Evaluation objective de la qualité vocale en contexte de conversation

Marie Guéguin

▶ **To cite this version:**

Marie Guéguin. Evaluation objective de la qualité vocale en contexte de conversation. Traitement du signal et de l'image [eess.SP]. Université Rennes 1, 2006. Français. NNT : . tel-00132550

HAL Id: tel-00132550
https://theses.hal.science/tel-00132550

Submitted on 21 Feb 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre: 3422

THÈSE

présentée

devant l'Université de Rennes 1

pour obtenir

le grade de : Docteur de l'Université de Rennes 1
Mention Traitement du Signal et Télécommunications

par

Marie Guéguin

Équipe d'accueil : Laboratoire Traitement du Signal et de l'Image (LTSI)
École Doctorale : MAthématiques, Télécommunications, Informatique, Signal, Systèmes, Electronique (MATISSE)
Composante universitaire : UFR Structure et Propriétés de la Matière

Titre de la thèse :

Évaluation objective de la qualité vocale en contexte de conversation

soutenue le 4 décembre 2006 devant la commission d'examen

M. :	Gérard	Bailly	Président
MM. :	Christophe	d'Alessandro	Rapporteurs
	Sebastian	Möller	
MM. :	Vincent	Barriac	Examinateurs
	Gérard	Faucon	
	Régine	Le Bouquin Jeannès	

Remerciements

Le travail matérialisé par ce mémoire de thèse a été financé dans le cadre d'une bourse CIFRE. Mes travaux se sont déroulés à la fois dans les locaux de France Télécom R&D situés à Lannion et dans ceux du Laboratoire Traitement du Signal et de l'Image (LTSI) situés à Rennes.

J'adresse ma gratitude à Monsieur Gérard Bailly pour l'honneur qu'il m'a fait en présidant le jury de cette thèse. Mes très vifs remerciements vont à Messieurs Christophe d'Alessandro et Sebastian Möller pour avoir accepté d'être les rapporteurs de ma thèse et pour leur intérêt dans mon travail.

Je ne saurais trop remercier mes directeurs de thèse et encadrants : Régine Le Bouquin Jeannès et Gérard Faucon pour le LTSI, Vincent Barriac et Valérie Gautier-Turbin pour France Télécom R&D. Ils m'ont accordé toute leur confiance et ont su me soutenir dans les moments de doute. Régine, Gérard, vos qualités humaines et scientifiques ont fait que ces trois années de thèse passées sous votre direction ont été très enrichissantes pour moi.

Je remercie toutes les personnes qui ont collaboré à ce travail, notamment lors de la mise en place (vraiment pas simple...) des tests subjectifs à France Télécom R&D et qui leur a demandé du temps et de la patience : Laetitia Gros, Philippe Saliou, Caroll Rattazzi, Nicolas Le Faucheur, Patrice Odermatt et Christophe Le Maguer. Merci à l'ensemble de l'équipe MOV, qui m'a supportée avant ma soutenance (et félicitée après !). J'ai une pensée particulière pour André Gilloire, qui m'a permis d'effectuer cette thèse et dont les encouragements m'ont touchée.

Mes remerciements les plus chaleureux s'adressent aux membres du LTSI pour l'ambiance joyeuse (mais néanmoins studieuse !) qui y règne et les pauses café animées. J'exprime toute mon amitié aux thésards (et associés) présents et passés du LTSI, que j'ai côtoyés de loin ou de plus près : Delphine, Antoine, Virginie, Vincent, Fabienne, Manu & Delphine, Amar, Cemil, Oscar, Phong, qui, grâce à leur bonne humeur et à leur amitié, m'ont permis de passer trois années inoubliables.

Comment ne pas remercier ma famille, pour m'avoir encouragée au cours de cette thèse et, au-delà, pour son amour qui dure depuis longtemps déjà... Merci à mes parents, Philippe & Isabelle, pour m'avoir appris la persévérance, la ténacité et tout le reste ; à mes deux frères adorés, Renaud (associé à Aurélia) & Maxime, à qui je promets d'être plus disponible à présent ; à ma grand-mère, mes oncles, tantes, cousins et cousines qui m'ont toujours témoigné de leur soutien et dont certains ont pu faire le déplacement jusqu'à Rennes en ce lundi de décembre 2006... Un énorme merci à Morgane pour son amitié infaillible !

