie engendrée par les pics aux instants de fermeture.

C.2.2 Mesure du Gauchissement

Nous utilisons cette information afin de détecter les zones d'activité de la glotte et donc, les zones d'estimation des GCI/GOI. Les indicateurs appelés "Gauchissement" (skewness) et "Aplatissement" (kurtosis) donnent des indications sur le

caractère non gaussien d'une distribution [Kendall 1969]. Le Gauchissement d'une variable aléatoire X de dimension N est relative au moment statistique du 3ème ordre :

$$Gauchissement(X) = \frac{E[(X - E[X])^3]}{\sigma_X^3} \tag{C.2}$$

Avec l'espérance définie classiquement par :

$$E_N[X] = \sum_{i=1}^{N} X_i P[X = X_i] \tag{C.3}$$

avec $P[X = X_i \mid \forall i \in [1 : N]] = \frac{1}{N}$ dans notre cas puisque nous ne possédons aucune connaissance a priori sur la distribution de la dEGG si ce n'est qu'elle n'est pas gaussienne.

Et la variance :

$$\sigma_N[X] = \sqrt{E_N[(X - E_N[X])^2]} \tag{C.4}$$

Le Gauchissement mesure le caractère asymétrique d'une loi de part et d'autre de sa valeur moyenne. Pour une loi gaussienne, le Gauchissement doit se rapprocher de zéro. Le Gauchissement calculé sur un ensemble d'échantillons du signal dEGG nous permet de mesurer l'éloignement de la distribution des points à une distribution gaussienne. Considérons le signal dEGG stationnaire sur un horizon temporel d'une durée environ deux fois plus grande que la plus longue période fondamentale de la phrase (N_0 points). On mesure sur chaque trame consécutive assez proche, la moyenne, la variance et le Gauchissement sur une fenêtre glissante de taille N_0 points. Chaque trame possédant une valeur de Gauchissement supérieure à un seuil arbitraire est considérée comme présentant une activité glottique.

C.2.3 Robustesse vis à vis du bruit

Le rapport signal sur bruit (noté RSB ou encore SNR pour Signal to Noise Ratio) chiffre le rapport entre le signal utile et le bruit. Ce rapport est généralement exprimé en décibels par :

$$RSB = 20 * log(\frac{energy(signal)}{energy(bruit)}) \tag{C.5}$$

avec

$$energy(x) = \sum x[n]^2 \tag{C.6}$$

Dans ce contexte, le signal est la dEGG et le bruit est un bruit blanc gaussien centré de distribution $\mathcal{N}(0, \sigma_k)$. Nous avons fait varier le RSB de manière artificielle en ajoutant un tel bruit au signal dEGG dans des rapports d'énergie divers. Ceci dans le but de montrer que la mesure du Gauchissement nous permet d'accéder aux

zones d'activité glottique de manière plus fiable que la mesure directe de l'énergie de la dEGG en présence de bruit.

La figure C.1 représente le signal dEGG, et les mesures du Gauchissement et de l'énergie de la dEGG pour différents RSB (de 20 dB à 60 dB par pas de 10 dB). On observe que pour de faibles RSB, une comparaison entre un seuil fixe et l'énergie ne nous permet pas d'identifier les zones d'activité glottique à cause de l'énergie du bruit. Tandis qu'une comparaison entre le Gauchissement et un seuil fixe et arbitraire souffre beaucoup moins de la variation du RSB. Ce seuil n'étant ni dépendant de l'énergie du signal dEGG, ni du RSB, nous choisissons d'utiliser la mesure du Gauchissement pour définir les zones d'activité glottique.

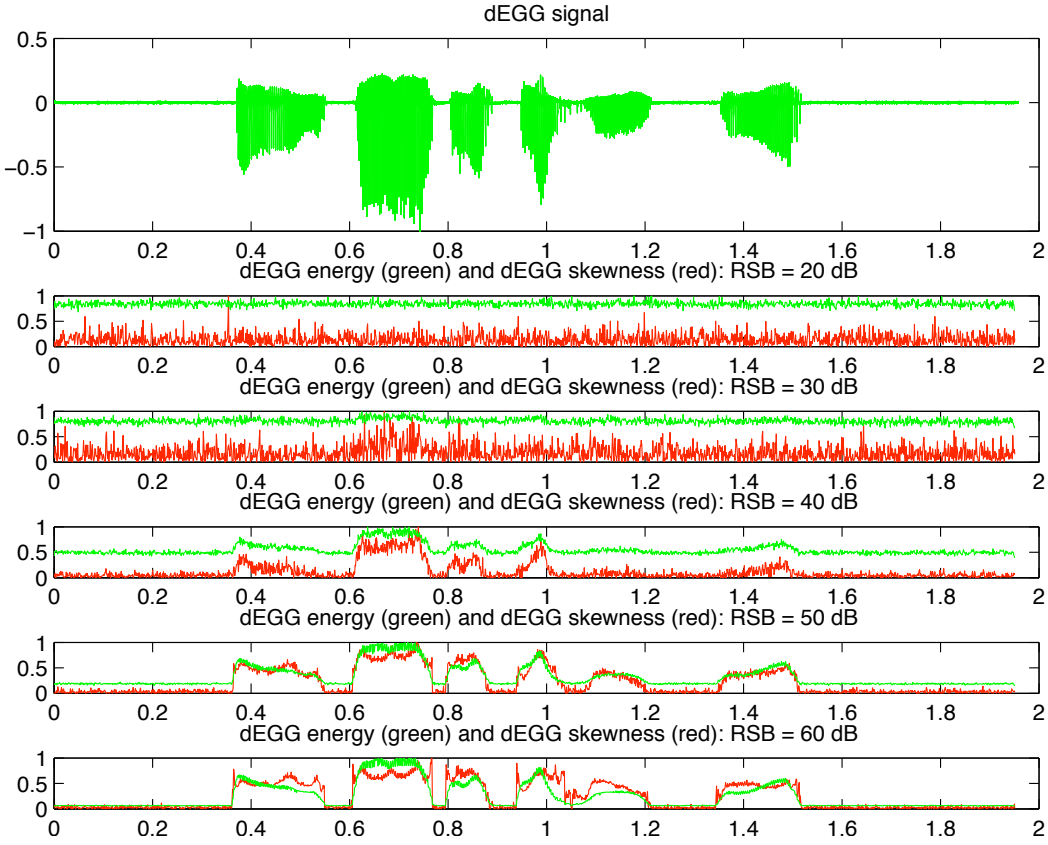

FIG. C.1: Mesure de détection de l'activité glottique robuste au RSB.

C.3 Marquage des instants de fermeture de la glotte : GCI

Dans chaque zone d'activité glottique, on cherche à déterminer les instants de fermeture de la glotte. Pour cela, nous cherchons les maxima locaux du signal dEGG correspondant aux changements abrupts du signal EGG. La même fenêtre glissante

d'un horizon temporel d'environ deux périodes fondamentales est déroulée sur le signal dEGG. On estime pour chacune d'elle la position du maximum local. Puis on fait un histogramme de ces positions, un seuil nous permet de ne garder que les maxima correspondants aux pics proéminents du signal et donc au GCI. La figure C.2 montre ce procédé. Ceux dont la fréquence de répétition est supérieur à un seuil dépendant d'un rapport entre la période fondamentale et le pas d'avancement de la fenêtre sont considérés comme GCI.

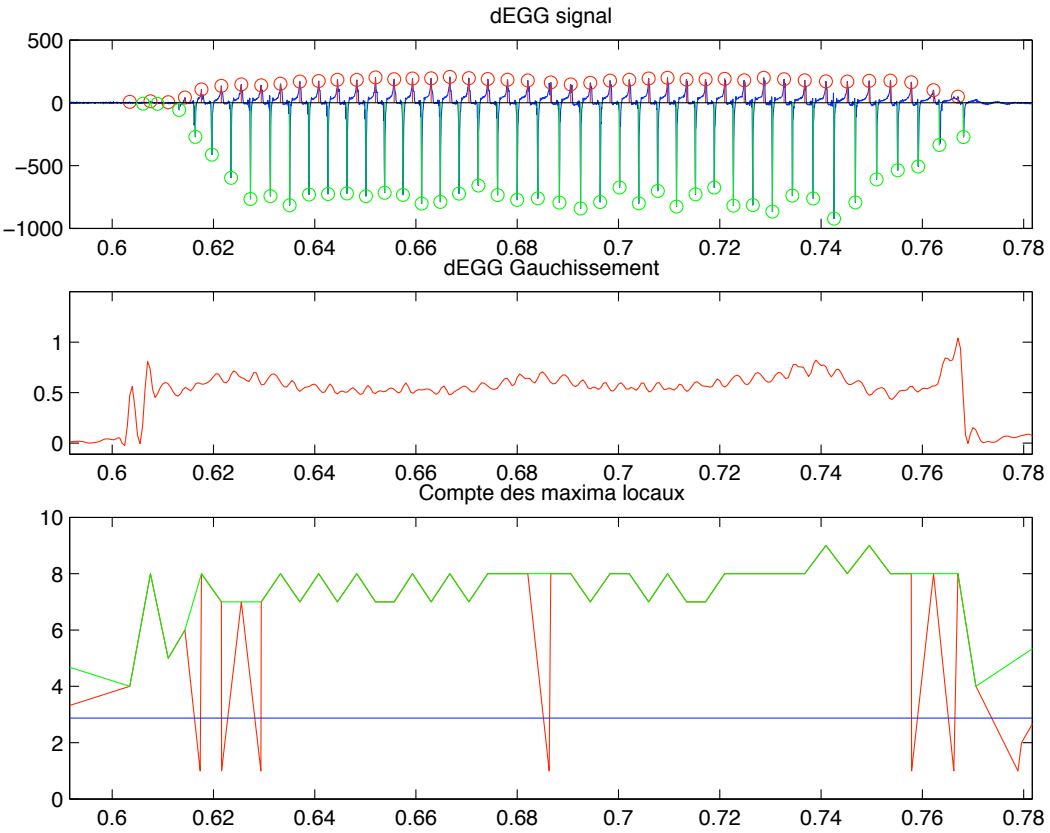

FIG. C.2: Signal dEGG marqué, gauchissement local du signal dEGG, compte des maxima locaux (ceux en vert sont les maxima locaux correspondant aux GCI).

C.4 Marquage des instants d'ouverture de la glotte : GOI

Une fois les instants de fermeture de la glotte GCI détectés, nous possédons un ensemble de périodes fondamentales consécutives ou non dans lesquelles nous cherchons les instants d'ouverture de la glotte GOI. Les GOI sont détectables de la même manière que les GCI puisqu'ils correspondent aussi à une forte variation de l'EGG. Ainsi on recherche dans chacune des périodes fondamentales marquées, un maximum local de la valeur absolue du signal dEGG. Afin d'éviter les valeurs

aberrantes (un GCI présente parfois deux pics proéminents proches dans le signal dEGG), nous restreignons la zone de recherche du GOI : nous interdisons qu'un GOI soit placé à côté d'un GCI dans un rayon de ±10% de la période fondamentale locale. Ce choix contraint le quotient ouvert Oq à prendre une valeur entre [0.1 : 0.9].