Enfin, un merci très spécial à Julien pour être là.

Table des matières

Remerciements **1**

Table des matières **3**

Glossaire **9**

Introduction **13**

1 État de l'art **17**
Introduction . . . 17
1.1 Qualité vocale dans les télécommunications . . . 17
1.1.1 Considérations sur la qualité vocale . . . 17
1.1.1.1 Subjective par nature . . . 17
1.1.1.2 Critères de qualité . . . 18
1.1.1.3 Contexte . . . 18
1.1.1.4 Synthèse . . . 21
1.1.2 Évolution des systèmes de télécommunications . . . 21
1.1.2.1 Téléphonie classique . . . 22
1.1.2.2 Systèmes numériques . . . 22
1.1.2.3 Systèmes mobiles . . . 22
1.1.2.4 Réseaux en mode paquet . . . 23
1.1.2.5 Synthèse . . . 23
1.2 Évaluation subjective de la qualité vocale . . . 23
1.2.1 Choix des sujets . . . 23
1.2.2 Méthodes normalisées . . . 24
1.2.2.1 Essais d'opinion d'écoute . . . 24
1.2.2.2 Tests de parole et d'écoute . . . 25
1.2.2.3 Essais d'opinion de conversation . . . 25
1.2.2.4 Synthèse . . . 26
1.2.3 Effets subjectifs des différentes dégradations sur la qualité vocale . . . 27
1.2.3.1 Échos . . . 27
1.2.3.2 Délai . . . 29
1.2.3.3 Distorsion de la parole due au codage . . . 30
1.2.3.4 Bruits . . . 32
1.2.3.5 Pertes de paquets . . . 33
1.2.3.6 Distorsion de l'effet local . . . 36
1.2.3.7 Variations dans le temps des dégradations . . . 36
1.2.3.8 Double parole . . . 37
1.2.3.9 Dispositifs de traitement du signal . . . 37
1.2.3.10 Synthèse . . . 38
1.2.4 Limites de l'évaluation subjective . . . 38
1.3 Modèles objectifs de la qualité vocale . . . 38

1.3.1 Modèles paramétriques . . . 39
1.3.1.1 Modèle E . . . 39
1.3.1.2 Modèle CCI . . . 40
1.3.1.3 Modèle P.564 . . . 40
1.3.1.4 Avantages et limites . . . 40
1.3.2 Modèles basés sur les signaux avec référence . . . 41
1.3.2.1 Transformation par représentation interne . . . 41
1.3.2.2 Modèle PESQ . . . 44
1.3.2.3 Modèle PESQM . . . 44
1.3.2.4 Avantages et limites . . . 45
1.3.3 Modèles basés sur les signaux sans référence . . . 45
1.3.4 Évaluation des mesures objectives de la qualité vocale . . . 47
Conclusion . . . 47

2 Problématique, objectifs et méthode proposée 49
Introduction . . . 49
2.1 Problématique . . . 49
2.2 Objectifs . . . 50
2.3 Méthode proposée . . . 51
2.3.1 Partie intégration . . . 51
2.3.2 Partie mesure . . . 52
Conclusion . . . 53

3 Construction du modèle 55
Introduction . . . 55
3.1 Méthodologie de test proposée . . . 55
3.1.1 Protocole . . . 55
3.1.2 Déroulement des tests . . . 56
3.1.3 Choix des conditions de test . . . 56
3.1.4 Montage expérimental et enregistrement . . . 56
3.1.5 Analyse des résultats subjectifs et rejet des sujets aberrants . . . 57
3.2 Tests subjectifs réalisés . . . 57
3.2.1 Test 1 : délai et écho . . . 57
3.2.1.1 Objectifs . . . 57
3.2.1.2 Conditions et facteurs expérimentaux . . . 58
3.2.1.3 Analyse des résultats . . . 59
3.2.1.4 Synthèse . . . 63
3.2.2 Test 2 : pertes de paquets et bruit . . . 64
3.2.2.1 Objectif . . . 64
3.2.2.2 Conditions et facteurs expérimentaux . . . 64
3.2.2.3 Analyse des résultats . . . 66
3.2.2.4 Synthèse . . . 70
3.2.3 Test 3 : bruit . . . 70
3.2.3.1 Objectif . . . 70
3.2.3.2 Conditions et facteurs expérimentaux . . . 71
3.2.3.3 Analyse des résultats . . . 72
3.2.3.4 Synthèse . . . 74
3.2.4 Test 4 : écho, délai et pertes de paquets . . . 74
3.2.4.1 Objectif . . . 74
3.2.4.2 Conditions et facteurs expérimentaux . . . 74
3.2.4.3 Analyse des résultats . . . 75
3.2.4.4 Synthèse . . . 80