C.5 Mesure de la période fondamentale locale : T0

Elle est directement déduite du marquage des instants GCI, puisqu'elle correspond à la durée entre deux instants de fermeture de la glotte consécutifs.

$$T0(T) = GCI(T+1) - GCI(T) \tag{C.7}$$

On peut visualiser la différence entre la fréquence fondamentale déduite de cette mesure et le résultat lissé fournie par l'algorithme yin dans la figure C.3.

C.6 Mesure du Quotient Ouvert : Oq

Cette mesure est calculée à partir des marques GCI-GOI sur chaque période fondamentale. En effet, une valeur du Quotient Ouvert Oq est le rapport entre la durée d'ouverture et la période fondamentale :

$$OQ(T) = \frac{T0(T) - (GOI(T) - GCI(T))}{T0(T)} \tag{C.8}$$

Un exemple de son évolution au cours d'une phrase est donné dans la figure C.3.

C.7 Post-processing des signaux EGG

L'enregistrement des signaux EGG et audio est simultané. Le plus souvent, ces deux signaux sont stockés dans un fichier stéréo dans lequel chaque signal occupe un canal. Cependant, cela ne veut pas pour autant dire qu'ils soient synchronisés. Dans le modèle source-filtre de la parole, le signal EGG est lié à l'activité de la source glottique. Cette source, filtrée par le conduit vocal et rayonnée aux lèvres, produit le son de parole voisé. La vibration acoustique est captée par un microphone placé à une distance inférieure à un mètre des lèvres. Le transfert aérien du signal acoustique provoque un délai de l'ordre de la milliseconde (2.9 ms par mètre) entre le signal audio capté et le signal EGG pris à la source. C'est pourquoi il est nécessaire de retrouver la correspondance temporelle entre le signal EGG et le signal d'excitation filtré. De plus, les variétés des enregistrements intra et inter locuteurs obligent à une synchronisation automatique des signaux. Il est considéré que la distance entre les lèvres et le micro durant la prononciation d'une phrase reste constante. Ainsi on cherche un délai global sur toute une phrase. On spécifie un délai maximal : $d'elai_{max} \sim 1000ms$. Puis on effectue deux types de corrélation sur l'horizon temporel correspondant $[-d'elai_{max} : +d'elai_{max}]$. La première corrélation étudiée

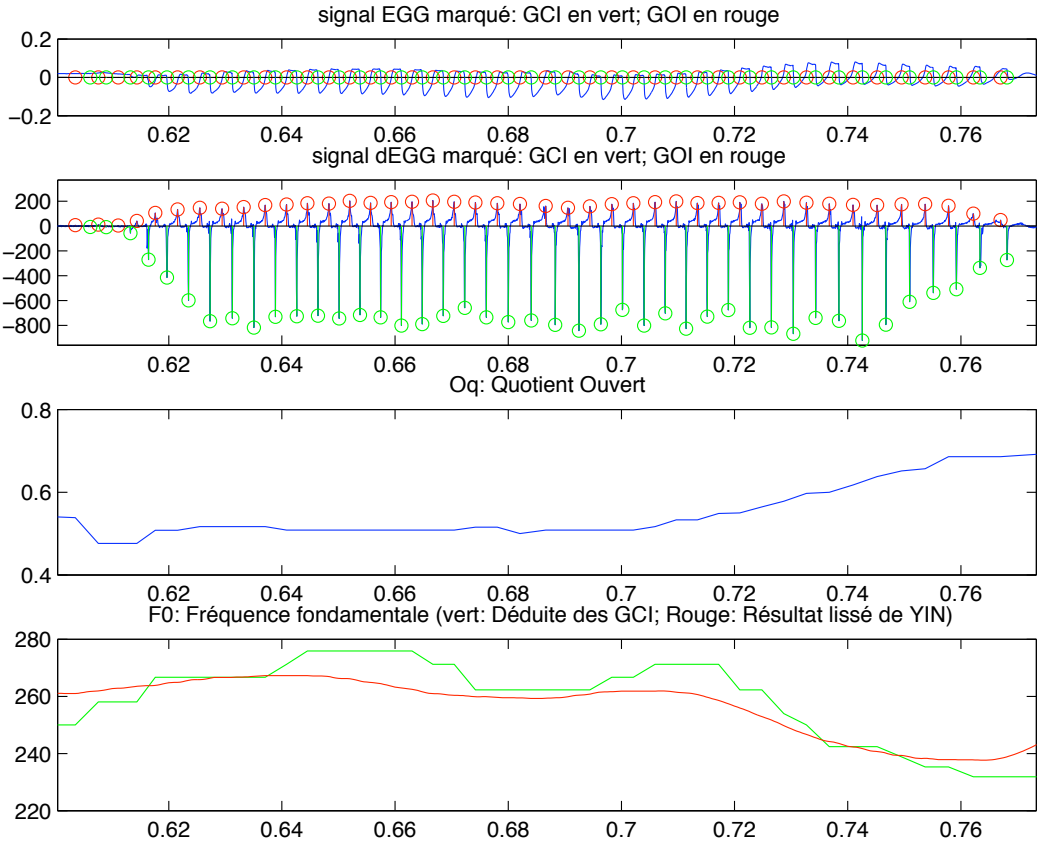

FIG. C.3: Signal EGG marqué par les GCI et les GOI ; signal dEGG marqué par les GCI et les GOI ; Oq : Quotient ouvert ; Deux estimations de la fréquence fondamentale F0.

utilise les marqueurs PSOLA et les GCI (voir section C.2). La seconde corrélation met directement en jeu les signaux dEGG et audio.

C.7.1 Corrélation entre les marqueurs PSOLA et les GCI

La méthode PSOLA (Pitch Synchronous OverLapp and Add) repose sur une estimation directe des GCI à partir du signal audio. Nous avons utilisé une méthode PSOLA minimisant une norme de Fröbenius [Peeters 2001] pour déterminer les GCI à partir du signal audio. Puis nous calculons la corrélation entre ces marques et les GCI calculés précédemment à partir du signal EGG. L'indice temporel du maximum de la valeur absolue de la fonction de corrélation correspond au délai entre le signal audio et le signal dEGG recherché. La figure C.4 présente l'histogramme des 816 délais EGG/audio obtenus par cette méthode. Le délai moyen ainsi obtenu est de l'ordre de 1.688 ms (variance de 0.147 ms) ce qui correspond à une distance d'environ 57 cm. La longueur moyenne du conduit vocal étant de 17 cm [Fritz 2004], nous en déduisons que le microphone a été placé à proximité de la bouche des locuteurs ($\sim$ 40 cm).

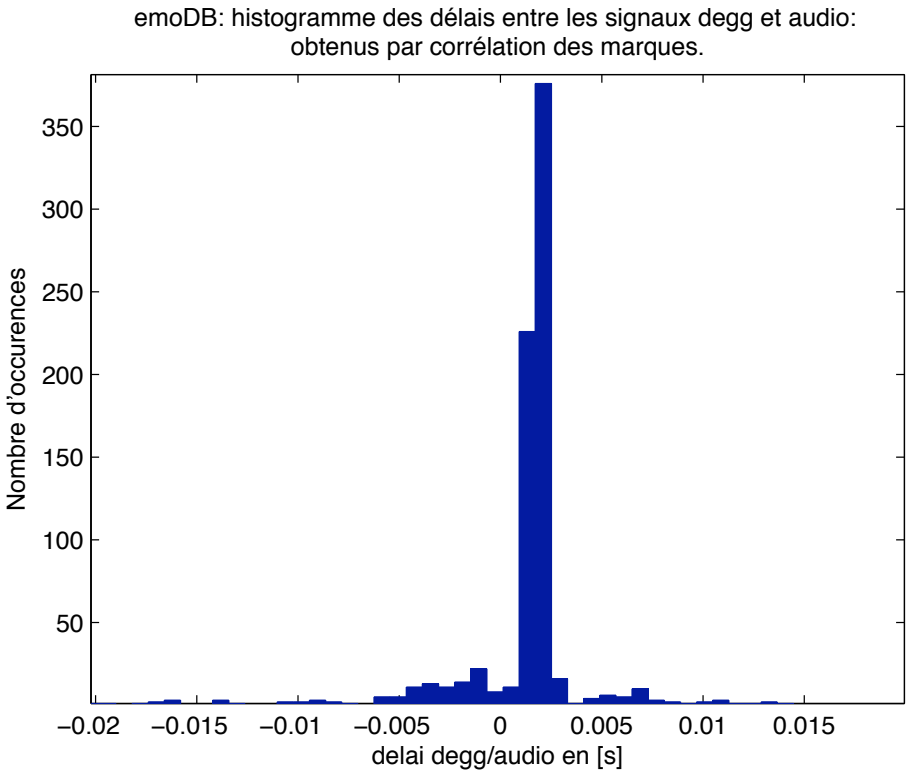

FIG. C.4: histogramme des délais entre les signaux dEGG et audio : obtenus par corrélation des marqueurs GCI et des marqueurs PSOLA.

C.7.2 Corrélation entre les signaux dEGG et résiduel

Le signal dEGG peut être associé à une image du signal d'excitation glottique dans la phase de fermeture [Henrich 2001]. Le signal résiduel est obtenu grâce à un filtrage inverse destiné à produire un signal résiduel à phase minimale [Roebel 2005b, Villavicencio 2006]. Un exemple de signal résiduel est donnée dans la figure C.5.

Puis on effectue la corrélation du signal audio et du signal dEGG sur l'horizon temporel correspondant [- délai-max : + délai-max]. L'indice temporel du maximum de la valeur absolue de la fonction de corrélation correspond au délai entre le signal audio et le signal dEGG recherché.

Le délai moyen ainsi obtenu est de l'ordre de 1.843 ms (variance de 0.023 ms) ce qui correspond à une distance d'environ 46 cm.

On observe sur la figure C.7, la distribution des délais obtenus par cette méthode.

C.7.3 Corrélation entre les signaux EGG et audio

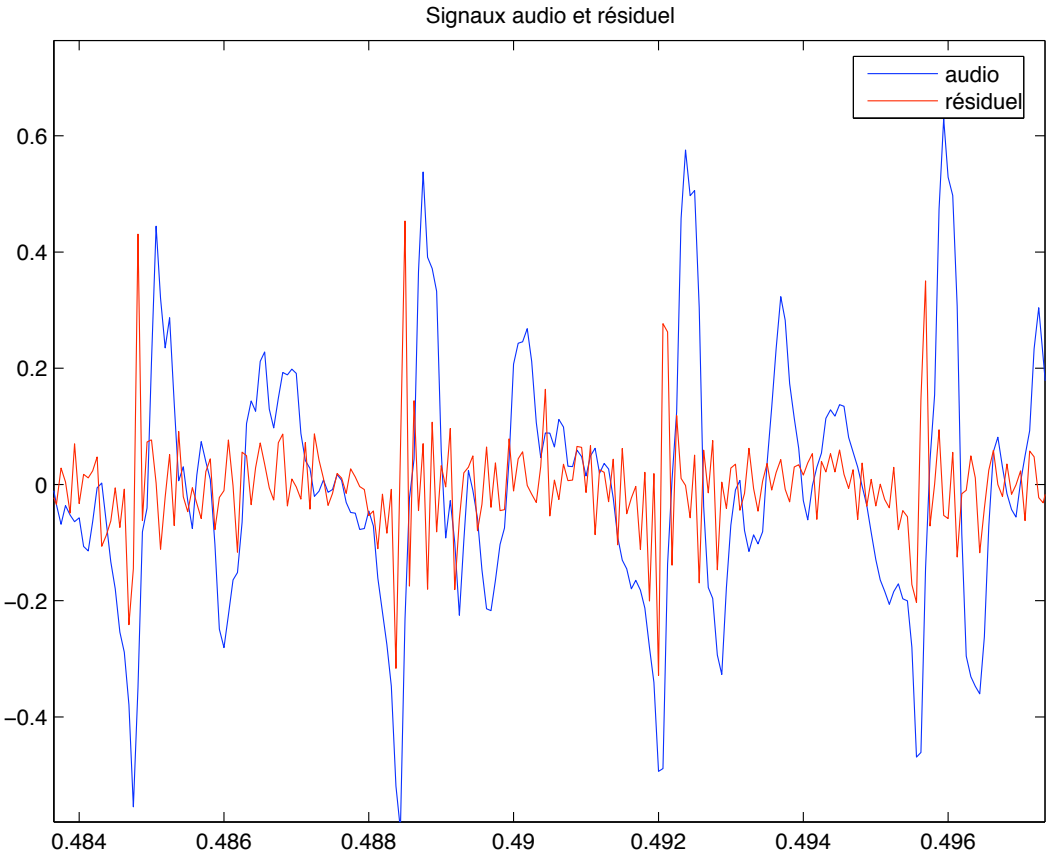

FIG. C.5: Exemple de signal résiduel obtenu par filtrage inverse du signal audio.

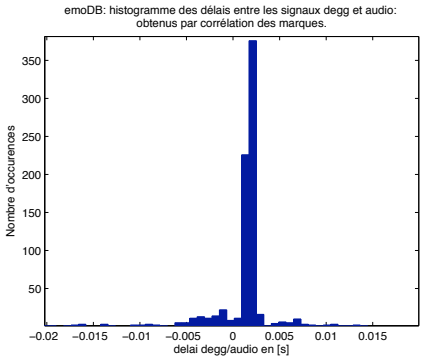

FIG. C.6: histogramme des délais obtenus par corrélation des marqueurs GCI et des marqueurs PSOLA.

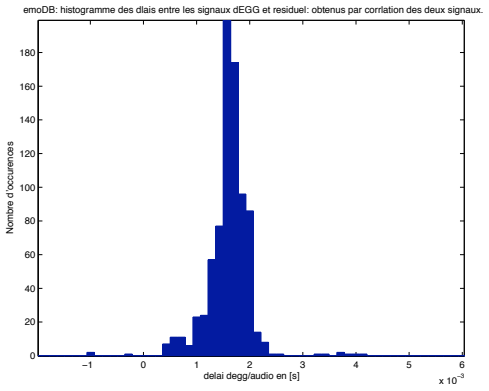

FIG. C.7: histogramme des délais obtenus par corrélation des signaux dEGG et résiduel.

Les deux méthodes présentées précédemment donnent un délai moyen d'environ 1.750 ms qui correspond à une distance moyenne entre la bouche d'un locuteur et le microphone d'environ 50 cm. Cette distance concorde avec les enregistrements dans le sens où les locuteurs étaient placés à environ 50 cm du microphone, comme le montre la photo de la figure C.8.

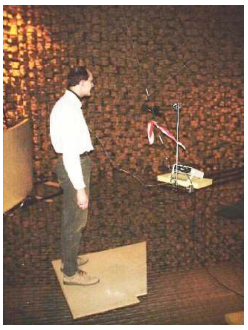

FIG. C.8: Photo prise durant les enregistrements de la base emodb, tirée de http://database.syntheticspeech.de/

Annexe D

Annexe : Synthèse semi-paramétrique de rires

Sommaire

D.1 Résumé du chapitre

This paper describes a semi-parametric speaker-like laughter synthesis method. A large corpus of spontaneous laughter is presented. An attempt to use traditional automatic segmentation on the data is discussed. Significant results from the statistical analysis of the corpus are then presented, with concern to the static and dynamic acoustic characterizations of bouts and syllables. Interestingly, laughter prosody seems to be guided by the same physiological constraints as verbal speech. After this analysis part, a method for synthesizing laughter from any neutral utterance using information from the previous results is described. A TTS algorithm selects some phones that are duplicated to create a homotype series. Finally, speech processings modify the prosody of this series, providing a realistic high quality speaker-like bout of laughter

D.2 Introduction

A database management system for speech is being constructed to allow the manipulation of large corpora for various artistic objectives [Veaux 2008]. One of our objectives is high quality expressive Text-To-Speech synthesis. This objective is divided into two parts : high quality neutral TTS synthesis and high quality expressive speech transformation. The latter part requires statistical context-dependent analysis of prosodic parameters according to expressivity [Beller 2006b] [Beller 2007a]. This is achieved by para-linguistic speech manipulations such as articulation degree modification [Beller 2008a]. This modification can then be applied to either synthesized or spoken speech. It is currently being used by film and theater directors and also in dubbing studios. It is within this outline that the study is conducted.

The analysis of a very large corpus of naturally-occurring conversational speech [Campbell 2005] reveals that approximately one in ten utterances contain laughter. Therefore laughter is a powerful means of emotion expression which is beginning to be analyzed and used in speech synthesis [Schroeder 2004]. Acoustic studies on spontaneous [Campbell 2005], semi-spontaneous [Bachorowski 2001] and simulated corpora exhibit interesting acoustic features, tendencies, and variabilities. Despite the youth of this new research topic, the need of a common terminology for laughter description has already been adressed and partly solved [Trouvain 2003]. In order to compare this study to others, we refer to the terminology of Trouvain et al. for the definition of the terms used in this paper.

Although the acoustic of laughter is highly variable, some regularities can be observed with regard to its temporal structure. Laughter bouts are typically initiated by one or two singular elements (i.e. non-repeated, with large variability in acoustic parameters). These are often followed by a succession of *syllables* with predictable similarity, i.e. a homotype series [Kipper 2007]. The overall temporal behavior can be captured by a parametric model based on the equations that govern the simple harmonic motion of a mass-spring system [Sundaram 2007].

Our main goal is to apply a desired expressivity to a spoken or synthesised neutral utterance. In the case of happiness, adding speaker-specific laughter as a para-verbal burst to the transformed utterance makes the result more likely to be perceived as the intended expressivity. Unfortunately, no laughter is present in the neutral utterance and one must synthesize it taking in to account only the verbal content of the utterance. This paper explains a semi parametric method that is able to provide speaker-like laughter from a neutral utterance using a corpus based analysis of the dynamics of laughter.

D.3 General Overview

First, a large corpus of spontaneous laughter [Campbell 2005] is presented. An attempt to use traditional automatic segmentation of the data is discussed. Significant results from the statistical analysis of the corpus are then presented, with concern to the static and dynamic acoustic characterizations of bouts and syllables. After this analysis part, a method for synthesizing laughter from a new neutral utterance using information from the previous result is presented. Then, a rule-based selection algorithm picks up some phones that are duplicated to create a homotype series. Finally, speech processing modify the prosody of this series, providing a realistic high quality bout of laughter.

D.4 Corpus

The data came from a large corpus of spontaneous Japanese conversational speech [Campbell 2007b]. Two sets of laughter bouts have been extracted : one of a male speaker JMA and one of a female speaker JFA. Corpora consist of 1150 bouts of JMA and 953 bouts of JFA recorded with a head-mounted Sennheiser HMD-410 close-talking dynamic microphone and using DAT (digital audio tape) at a sampling rate of 48kHz.

D.5 Analysis part

D.5.1 Automatic Segmentation

The bouts were automatically labeled by an "unsupervised" automatic HMM-based segmentation system [Morris 2006] trained on a neutral multi-speaker database [Lamel 1991]. A first attempt presented bad segmentation results. This was because of the lack of breathing in the training corpus, a lot of devoiced vowels and breathing had been tagged as voiced fricatives or liquids, as shown by figure D.2. In order to circumvent this problem, we removed the corresponding models of voiced consonants leading to a supervised automatic HMM-based segmentation. Although breathing was wrongly marked as occlusive, vowels seemed have a better response to the automatic segmentation. Results of supervised automatic segmentation led

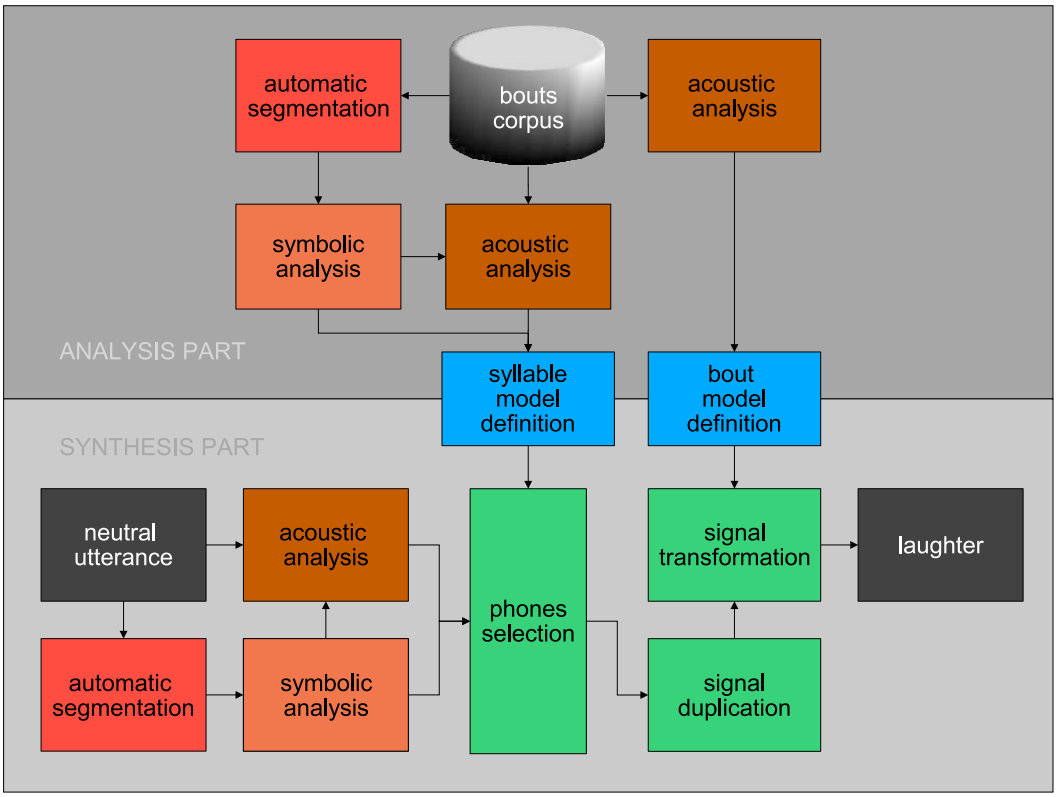

FIG. D.1: Overview of the semi-parametric synthesis method.