3.3 Relation entre les différentes composantes de la qualité vocale 81
3.3.1 Test 1 : délai et écho 82
3.3.2 Test 2 : pertes de paquets et bruit 84
3.3.3 Test 3 : bruit 85
3.3.4 Test 4 : écho, délai et pertes de paquets 87
3.3.5 Détection des dégradations 89
3.3.6 Tous tests 90
3.3.6.1 Apprentissage 90
3.3.6.2 Validation 91
Conclusion 93

4 Outils de mesure **95**
Introduction 95
4.1 Optimisation du modèle de qualité de locution PESQM 96
4.1.1 Étude préliminaire de PESQM sur deux tests de locution de la littérature 96
4.1.2 Test de locution 98
4.1.2.1 Protocole 98
4.1.2.2 Analyse des résultats 100
4.1.2.3 Enregistrement des signaux de test 105
4.1.2.4 Vérification de la reproductibilité des notes subjectives entre les deux sessions 106
4.1.3 Étude de PESQM sur les résultats de notre test de locution 106
4.1.4 Optimisation et validation de PESQM 109
4.1.4.1 Optimisation 109
4.1.4.2 Choix des paramètres optimaux 109
4.1.4.3 Version optimisée de PESQM appliquée au test de locution . 112
4.1.4.4 Synthèse 113
4.2 Découpage des signaux de conversation 113
Conclusion 115

5 Résultats et performances du modèle objectif **119**
Introduction 119
5.1 Application à des signaux de test 120
5.1.1 Performances du modèle objectif de qualité d'écoute (PESQ) 120
5.1.2 Performances du modèle objectif de qualité de locution (PESQM) . . . 121
5.1.3 Performances du modèle objectif de qualité de conversation (CONV) . 123
5.1.4 Performances du modèle objectif de qualité de conversation (CONV) avec détection du bruit 125
5.1.5 Synthèse 126
5.2 Application à des signaux de conversation 127
5.2.1 Performances du modèle objectif de qualité d'écoute (PESQ) 128
5.2.2 Performances du modèle objectif de qualité de locution (PESQM) . . . 128
5.2.3 Performances du modèle objectif de qualité de conversation (CONV) . 129
5.2.4 Performances du modèle objectif de qualité de conversation (CONV) avec détection du bruit 130
5.2.5 Synthèse 131
5.3 Étude de l'interactivité 132
5.3.1 Motivations 132
5.3.2 Application 134
5.3.3 Synthèse 136
Conclusion 136

Conclusions et perspectives **139**

A Notions de psychoacoustique **143**
A.1 Physique du phénomène sonore . . . 143
A.2 Capacités sensorielles et dimensions de la perception auditive . . . 143
A.2.1 Bandes critiques . . . 143
A.2.2 Masquage . . . 143
A.2.3 Audiogramme masqué . . . 144
A.3 Perception de l'intensité acoustique . . . 144
A.4 Perception de la hauteur . . . 145
A.5 Échelles naturelles de la membrane basilaire . . . 145

B Modèle objectif PESQ **147**
B.1 Domaine d'application et limitations de PESQ . . . 147
B.2 Principe . . . 148
B.2.1 Échelonnement et alignement temporel . . . 148
B.2.1.1 Échelonnement du niveau . . . 148
B.2.1.2 Filtrage du système IRS . . . 149
B.2.1.3 Alignement temporel . . . 149
B.2.2 Modèle psychoacoustique . . . 150
B.2.2.1 Initialisations et calibrations . . . 150
B.2.2.2 Transformation temps-fréquence . . . 151
B.2.2.3 Prédistorsion et densité de puissance fondamentale . . . 151
B.2.2.4 Compensations . . . 152
B.2.2.5 Densité de sonie . . . 153
B.2.2.6 Densité de perturbation . . . 153
B.2.2.7 Traitement de l'asymétrie . . . 153
B.2.2.8 Accentuation des parties de silence . . . 153
B.2.2.9 Intégration en temps et fréquence . . . 153
B.2.2.10 Calcul du score PESQ . . . 153
B.2.3 Performances . . . 154

C Modèle objectif PESQM **155**
C.1 Principe . . . 155
C.2 Équations . . . 156
C.2.1 Initialisations et calibrations . . . 156
C.2.2 Fenêtrage et densité spectrale de puissance . . . 157
C.2.3 Prédistorsion et densité de puissance fondamentale . . . 158
C.2.4 Étalement dans le domaine fréquentiel . . . 159
C.2.5 Densité de sonie . . . 159
C.2.6 Densité de perturbation due au bruit . . . 160
C.2.7 Suppression du bruit . . . 160
C.2.8 Calcul du score PESQM . . . 160
C.3 Performances . . . 161

D Régression linéaire **163**
D.1 Modèle . . . 163
D.2 Suppositions . . . 163
D.3 Statistiques . . . 164
D.3.1 Sommes des carrés (SC) . . . 164
D.3.2 Coefficient de détermination . . . 164
D.3.3 Coefficient de détermination ajusté . . . 164
D.3.4 Test de Fisher . . . 164

D.3.5 Table d'analyse de variance (ANOVA) . . . 164
D.3.6 Intervalles de confiance des coefficients estimés . . . 165
D.3.7 Test de nullité d'un coefficient . . . 165
D.4 Analyse des résidus . . . 165
D.5 Multicolinéarité . . . 166
D.6 Sélection des régresseurs . . . 166
D.6.1 Backward elimination . . . 166
D.6.2 Forward selection . . . 166
D.6.3 Stepwise regression . . . 167
D.7 Méthode de bootstrap . . . 167

E Matériel de test **169**

F Liste de publications associées **171**

Bibliographie **178**

Glossaire

Vocabulaire spécifique

Ce travail de recherche a été effectué dans le cadre d'une convention CIFRE entre la division R&D de France Télécom et le laboratoire LTSI de l'Université de Rennes 1. L'expérience a montré qu'il existe parfois un écart considérable entre la précision et le sens donnés à certains mots ou expressions dans le monde des télécommunications et le monde scientifique. Ce travail apportant une réponse à une problématique d'opérateur de télécommunications, il a semblé naturel de conserver les expressions communément utilisées par les acteurs du monde des télécommunications. Cependant, il s'avère indispensable avant d'aller plus avant de préciser ou clarifier le sens de certaines expressions utilisées tout au long de cet ouvrage et identifiées comme potentiellement confuses :

Qualité vocale	Expression utilisée pour désigner la qualité de transmission téléphonique d'un signal de parole. L'étude de la qualité vocale dans notre contexte n'aborde pas par conséquent la notion d'intelligibilité du signal de parole ou encore de qualité de la voix.
Évaluation subjective	Évaluation basée sur la perception humaine, réalisée par la collecte d'opinions de sujets ou utilisateurs des services évalués.
Évaluation objective	Évaluation réalisée par une mesure instrumentale.
Qualité de locution	Expression utilisée pour désigner la qualité perçue par un sujet pendant les phases de parole qui lui sont propres (*i.e.* aucune autre personne alors en phase de parole).

Abréviations

ACR	Évaluation par catégories absolues (*Absolute Category Rating*)
ADPCM	Modulation par impulsions et codage différentiel adaptatif (*Adaptive Differential Pulse Code Modulation*)
ANOVA	Analyse de variance (*ANalysis Of VAriance*)
CCR	Évaluation par catégories de comparaison (*Comparison Category Rating*)
CELP	Prédiction linéaire avec excitation par code (*Codebook Excited Linear Prediction*)
CODEC	COdeur/DECodeur
dB	Décibel
DAV	Détection d'Activité Vocale