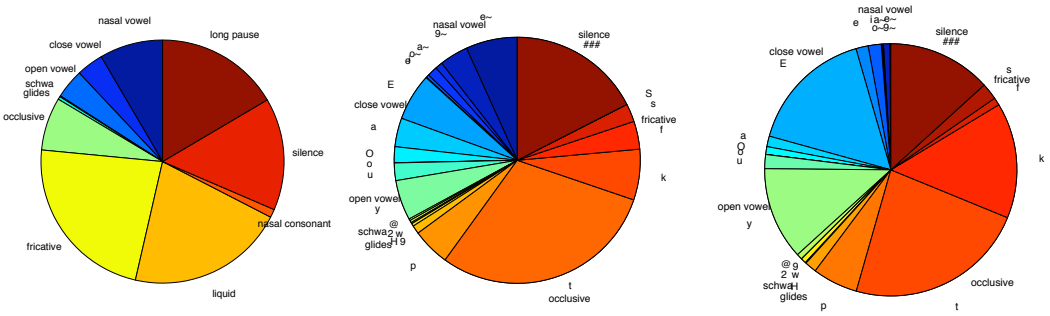

FIG. D.2: Pies of phonetic distributions issued of automatic segmentation. Left : Unsupervised segmentation of JMA. Center : Supervised segmentation of JMA. Right : Supervised segmentation of JFA.

to 4273 (JMA) and 5487 (JFA) voiced segments used in following segment analysis. JFA's vowel distribution confirms theoretical prediction that laughter is mainly based on central vowels [Szameitat 2007].

D.5.2 Acoustic features

The observed variability highlights the need for large sample sizes when studying laughter [Bachorowski 2001]. Therefore it was important to use statistical analysis over computed continuous acoustic features. For each bout and segment, three types of acoustic features are computed :

- continuous features : energy, loudness, voicing coefficient, pitch (f0), formant frequencies and Rd [Fant 1997] are data evolving during segment time span.
- static features : mean and standard deviation of previous continuous features are computed. Duration
- dynamic features : 2-order polynomials of Legendre, model temporal evolution of continuous feature trajectories by slope and curve values.

D.5.3 Segment analysis

Segment analysis exhibits interesting values reported and commented in table D.1. laughter syllables are predominantly based on central vowels (/e/ →/a/). They show higher formant frequencies than normal speech vowels because of extreme positions adopted by the vocal tract during laughter in combination with physiological constraints accompanying production of a pressed voice [Szameitat 2007].

D.5.4 Bout analysis

Acoustic features computed on bouts show several tendencies summed up in the figure D.3. Interestingly, some common aspects of verbal speech prosody seem to be present in laughter prosody like negative pitch slope, negative loudness slope,

acoustic feature	JMA value (std)	JFA value (std)	comment
voicing coef. mean	0.18 (0.14)	0.16 (0.15)	weakly voiced
Rd mean	1.24 (0.35)	1.37 (0.31)	pressed voice quality
f0 mean f0 slope mean	154 (49) -4.8 (13.1)	307 (104) -4.7 (42.3)	normal register not significant for JFA
duration mean mean number of periods	0.08 (0.03) 11.8 (6.5)	0.09 (0.05) 27.8 (21.5)	sexe independent mean(duration) * mean(f0)
1^{st} formant freq. mean	305 (193)	351 (114)	central
2^{nd} formant freq. mean	1486 (356)	1524 (570)	vowel
3^{rd} formant freq. mean	2588 (435)	2622 (605)	/e/

Tab. D.1: Segment acoustic analysis of JMA and JFA vowels.

correlation between f0 mean and loudness mean, and positive vowel duration slope. This last aspect relative to final lengthening has to be further confronted to other models that don't take it into account [Sundaram 2007]. Mean number of vowels per bout is 5 (5) for JMA and 6 (6) for JFA which correspond to the mean number of syllables that compose verbal prosodic groups. The positive 1^{st} formant frequency slope mean is interpretable as a progressive jaw opening during laughter [Sundberg 1995] [Szameitat 2007]. Not only variations (social functions) of laughter are brought up by its type [Campbell 2007b], but also by its prosody which seems to be guided by the same physiological constraints as verbal speech.

D.6 Synthesis part

In order to provide a speaker-like laughter from only one utterance, a unit selection method is used. However no laughter is present in the utterance and one must combine a parametric method to generate realistic laughter from a few units. Observations made in the analysis part are all taken into account for synthesizing laughter. The first phase is composed of automatic segmentation, symbolic analysis and acoustic analysis of the neutral utterance to add laughter. The second phase is a selection algorithm that extracts three segmented phones of the utterance. The third phase designs bout attack and syllables from these phones. The fourth stage is bout prosody synthesis using a parametric model guided by previous bout analysis. Finally, the last phase is prosodic modification of a duplicated signal by a speech processing algorithm. The overall synthesis process is exemplified in figure D.4 and

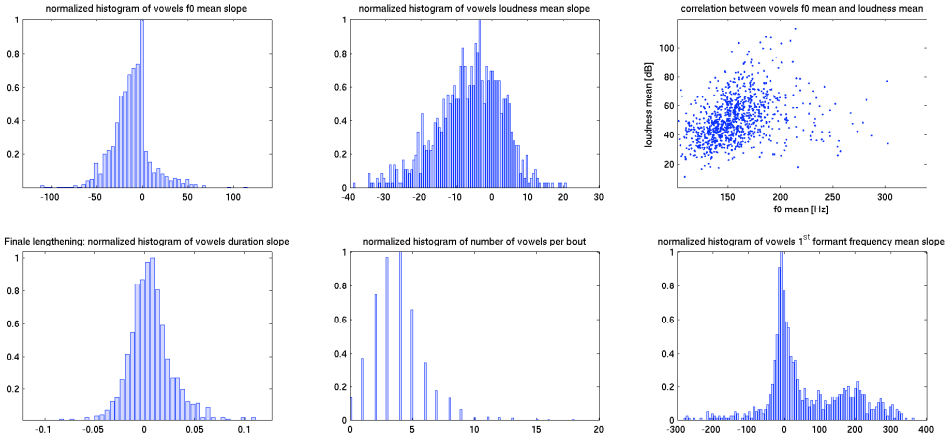

FIG. D.3: Bout acoustic analysis of JMA laughter.

is parameterized by values explained in table D.2.

D.6.1 Phones selection

Once the neutral utterance is phonetically segmented, a unit selection algorithm identifies three phones. The first phone is arbitrarily the occlusive that possesses the maximum positive loudness slope for starting the bout (attack). The second selected phone is also an occlusive (because of automatic segmentation analysis results) but with the minimum absolute loudness slope, which emulates the breathing part of syllables. The third selected phone is a vowel, preferably /a/, then central vowel, then nasal vowel, which satisfies following acoustic constraints : minimum f0 slope, minimum voicing coefficient and minimum Rd mean. The three phones are balanced regarding attack, breath and vowel relative loudness parameters (P1,P2,P3).

D.6.2 Signal duplication

The vowel is truncated if its duration is longer than the maximum number of periods parameter (P4), using pitch and duration. The first syllable, called attack, is made from the concatenation of the first occlusive and the vowel. Other syllables are made from the concatenation of the breathing (second occlusive) and the vowel. Before duplication of this syllable in a number of syllables (P5), the syllable is energy-windowed by a Tukey's window that eliminates occlusive attack and fades out the vowel.

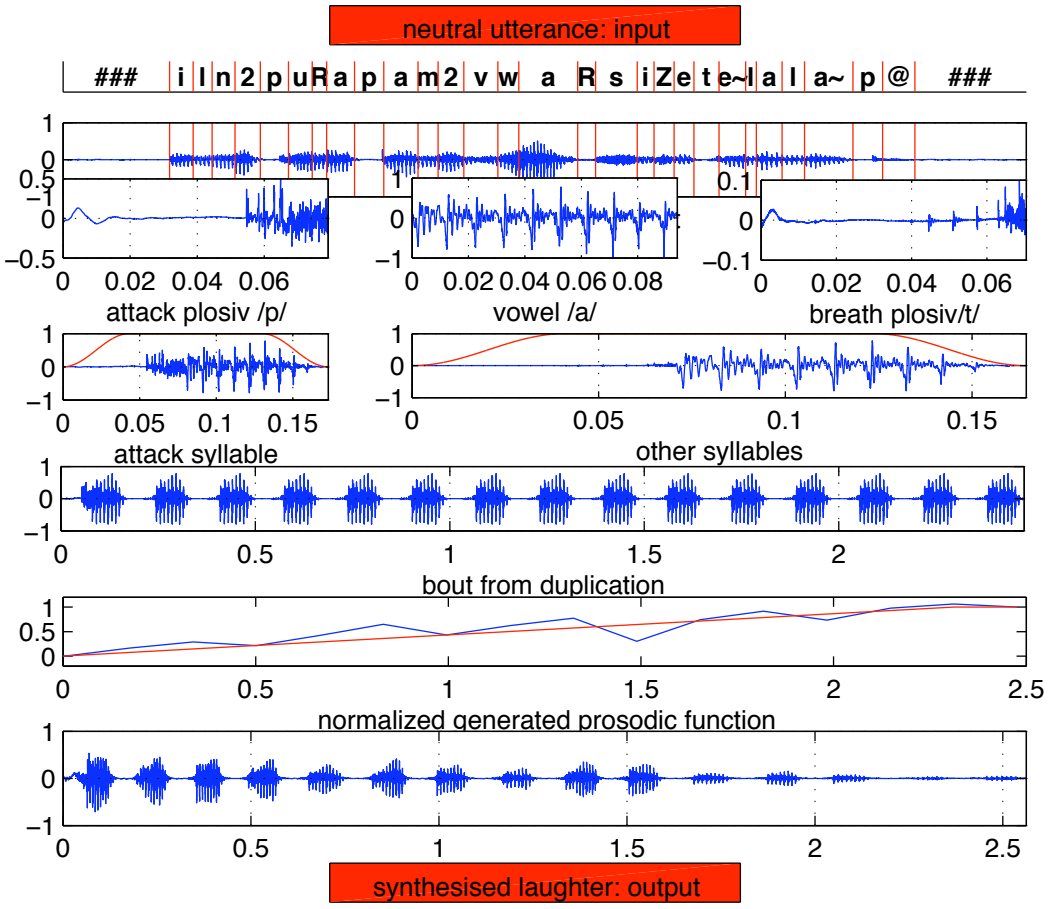

FIG. D.4: Example of a laughter synthesis from a neutral utterance. The number of syllables is 15 to supply a better visualization of the normalized generated prosodic function. Neutral utterance, bout from duplication and synthesised laughter can be listened to in attached sound files.

Param. ID	Name	Def. Val.	Unity
P1	attack relative loudness	0.5	normalized
P2	breath relative loudness	0.1	normalized
P3	vowel relative loudness	1.0	normalized
P4	maximum number of periods during vowel	15	integer
P5	number of syllables	5	integer
P6	time stretch start and end values	0.8 →1.2	slower if > 1
P7	transposition start and end values	1.5 →0.8	higher if > 1
P8	gain start and end values	1 →0	louder if > 1
P9	formant warping function start and end values	400 →500	[Hz] displaced frequency zone

TAB. D.2: Parameters default values that drive synthesis process.

D.6.3 Bout prosody generation

Bout prosody generation comes from an empiric parameterizable mathematical model (as [Sundaram 2007]) that is inspired by the bout analysis part and that is used for providing every transformations factors. A normalized linear function takes decreasing values over the number of syllables and gives the overall movement of the laughter. A triangularly windowed random signal is added to the linear function, in order to simulate laughter variability [Campbell 2007b].

D.6.4 Signal transformation

The same generated abstract prosodic function is used to provide transformation factors used by phase vocoder technology [Bogaards 2004] to transform the duplicated signal. Every syllable is time-stretched, transposed, gained and frequency warped to modify respectively rhythm, intonation, loudness and articulation degree [Beller 2008a] (jaw opening in this case) of the laughter bout. The same function is used to generate all factors to respect the natural and physiological correlation of speech production that seems to be as relevant as in verbal speech (see part D.5.4). Furthermore, parameters range (P6, P7, P8, P9) can be automatically estimated on the neutral utterance.

D.6.5 Results

The randomness of the selected syllables and of the prosodic parameters make the results highly variable as demonstrated by some examples that can be heard in attached sound files or at the following address : `http://www.ircam.fr/anasyn/beller`. The quality of the resulting synthesis has

not yet been evaluated by perceptive tests, but informal characterization of provided laughter bouts encourages the method. Even if the adequacy of the synthesised laughter in the original statement always remain a difficulty [Schroeder 2004], the proposed method resolves partially the problem. The use of segment of the neutral utterance and the limitation of the bout prosody by the physiological constraints measured on the neutral utterance reduces the perceptual distance between the neutral utterance and the speaker-like synthesised laughter. These conditions are necessary but not sufficient because we believe that a part of the function of a laughter lies in the interaction of its prosody in that of the sentence to which it is attached.

D.7 Conclusion

In this paper, we presented our motivation for artistic laughter synthesis. Segmental and prosodic analyses were conducted on a laughter corpus of two Japanese speakers. Statistical acoustic feature analysis of the dynamics of laughter emphasize some natural tendencies reliable to the physiological constraints that prevail in verbal speech prosody. The results then lead to the design of a laughter bout prosodic prototype that encounters randomness to simulate variability of laughter. A semi-parametric method to synthesize speaker-like laughter from one neutral utterance was presented. Future works will now focus on the modification of the voice quality as laughter segments are significantly uttered with pressed voice. The presented method allows laughter-speech synthesis that is another part of our future directions.

D.8 Acknowledgments

The author would like to kindly thank N. Campbell (ATR) for providing the laughter corpus. This work was partially funded by the French RIAM network project VIVOS.

Bibliographie

[Aubergé 2004] V. Aubergé, N. Audibert et A. Rilliard. *E-Wiz : A trapper protocol for hunting the expressive speech corpora*. In LREC2004, pages 179–182, Lisbon, Portugal, 2004.

[Aubergé 2006] V. Aubergé, N. Audibert et A. Rilliard. *Auto-annotation : an alternative method to label expressive corpora*. In LREC2006 - Workshop on Emotional Corpora, pages 45–46, Genova, Italy, 2006.

[Averill 1980] J. R. Averill. Emotion : Theory, research and experience, vol. 1, chapitre A constructivist view of emotion, pages 305–339. New York : Academic Press, 1980.

[Bachorowski 2001] J.-A. Bachorowski, M. J. Smoski et M. J. Owren. *The acoustic features of human laughter* . JASA, vol. 110, pages 1581–1597, Septembre 2001.

[Bailly 1992] G. Bailly et C. Benoit. Talking machines : Theories, models and designs. Amsterdam, North-Holland, 1992.

[Bailly 1995] G. Bailly. *Characterising formant trajectories by tracking vocal tract resonances*. WASKOPF, 1995.

[Bailly 2005] G. Bailly et B Holm. Special issue on quantitative prosody modelling for natural speech description and generation, volume 46, chapitre SFC : a trainable prosodic model, pages 348–364. 2005.

[Ball 2003] E. Ball. *A bayesian heart : computer recognition and simulation of emotion*. In Payr S Trappl R Petta P, editeur, Emotions in humans and artifacts. Cambridge, The MIT Press, 2003.

[Barras 1998] C. Barras, E. Geoffrois, Z. Wu et M. Liberman. *Transcriber : a Free Tool for Segmenting, Labeling and Transcribing Speech*. In LREC1998, pages 1373–1376, 1998.

[Beaucousin 2006] V. Beaucousin, A. Lacheret, M.R. Turbelin, M. Morel, B. Mazoyer et N. Mazoyer. *FRMI Study of Emotional Speech Comprehension*. Cerebral Cortex, 2006.

[Beller 2004] G. Beller. *un synthétiseur vocal par sélection d'unités*. rapport de stage, IRCAM, Paris, 2004.

[Beller 2006a] G. Beller et A. Marty. *Talkapillar : outil d'analyse de corpus oraux*. In Rencontres Jeunes Chercheurs de L'Ecole Doctorale 268, pages 97–100. Paris 3 Sorbonne-Nouvelle, 2006.

[Beller 2006b] G. Beller, D. Schwarz, T. Hueber et X. Rodet. *Speech Rates in French Expressive Speech*. In Speech Prosody 2006, pages 672–675, Dresden, 2006. SproSig, ISCA.

[Beller 2007a] G. Beller. *Context Dependent Transformation of Expressivity in Speech Using a Bayesian Network*. In ParaLing, pages 48–41, Germany, Saarbrücken, August 2007.

[Beller 2007b] G. Beller. *Influence de l'expressivité sur le degré d'articulation*. In RJCP, Rencontres Jeunes Chercheurs de la Parole, pages 24–27, 2007.

[Beller 2008a] G. Beller, N. Obin et X. Rodet. *Articulation Degree as a Prosodic Dimension of Expressive Speech*. In Speech Prosody 2008, pages 681–684, Campinas, 2008.

[Beller 2008b] G. Beller, C. Veaux et X. Rodet. *IrcamCorpusExpressivity : Non-verbal Words and Restructurings*. In LREC2008 - workshop on emotions, 2008.

[Beller 2009a] G. Beller. *Transformation of Expressivity in Speech*. In Peter Lang, editeur, The Role of Prosody in the Expression of Emotions in English and in French. Peter Lang, 2009.

[Beller 2009b] G. Beller, C. Veaux, G. Degottex, N. Obin, P. Lanchantin et X. Rodet. *IRCAM Corpus Tools : Système de Gestion de Corpus de Parole*. TAL, 2009.

[Bilhaut 2006] F. Bilhaut et A. Widlöcher. *LinguaStream : An Integrated Environment for Computational Linguistics Experimentation*. In 11th Conference of the European Chapter of the Association of Computational Linguistics (Companion Volume), pages 95–98, Trento, Italy, 2006.

[Bird 2000] S. Bird, D. Day, J. Garofolo, J. Henderson, C. Laprun et M. Liberman. *ATLAS : A Flexible and Extensible Architecture for Linguistic Annotation*. In in Proceedings of the Second International Conference on Language Resources and Evaluation, pages 1699–1706, 2000.

[Bird 2001] S. Bird et M. Liberman. *A formal framework for linguistic annotation*. Speech Commun., vol. 33, no. 1-2, pages 23–60, 2001.

[Blankinship 2001] E. Blankinship et R. Beckwith. Uist '01 : Proceedings of the 14th annual acm symposium on user interface software and technology, chapitre Tools for expressive text-to-speech markup, pages 159–160. ACM, New York, NY, USA, 2001.

[Boersma 2001] P. Boersma et D. Weenink. *Praat, a system for doing phonetics by computer*. In Glot international, volume 5-9 of *10*, pages 341–345, 2001.

[Bogaards 2004] N. Bogaards, A. Roebel et X. Rodet. *Sound Analysis and Processing with AudioSculpt 2*. In ICMC, Miami, USA, Novembre 2004.

[Bozkurt 2005] B. Bozkurt et L. Couvreur. *On The Use of Phase Information for Speech Recognition*. In EUSIPCO, 2005.

[Bresin 2000] R. Bresin. *Virtual Virtuosity. Studies in Automatic Music Performance*. PhD thesis, TMH, KTH, dec 2000.

[Bresin 2003] R. Bresin et S. Dahl. The sounding object, chapitre Experiments on gestures : walking, running, and hitting, pages 111–136. Florence, Italy : Mondo Estremo., 2003.

[Brun 1698] Le Brun. *conférence sur l'expression générale et particulière*. In Académie Royale de Peinture et de Sculpture de Paris, 1698.

[Buck 1984] R. Buck. The communication of emotion. Guilford Press, New York, 1984.

[Bulut 2007] M. Bulut, S. Lee et S. Narayanan. *a statistical approach for modeling prosody features using pos tags for emotional speech synthesis*. In ICASSP, 2007.

[Burkhardt 2006] F. Burkhardt, N. Audibert, L. Malatesta, O. T urk, L. Arslan et V. Auberge. *Emotional Prosody - Does Culture Make A Difference ?* In Speech Prosody, 2006.

[Caelen-Haumont 2004] G. Caelen-Haumont. *Valeurs pragmatiques de la proéminence prosodique lexicale : de l'outil vers l'analyse*. In JEP-TALN, 2004.

[Campbell 2003] N. Campbell et P. Mokhtari. *Voice quality : the 4th prosodic dimension*. In XVth ICPhS, volume 3, pages 2417–2420, Barcelona, 2003.

[Campbell 2005] N. Campbell, H. Kashioka et R. Ohara. *No laughing matter*. In Interspeech, pages 465–468., 2005.

[Campbell 2007a] N. Campbell. *Changes in voice quality due to social conditions*. In ICPhS, Saarbrücken, August 2007.

[Campbell 2007b] N. Campbell. *Whom we laugh with affects how we laugh*. In Interdisciplinary Workshop on The Phonetics of Laughter, 2007.

[Campedel-Oudot 1998] M. Campedel-Oudot. *Speech processing using a sinusoidal + noise model Robust estimation of the spectral envelope*. PhD thesis, Ecole nationale supérieure des télécommunications, Paris, FRANCE, 1998.

[Cassidy 2001] S. Cassidy et J. Harrington. *Multi-level annotation in the Emu speech database management system*. Speech Commun., vol. 33, 1-2, pages 61–77, 2001.

[Chafe 1992] W. Chafe. *The importance of corpus linguistics to understanding the nature of language*. In Jan Svartvik, editeur, Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82, pages 79–97. Berlin - New York : Mouton de Gruyter, 1992.

[Changeux 1983] J. P. Changeux. L'homme neuronal. Hachette, Paris, 1983.

[Cheveigné 2002] A. De Cheveigné et H. Kawahara. *YIN, a Fundamental Frequency Estimator for Speech and Music*. JASA, vol. 111, pages 1917–1930, 2002.

[Chung 2000] S.-J. Chung. *L'expression et la perception de l'émotion extraite de la parole spontannée : évidences du coréen et de l'anglais*. phonétique, Université PARIS III - Sorbonne Nouvelle : ILPGA, Paris, 2000.

[Combescure 1981] P. Combescure. *20 listes de dix phrases phonétiquement équilibrées*. Revue d'Acoustique, vol. 56, pages 34–38, 1981.

[Cunningham 2002] H. Cunningham, D. Maynard, K. Bontcheva et V. Tablan. *GATE : A framework and graphical development environment for robust NLP tools and applications*. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, pages 168–175, 2002.

[D'Alessandro 1995] C. D'Alessandro et P. Mertens. *Automatic pitch contour stylization using a model of tonal perception*. In Computer Speech and Language, volume 9, pages 257–288, 1995.

[D'Alessandro 1997] C. D'Alessandro et B. Doval. *Spectral representation and modelling of glottal flow signals*. ESCA, 1997.

[D'Alessandro 2003a] C. D'Alessandro et B. Doval. *Voice quality modification for emotional speech synthesis.* In Eurospeech, Geneva, Switzerland, 2003.

[D'Alessandro 2003b] C. D'Alessandro, B. Doval et K. Scherer. *Voice quality : functions, analysis and synthesis*. In ISCAtutorial and research workshop VOQUAL'03, 2003.

[D'Alessandro 2004] C. D'Alessandro. L'évaluation des systèmes de traitement de l'information, chapitre L'évaluation des systèmes de synthèse de la parole, pages 215–239. Hermès, Lavoisier, Paris, 2004.

[Darwin 1965] C. Darwin. Expression of emotion in man and animals. Konrad Lorenz, 1965.

[Degottex 2008a] G. Degottex, E. Bianco et X. Rodet. *Usual to particular phonatory situations studied with high-speed videoendoscopy*. In The 6th International Conference on Voice Physiology and Biomechanics, pages 19–26, 2008.

[Degottex 2008b] G. Degottex et X. Rodet. *Voice source and vocal tract separation*. to be published, 2008.

[Delattre 1955] P. Delattre, M. Liberman et F. Cooper. *Acoustic loci and transitional cues for consonants*. JASA, vol. 27, pages 769–773, 1955.

[Descartes 1948] R. Descartes. Les passions de l'âme, oeuvres philosophiques et morales. bibliothèque des lettres, 1948.

[Desjardins 2001] Lucie Desjardins. Le corps parlant : savoirs et représentation des passions au xviie siècle. Presses Université Laval, 2001.

[Devillers 2003a] L. Devillers, L. Lamel et I. Vasilescu. *Emotion detection in Task-oriented spoken dialogs*. In IEEE ICME, 2003.

[Devillers 2003b] L. Devillers, I. Vasilescu et C. Mathon. *Prosodic cues for perceptual emotion detection in task-oriented Human-Human corpus*. In ICPhS, 2003.

[Devillers 2005] L. Devillers, L. Vidrascu et L. Lamel. *Emotion detection in real-life spoken dialogs recorded in call center*. Journal of Neural Networks, special issue on Emotion and Brain, vol. 18, no. 4, pages 407–422, 2005.

[Douglas-cowie 2000] E. Douglas-cowie, R. Cowie et M. Schroeder. *A New Emotion Database : Considerations, Sources and Scope*. In SpeechEmotion2000, pages 39–44, 2000.

[Douglas-cowie 2007] E. Douglas-cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir et K. Karpouzis. *The HUMAINE Database : Addressing the Collection and Annotation of Naturalistic and*. In Induced Emotional Data, Affective Computing and Intelligent Interaction, pages 488–500, 2007.

[Doval 2000] B. Doval et C. D'Alessandro. *Spectral Correlates of Glottal Waveform models : An analytic study*. Rapport technique, LIMSI, 2000.

[Doval 2003] B. Doval, N. D'Alessandro et N. Henrich. *The voice source as a causal/anticausal linear filter*. In VOQUAL, August 2003.

[Doval 2006] B. Doval, C. D'Alessandro et N. Henrich. *The Spectrum of Glottal Flow Models*. Acta Acustica united with Acustica, vol. 92, no. 6, pages 1026–1046, November/December 2006.

[Duncan 1989] G. Duncan, B. Yegnanarayanan et Hema A. Murthy. *A non Parametric Method of Formant Estimation Using Group Delay Spectra*. In IEEE, 1989.

[Durand 2005] J. Durand, B. Laks et C. Lyche. *Un corpus numérisé pour la phonologie du français*. In G. Williams, editeur, La linguistique de corpus, pages 205–217. Presses Universitaires de Rennes, 2005.

[Dutoit 1997] T. Dutoit. *High-Quality Text-to-Speech Synthesis : an Overview*. Journal of Electrical & Electronics Engineering, Australia : Special Issue on Speech Recognition and Synthesis, vol. 17, no. 1, pages 25–37, 1997.

[Dutoit 2003] T. Dutoit et Y. Stylianou. Handbook of computational linguistics, chapitre Chapter 17 : Text-to-Speech Synthesis, pages 323–338. Oxford University Press, 2003.

[Eide 2004] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny et J. Pitrelli. *A Corpus-Based Approach to <Ahem/> Expressive Speech Synthesis*. In 5th ISCA Speech Synthesis Workshop, 2004.

[Ekman 1999a] P. Ekman. The handbook of cognition and emotion, chapitre Basic Emotions. John Wiley & Sons, Ltd., 1999.

[Ekman 1999b] P. Ekman. The handbook of cognition and emotion, chapitre Facial Expressions, pages 301–320. John Wiley & Sons, Ltd., 1999.

[Eriksson 2007] E.J. Eriksson, R.D. Rodman et R.C. Hubal. Speaker classification i, volume 4343 of *Lecture Notes in Computer Science*, chapitre Emotions in Speech : Juristic Implications, pages 152–173. 2007.

[Evangelista 2000] Gianpaolo Evangelista. *Real-time time-varying frequency warping via short-time Laguerre transform*. In Proc. DAFx00, pages 7–12, 2000.

[Evangelista 2001a] G. Evangelista et S. Cavaliere. *Audio Effects Based on Biorthogonal Time-Varying Frequency Warping*. EURASIP Journal on Applied Signal Processing, vol. 1, pages 27–35, 2001.

[Evangelista 2001b] G. Evangelista et S. Salvatore. *Time-Varying Frequency Warping : Results and Experiments*. Rapport technique, 2001.

[Fant 1960] G. Fant. Acoustic theory of speech production. Mouton, The Hague, 1960.

[Fant 1985] G. Fant, J. Liljencrants et Q. Lin. *A four-parameter model of glottal flow*. STL-QPSR, vol. 4, pages 1–13, 1985.

[Fant 1995] G. Fant. *Quarterly Progress and Status Report : The LF-model revisited. Transformations and frequency domain analysis*. 36 2-3, Dept. for Speech, Music and Hearing, KTH University, 1995. 119-156.

[Fant 1997] G. Fant. *The voice source in connected speech,*. Speech Communication, vol. 22, pages 125–139, 1997.

[Feng 2003] Z. Feng et Pan Y. *Popular music retrieval by detecting mood*. In SIGIR, 2003.

[Fitrianie 2006] S. Fitrianie et L.J.M. Rothkrantz. *Constructing Knowledge for Automated Text-Based Emotion Expressions*. In CompSysTech2006, 2006.

[Fónagy 1972a] I. Fónagy et E. Bérard. *Il est huit heure : Contribution à l'analyse sémantique de la vive voix*. Phonetica, vol. 26, pages 157–192, 1972.

[Fonàgy 1972b] I. Fonàgy et K. Magdics. Intonation, chapitre Emotional patterns in language and music, pages 286–312. Pinguin Books, 1972.

[Fónagy 1983] I. Fónagy. La vive voix : essais de psycho-phonétique. Payot, Paris, 1983.

[Fougeron 1998] C. Fougeron et S.A. Jun. *Rate effects on French intonation : prosodic organization and phonetic realization*. Journal of Phonetics, vol. 26, pages 45–69, 1998.

[Friberg 1999] A. Friberg et J. Sundberg. *Does music performance allude to locomotion ? A model of final ritardandi derived from measurements of stopping runners.* JASA, vol. 105(3), pages 1469–1484, 1999.

[Frijda 1986] N. H. Frijda. The emotions. Cambridge University Press, 1986.

[Fritz 2004] C. Fritz. *La clarinette et le clarinettiste : Influence du conduit vocal sur la production du son*. PhD thesis, Université Paris 6 et Université de New South Wales, 2004.

[Fujisaki 1981] H. Fujisaki. *Dynamic characteristics of voice fundamental frequency in speech and singing. Acoustical analysis and physiological interpretations*. Rapport technique, Dept. for Speech, Music and Hearing, 1981.

[Gabrielson 1996] A. Gabrielson et P.N. Juslin. *Emotional expression in music performance : Between the performers intention and the listeners experience*. Psychology of music, vol. 24, pages 68–91, 1996.

[Galarneau 2001] A. Galarneau, P. Tremblay et P. Martin. *dictionnaire de la parole*. Rapport technique, Laboratoire de Phonétique et Phonologie de l'Université Laval à Québec, 2001.

[Garnier 2007] M. Garnier. *Communiquer en environnement bruyant : de l'adaptation jusqu'au forcage vocal*. PhD thesis, University of Paris VI, 2007.

[Gauvain 1996] J.L. Gauvain et L. Lamel. *Large Vocabulary Continuous Speech Recognition : from Laboratory Systems towards Real-World Applications*, 1996.

[Gendrot 2002] C. Gendrot. *Ouverture de la glotte, Fo, intensité et simulations émotionnelles : le cas de la joie, la colère, la surprise, la tristesse et la neutralité*. In XXIVèmes Journées d'Étude sur la Parole, 2002.

[Gendrot 2004] C. Gendrot et M. Adda-Decker. *Analyses formantiques automatiques de voyelles orales : évidence de la réduction vocalique en langues franàaise et allemande*. In MIDL, 2004.

[Gobl 2003] C. Gobl et A.N. Chasaide. *The role of voice quality in communicating emotion, mood and attitude*. Speech Communication, vol. 40, no. 1-2, pages 189–212, 2003.

[Gouvea 1997] E.B. Gouvea et R.M. Stern. *Speaker normalization through formant-based warping of the frequency scale*. In Eurospeech, pages 1139–1142, 1997.

[Gouvea 1999] E.B. Gouvea. *acoustic-feature-based frequency warping for speaker normalization*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 1999.

[Grobet 2001] A. Grobet et A.C. Simon. *Diffèrents critères de dèfinition des unitès prosodiques maximales*. Cahiers de Linguistique Française, vol. 23, pages 143–163, 2001.

[Gussenhoven 1988] C. Gussenhoven et A. C. M. Rietveld. *Fundamental Frequency Declination in Dutch : Testing Three Hypotheses*. Journal of Phonetics, vol. 16, pages 355–369, 1988.

[Gut 2004] U. Gut, J-T. Milde, H. Voormann et U. Heid. *Querying Annotated Speech Corpora*. In Proceedings of Speech Prosody 2004, pages 569–572, Nara, Japan, 2004.

[Haermae 2000] A. Haermae, M. Karjalainen, L. Savioja, V. Vaelimaeki, U. K. Laine et J. Huopaniemi. *Frequency-warped signal processing for audio applications*. In Audio Engineering Society, editeur, AES 108th Convention, February 2000.

[Haermae 2001] A. Haermae. *frequency-warped autoregressive modeling and filtering*. PhD thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 2001.

[Hastie 2001] T. Hastie, R. Tibshirani et J. Friedman. The elements of statistical learning. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[Henrich 2001] N. Henrich. *Etude de la source glottique en voix parlée et chantée*. PhD thesis, Université Paris 6, Paris, France, nov 2001.

[Henrich 2002] N. Henrich, C. D'Alessandro et B. Doval. *Glottal Flow Models : Waveforms, Spectra And Physical Measurements*, 2002.

[Hevner 1936] K. Hevner. *Experimental studies of the elements of expression in music*. American Journal of Psychology, vol. 48, pages 246–268, 1936.

[Hirst 1993] D. Hirst et R. Espesser. *Automatic modelling of fundamental frequency using a quadratic spline function*. Travaux de l'Institut de Phonétique d'Aix, vol. 15, pages 71–85, 1993.

[Hirst 2000] D. Hirst, A. Di Cristo et R. Espesser. Prosody : Theory and experiment, chapitre Levels of representation and levels of analysis for intonation, pages 51–87. Kluwer Academic, M. Horne (ed), 2000.

[Horri 1982] Y. Horri. *Jitter and Shimmer Differences among Sustained Vowel Phonations*. Journal of Speech and Hearing Research, vol. 25, pages 12–14, 1982.

[Hozjan 2006] V. Hozjan et Z. Kacic. *A rule-based emotion-dependent feature extraction method for emotion analysis from speech*. The Journal of the Acoustical Society of America, vol. 119, no. 5, pages 3109–3120, May 2006.

[Hsia 2007] Chi-Chun Hsia, Chung-Hsien Wu, et Jian-Qi Wu. *conversion function clustering and selection for expressive voice conversion*. In ICASSP, 2007.

[Hu 2006] X. Hu, J.S. Downie et A.F. Ehmann. *Exploiting Recommended Usage Metadata : Exploratory Analyses*. In ISMIR, 2006.

[Hu 2007] Xiao Hu et J. Stephen Downie. *Exploring mood metadata : Relationships with genre, artist and usage metadata*. In ISMIR, 2007.

[James 2007] W. James. What is an emotion ? Wilder Publications, 2007.

[Johnstone 1999] T. Johnstone et K. R. Scherer. *The effects of emotions on voice quality*. In XIVth ICPhS, pages 2029–2032, 1999.

[Juslin 2003] P.N. Juslin. *Five facets of musical expression : a psychologist's perspective on music performance*. Psychology of Music, vol. 31(3), pages 273–302, 2003.

[Juslin 2004] P.N. Juslin et P. Laukka. *Expression, perception, and induction of musical emotions : A review and a questionnaire study of everyday listening*. Journal of New Music Research, vol. 33, no. 3, pages 217–238, 2004.

[Juslin 2006] P.N. Juslin, J. Karlsson, Lindstrom E., A. Friberg et E Schoonderwaldt. *Play It Again With Feeling : Computer Feedback in Musical Communication of Emotions*. Journal of Experimental Psychology, vol. 12, no. 2, pages 79–95, 2006.

[Kaehler 2000] B. Kaehler, J. Smith et J. Wolfe. *Longueur de confusion sur la plage vocalique*. In JEP XXIII, Juin 2000.

[Kain 2007] A. Kain, Q. Miao et J.P.H. van Santen. *Spectral Control in Concatenative Speech Synthesis*. In SSW6, 2007.

[Kates 2005] J.M. Kates et K.H. Arehart. *Multichannel Dynamic-Range Compression Using Digital Frequency Warping*. EURASIP Journal on Applied Signal Processing, vol. 18, pages 3003–3014, 2005.

[Kawahara 2005a] H. Kawahara, A. de Cheveigne, H. Banno, T. Takahashi et T. Irino. *Nearly Defect-free F0 Trajectory Extraction for Expressive Speech Modifications based on STRAIGHT*. In Interspeech2005, pages 537–540, Lisboa, 2005.

[Kawahara 2005b] H. Kawahara et T. Irino. Speech separation by humans and machines, volume Engineering, chapitre Underlying Principles of a High-quality Speech Manipulation System STRAIGHT and Its Application to Speech Segregation, pages 167–180. 2005.

[Kendall 1969] M. G. Kendall et A. Stuart. The advanced theory of statistics, volume 1. Charles Griffin, 1969.

[Kipper 2007] S. Kipper et D. Todt. *Series of similar vocal elements as a crucial acoustic structure in human laughter*. In Interdisciplinary Workshop on The Phonetics of Laughter, 2007.

[Klatt 1980] D.H. Klatt. *Software for a cascade/parallel formant synthesizer*. Journal of the Acoustic Society of America, vol. 67, no. 3, pages 971–995, March 1980.

[Klingholz 1985] F. Klingholz et F. Martin. *Quantitative spectral evaluation of shimer and jitter*. Journal of Speech and Hearing Research, vol. 28, pages 169–174, 1985.

[Kreutz 2008] G. Kreutz, U. Ott, D. Teichmann, P. Osawa et D. Vaitl. Psychology of music, volume 36, chapitre Using music to induce emotions : Influences of musical preference and absorption, pages 101–126. 2008.

[Lacheret-Dujour 1999] A. Lacheret-Dujour et F. Beaugendre. La prosodie du francais. CNRS langage, 1999.

[Lai 2004] C. Lai et S. Bird. *Querying and updating treebanks : A critical survey and requirements analysis*. In In Proceedings of the Australasian Language Technology Workshop, pages 139–146, 2004.

[Lamel 1991] L.F. Lamel, J.-L. Gauvain et M. Eskénazi. *Bref, a large vocabulary spoken corpus for French*. In EuroSpeech, pages 505–508, 1991.

[Lanchantin 2008] P. Lanchantin, A. C. Morris, X. Rodet et C. Veaux. *Automatic Phoneme Segmentation with Relaxed Textual Constraints*. In LREC2008, Marrakech, Morocco, 2008.

[Lazarus 1991] R.S. Lazarus. Emotion and adaptation. Oxford University Press, 1991.

[LeDoux 2005] Joseph LeDoux. Le cerveau des émotions. Lavoisier, 2005.

[Lee 1996] L. Lee et R. C. Rose. *Speaker normalization using efficient frequency warping procedures*. In ICASSP1996, pages 353–356, 1996.

[Levenson 1990] R. W. Levenson, P. Ekman et W. V. Friesen. *Voluntary facial action generates emotion specific autonomic nervous system activity*. Psychophysiology, vol. 27, pages 363–384, 1990.

[Li 2003] Li et Ogihara. *Detecting emotion in music*. In ISMIR, 2003.

[Lindblom 1983] B. Lindblom. Economy of speech gestures, volume The Production of Speech. Spinger-Verlag, New-York, 1983.

[Lolive 2006] D/ Lolive, N. Barbot et O. Boeffard. *Modélisation B-spline de contours mélodiques avec estimation du nombre de paramètres libres par un critère MDL*. In JEP, 2006.

[Lu 1999] H.L. Lu et J.O. Smith. *Joint Estimation of Vocal Tract Filter and Glottal Source Waveform via Convex Optimization*. IEEE-WASPAA, 1999.

[Lu 2002] Hui-Ling Lu. *Toward a High-quality Singing Synthesizer with Vocal Texture Control*. PhD thesis, Stanford, 2002.

[Lu 2006] L. Lu et S. Zhang. *Automatic Mood Detection and Tracking of Music Audio Signals*. IEEE Transactions On Audio, Speech, And Language Processing, vol. 14, no. 1, JANUARY 2006.

[Luck 2008] G. Luck, P. Toiviainen, J. Erkkila, O. Lartillot, K. Riikkila, A. Makela, K. Pyhaluoto, H. Raine, L. Varkila et J. Varri. Psychology of music, volume 36, chapitre Modelling the relationships between emotional responses to, and musical content of, music therapy improvisations, pages 25–45. 2008.

[MacWhinney 2000] B. MacWhinney. The childes project : Tools for analyzing talk, third edition, volume Volume I : Transcription format and programs. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

[Maghbouleh 1998] A. Maghbouleh. *ToBI accent type recognition*. In ICSLP, 1998.

[Mandel 2006] P. Mandel et D. Ellis. *Support vector machine active learning for music retrieval*. Multimedia Systems, vol. 12, no. 1, Aug 2006.

[Mertens 2004] P. Mertens. *The Prosogram : Semi-Automatic Transcription of Prosody based on a Tonal Perception Model*. In Speech Prosody, 2004.

[Meyer 1956] L.B. Meyer. Emotion and meaning in music. Chicago University Press, 1956.

[Mixdorff 1999] H Mixdorff. *A novel approach to the fully automatic extraction of Fujisaki model parameters*. In ICASSP, 1999.

[Morlec 1997] Y. Morlec. *Génération multiparamétrique de la prosodie du français par apprentissage automatique.* PhD thesis, INPG, 1997.

[Morris 2006] A. Morris. *Automatic segmentation*. Rapport technique, IRCAM, 2006.

[Müller 2005] C. Müller. *A flexible stand-off data model with query language for multi-level annotation*. Annual Meeting of the Association for Computational Linguistics, , pages 109–112, 2005.

[Murphy 2001] K. Murphy. *The Bayes Net Toolbox for Matlab*. In Computing Science and Statistics, volume 33, 2001.

[Murray 2008] I.R. Murray et J.L. Arnott. *Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech*. Comput. Speech Lang., vol. 22, no. 2, pages 107–129, 2008.

[Murthy 1989a] H.A. Murthy. *Formant Extraction from Phase Weigthed Group Delay Function*. IEEE, 1989. voir notes du 9.3.07.

[Murthy 1989b] H.A. Murthy. *non parametric method using group delay function*. IEEE, 1989.

[Murthy 1989c] H.A. Murthy, K.V. Madhu Murthy et B. Yegnanarayana. *Formant extraction from phase using weighted group delay function*. In Electronics Letters, volume 25, pages 1609–1611. IEE, 1989.

[Murthy 1991a] H A. Murthy et B. Yegnanarayana. *Speech processing using group delay functions*. Elsevier Signal Processing, vol. 22, pages 259–267, 1991.

[Murthy 1991b] H.A. Murthy et B. Yegnanarayana. *Formant extraction from group delay function*. Elsevier Signal Processing, vol. 22, pages 209–221, 1991.

[Murthy 2003] H.A. Murthy et V.R. Rao Gadde. *The modified Group Delay Function and its Application to Phoneme Recognition*. In ICASSP, 2003.

[Naïm 2004] P. Naïm, P-H. Wuillemin, P. Leray, O. Pourret et A. Becker. Réseaux bayésiens. Eyrolles, Paris, 2004.

[Nakov 2005] P. Nakov, A. Schwartz, B. Wolf et M. Hearst. *Supporting annotation layers for natural language processing*. Annual Meeting of the Association for Computational Linguistics, pages 65–68, 2005.

[Obin 2008] N. Obin, J.P. Goldman, M. Avanzi et A. Lacheret-Dujour. *Comparaison de 3 outils de détection automatique de proéminence en français parlé*. In XXVIIèmes Journées d'Études de la Parole, pages 153–157, Avignon, France, 2008.

[Oostdijk 2000] N. Oostdijk. *The Spoken Dutch Corpus : Overview and first evaluation*. In LREC2000, pages 887–893, 2000.

[Palmer 1989] C. Palmer. *Mapping musical thought to musical performance*. Journal of experimental psychology. Human perception and performance, vol. 15, no. 2, pages 331–346, 1989.

[Patel 2008] A.D. Patel. Music, language, and the brain. Oxford University Press, 2008.

[Peeters 2001] G. Peeters. *Modeles et modification du signal sonore adaptés a ses caracteristiques locales*. Phd thesis, UPMC, Ircam - Centre Pompidou 1, Place Igor Stravinsky, 75004 Paris, July 2001. X. RODET Directeur de thèse.

[Peeters 2004] G. Peeters. *A large set of audio features for sound description (similarity and classication) in the CUIDADO project*. Rapport technique, IRCAM, 2004.

[Pfitzinger 2006] H.R. Pfitzinger. *Five Dimensions of Prosody : Intensity, Intonation, Timing, Voice Quality, and Degree of Reduction*. In H Hoffmann R. ; Mixdorff, editeur, Speech Prosody, numéro 40 de Abstract Book, pages 6–9, Dresden, 2006.

[Picard 1997] R. Picard. Affective computing. MIT Press, 1997.

[Piu 2007] M. Piu et R. Bove. *Annotation des disfluences dans les corpus oraux*. In RECITAL, 2007.

[Plutchik 1980] R. Plutchik. Emotion : Theory, research, and experience vol. 1. theories of emotion, chapitre A general psychoevolutionary theory of emotion, pages 3–33. New York : Academic, 1980.

[Pohle 2005] Pohle, Pampalk et Widmer. *Evaluation of Frequently Used Audio Features for Classification of Music into Perceptual Categories.* In CBMI, 2005.

[Potamianos 1997] A. Potamianos et R. C. Rose. *On Combining Frequency Warping and Spectral Shaping in Based Speech Recognition*. In Proc. ICASSP '97, pages 1275–1278, Munich, Germany, 1997.

[Rasamimanana 2007] N. H. Rasamimanana, F. Kaiser et F. Bevilacqua. *Transients control of violin players : relationships between bow acceleration and string irregular vibrations*. (in preparation), 2007.

[Rasamimanana 2008] N. H. Rasamimanana. *Geste instrumental du violoniste en situation de jeu : analyse et modélisation*. PhD thesis, Université Paris 6 - IRCAM UMR STMS, 2008.

[Rodet 1997] X. Rodet. *Musical Sound Signals Analysis/Synthesis : Sinusoidal+Residual and Elementary Waveform Models*. In TFTS, Août 1997.

[Roebel 2003] A. Roebel. *A new approach to transient processing in the phase vocoder*. DAFX, 2003.

[Roebel 2005a] A. Roebel et X. Rodet. *Real time signal transposition with envelope preservation in the phase vocoder*. In ICMC, 2005.

[Roebel 2005b] A. Roebel et X. Rodet. *Real Time Signal Transposition With Envelope Preservation In The Phase Vocoder*. In ICMC, 2005.

[Rossi 1981] M. Rossi, A. Di Cristo, D. Hirst, P. Martin et Y. Nishinuma. L'intonation. de l'acoustique à la sémantique. Klincksieck, 1981.

[Scherer 1984] K. R. Scherer. *Emotion as a multicomponent process : A model and some cross-cultural data*. Review of Personality and Social Psychology, vol. 5, pages 37–63, 1984.

[Scherer 1987] K. R. Scherer. *Toward a Dynamic Theory of Emotion : The Component Process Model of Affective States*. Geneva Studies in Emotion and Communication, vol. 1, page 1â98, 1987.

[Scherer 2005] K. R. Scherer. *What are emotions ? And how can they be measured ?* Social Science Information, vol. 44(4), pages 695–729, 2005.

[Scherer 2006] K. R. Scherer. Chinese spoken language processing, volume 4274, chapitre The Affective and Pragmatic Coding of Prosody, pages 13–14. 2006.

[Schoentgen 1995] J. Schoentgen et R. de Guchtenneere. *Time series analysis of jitter*. Journal of Phonetics, vol. 23, pages 189–201, 1995.

[Schroeder 2001] M. Schroeder. *Emotional Speech Synthesis–a Review*. In Eurospeech, Aalborg, pages 561–564, DFKI, Saarbrücken, Germany : Institute of Phonetics, University of the Sarland, 2001.

[Schroeder 2003] M. Schroeder. *Speech and Emotion Research : An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*. PhD thesis, University of Saarland, 2003.

[Schroeder 2004] M. Schroeder et J. Trouvain. *How (Not) to Add Laughter to Synthetic Speech*. In Workshop on Affective Dialogue Systems Kloster Irsee, 2004.

[Schroeder 2006] M. Schroeder, D. Heylen et I. Poggi. *Perception of non-verbal emotional listener feedback*. In Speech Prosody 2006, Dresden, Germany., 2006.

[Schwarz 2006] D. Schwarz, G. Beller, B. Verbrugghe et S. Britton. *Real-Time Corpus-Based Concatenative Synthesis with CataRT*. In DAFx, 2006.

[Shove 1995] P. Shove et B. Repp. *Musical motion and performance : theoretical and empirical perspectives*. J. Rink, pages 55–83, 1995.

[Shuang 2006] Shuang, Zhi-Wei, Bakis, Raimo, Shechtman, Slava, Chazan, Dan, Qin et Yong. *Frequency warping based on mapping formant parameters*. In Interspeech, numéro 1768, 2006.

[Sjölander 2000] K. Sjölander et J. Beskow. *WaveSurfer - An Open Source Speech Tool*. In International Conference on Spoken Language Processing, volume 4, pages 464–467, Beijing, China, 2000.

[Stanislavski 1966] C. Stanislavski. la formation de l'acteur. Pygmalion, 1966.

[Stylianou 1996] Y. Stylianou. *Decomposition of Speech Signals into a Deterministic and a Stochastic part*. In ICSLP '96, pages 1213–1216, 1996.

[Sundaram 2007] S. Sundaram et S. Narayanan. *Automatic acoustic synthesis of human-like laughter*. The Journal of the Acoustical Society of America, vol. 121, pages 527–535, January 2007.

[Sundberg 1995] J. Sundberg et J. Skoog. *Jaw opening, vowel and pitch*. In STL-QPSR, volume 36, pages 043–050, 1995.

[Szameitat 2007] D.P. Szameitat, C.J. Darwin, A.J. Szameitat, D. Wildgruber, A. Sterr, S. Dietrich et K. Alter. *Formant characteristics of human laughter*. In Interdisciplinary Workshop on The Phonetics of Laughter, 2007.

[Tao 2006] J. Tao, Y. Kang et A. Li. *Prosody conversion from neutral speech to emotional speech*. IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pages 1145 – 1154, July 2006.

[Taylor 2000] P. Taylor. *Analysis and Synthesis of Intonation using the Tilt Model*. Rapport technique, Centre for Speech Technology Research, University of Edinburgh, 2000.

[Taylor 2001] P. Taylor, A.W. Black et R. Caley. *Heterogeneous Relation Graphs as a Mechanism for Representing Linguistic Information*. Speech Communication, vol. 3, pages 153–174, January 2001.

[Toda 2001] T. Toda, H. Saruwatari et K. Shikano. *high quality voice conversion based on gaussian mixture model with dynamic frequency warping*. In Eurospeech, 2001.

[Trouvain 2003] J. Trouvain. *Segmenting Phonetic Units in Laughter*. In ICPhS, Barcelona, 2003.

[Umesh 1996] S. Umesh, L. Cohen, N. Marinovic et D. Nelson. *Frequency-Warping in Speech*. In ICSLP, volume 1, pages 414–417, Philadelphia, PA, 1996.

[van Santen 2003] J. van Santen, L. Black, G. Cohen, A. Kain, E. Klabbers, T. Mishra, J. de Villiers et X. Niu. *Applications of computer generated expressive speech for communication disorders*. In Eurospeech, 2003.

[Veaux 2008] C. Veaux, G. Beller et X. Rodet. *IrcamCorpusTools : an Extensible Platform for Spoken Corpora Exploitation*. In European Language Resources Association (ELRA), editeur, LREC2008, Marrakech, Morocco, may 2008.

[Vidrascu 2005] L. Vidrascu et L. Devillers. *Detection of Real-Life Emotions in Call Centers*. In Interspeech, 2005.

[Vieillard 2007] S. Vieillard, I. Peretz, N. Gosselin, S. Khalfa, L. Gagnon et B. Bouchard. *Happy, sad, scary and peaceful musical excerpts for research on emotions*. Cognition and Emotion, vol. 1, 2007.

[Villavicencio 2006] F. Villavicencio, A. Roebel et X. Rodet. *Improving LPC Spectral Envelope Extraction of Voiced Speech by True-Envelope Estimation*. In ICASSP, France, 2006.

[Vincent 2005a] D. Vincent, O. Rosec et T. Chonavel. *Estimation du signal glottique basée sur un modèle ARX*. In GRETSI, 2005.

[Vincent 2005b] D. Vincent, O. Rosec et T. Chonavel. *Estimation of LF glottal source parameters based on an ARX model*. In 9th European Conference on Speech Communication and Technology, pages 333–336, Lisbonne, Portugal, 2005.

[Vincent 2006] D. Vincent, O. Rosec et T. Chonavel. *Glottal closure instant estimation using an appropriateness measure of the source and continuity constraints*. In ICASSP, volume 1, pages 14–19, May 2006.

[Vincent 2007] D. Vincent, O. Rosec et T. Chonavel. *A new method for speech synthesis and transformation based on an arx-lf source-filter decomposition and hnm modeling*. In ICASSP, volume IV, pages 525–528, 2007.

[Vines 2005] A.W. Vines, C.L. Krumhansl, M.M. Wanderley, I.M.Dalca et D.J. Levitin. *Dimensions of emotion in expressive musical performance.* In The Neurosciences and Music II : From Perception to Performance, volume 1060, page 462â466, 2005.

[Wabnik 2005] S. Wabnik, G. Schuller, U. Kraemer et J. Hirschfeld. *Frequency Warping in Low Delay audio coding*. In ICASSP, 2005.

[Wheeldon 1994] L. Wheeldon. Cognition, volume 50, chapitre Do speakers have access to a mental syllabary, pages 239–259. 1994.

[Wightman 1995] C. Wightman et N. Campbell. *Improved labeling of prosodic structure*. In IEEE Trans. on Speech and Audio Processing, 1995.

[Wouters 2001] J. Wouters et M. Macon. *Control of spectral dynamics in concatenative speech synthesis*. In IEEE Transactions on Speech and Audio Processing, volume 9, pages 30–38, 2001.

[Yamagishi 2005] J. Yamagishi, K. Onishi, T. Masuko et T. Kobayashi. *Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis*. In IEICE Trans. on Inf. & Syst., volume E88-D, pages 503–509, March 2005.

[Yang 2006] Yang, Liu et Chen. *Music emotion classification : A fuzzy approach*. In ACMMM, 2006.

[Zajonc 1980] R. B. Zajonc. *Feeling and thinking : preferences need no inferences*. American Psychologist, vol. 35, pages 151–175, 1980.

[Zarader 1997] J.L. Zarader, B. Gas, J.C. Didiot et P. Sellem. *Neural Predictive Coding : application to phoneme recognition*. In ICONIP, 1997.

[Zellner 1998] B. Zellner. *Caractérisation du débit de parole en Français*. In Journées d'Etude sur la Parole, 1998.

[Zhan 1997] P. Zhan et M. Westphal. *Speaker Normalization Based on Frequency Warping*. In Proc. ICASSP '97, pages 1039–1042, Munich, Germany, 1997.

[Zhu 2004] Donglai Zhu et Kuldip K. Paliwal. *Product of Power Spectrum and Group Delay Function for Speech Recognition*. In ICASSP, 2004.

[Z.Inanoglu 2009] Z.Inanoglu et S. Young. *Data-driven emotion conversion in spoken English*. Speech Communication, vol. 51, no. 3, pages 268–283, 2009.

Liste des tableaux

Table des figures