DCR	Évaluation par catégories de dégradation (*Degradation Category Rating*)
DPCM	Modulation par impulsions et codage différentiel (*Differential Pulse Code Modulation*)
DSL	Ligne d'abonné numérique (*Digital Subscriber Line*)
DSLA	Analyseur numérique de niveau de parole (*Digital Speech Level Analyser*)
ETSI	Institut européen des normes de télécommunication (*European Telecommunications Standards Institute*)
FFT	Transformée de Fourier rapide (*Fast Fourier Transform*)
GSM	Système mondial de communications mobiles (*Global System for Mobile communications*)
HATS	Simulateur de tête et de torse (*Head And Torso Simulator*)
INMD	Dispositif de mesure en service sans intrusion (*In-service, Non-intrusive Measurement Device*)
IP	Protocole Internet (*Internet Protocol*)
IRS	Système de référence intermédiaire (*Intermediate Reference System*)
MOS	Note moyenne d'opinion (*Mean Opinion Score*)
MPE	Multi-Pulse Excitation
PABX	Autocommutateur privé (*Private Automatic Branch eXchange*)
PCM	Modulation par impulsions et codage (*Pulse Code Modulation*)
PLC	Masquage des pertes de paquets (*Packet Loss Concealment*)
QoS	Qualité de service (*Quality of Service*)
RNIS	Réseau Numérique à Intégration de Services
RPE	Excitation par impulsions régulières (*Regular Pulse Excitation*)
RSB	Rapport Signal-à-Bruit
RTC	Réseau Téléphonique Commuté
RTCP	Protocole de commande de transfert en temps réel (*Real-time Transfert Control Protocole*)
RTCP-XR	Rapports approfondis sur le protocole RTCP (*Real-time Transfert Control Protocol Extended Reports*)
RTP	Protocole de transfert en temps réel (*Real-Time Protocol*)
SCT	Essai de conversation bref (*Short Conversation Test*)
SPL	Niveau de pression sonore (*Sound Pressure Level*)
TELR	Equivalent à la sonie pour l'écho pour le locuteur (*Talker Echo Loudness Rating*)
UIT	Union Internationale des Télécommunications
UIT-T	Secteur normalisation des télécommunications de l'UIT
UMTS	Système universel de télécommunications mobiles (*Universal Mobile Telecommunications System*)

VoIP Téléphonie utilisant le protocole Internet (*Voice over Internet Protocol*)

WEPL Affaiblissement pondéré du trajet des courants d'écho (*Weighted Echo Path Loss*)

Définitions

dB SPL Décibel de l'intensité acoustique, défini par le rapport de la puissance par unité de surface du son que l'on mesure et une puissance par unité de surface de référence.

dB(A) Décibel SPL avec pondération A, prenant en compte la sensibilité de l'oreille humaine à certaines fréquences.

dBov Décibel par rapport au point de surcharge d'un système numérique.

dBr Niveau relatif de puissance, en décibels.

dBm Rapport en décibels d'un niveau de puissance à une puissance de référence de 1 mW.

dBm0 Niveau absolu du signal en décibels rapporté à un point de niveau relatif zéro.

dBm0p dBm mesuré au point zéro dBr, à pondération psophométrique.

Introduction

Les systèmes de télécommunications sont en constante évolution depuis plusieurs années : nous avons assisté à l'émergence de nouveaux types de transmission, tels que les réseaux mobiles (GSM, *Global System for Mobile communications*, et UMTS, *Universal Mobile Telecommunications System*) et les réseaux de type paquet (VoIP, *Voice over Internet Protocol*). Ces nouvelles technologies sont en pleine expansion du fait de la valeur ajoutée qu'elles apportent aux utilisateurs par rapport à la téléphonie classique, telle que la mobilité, ou la possibilité de transmettre non seulement la voix, mais aussi des données et du contenu multimédia, ou encore le coût réduit des appels longue distance. Cependant, contrairement à la qualité de la voix transmise sur le réseau téléphonique commuté (RTC) relativement stable et prévisible, la qualité de service (QoS, *Quality of Service*) de ces nouvelles technologies est généralement non garantie. En effet, elles sont non seulement sujettes à la plupart des dégradations rencontrées avec le RTC (écho, délai, distorsion de l'effet local, bruits, etc.), mais introduisent de nouvelles dégradations (distorsion de la parole due au codage, délais augmentés par le traitement numérique), dont certaines sont non linéaires (délai variable appelé « gigue » et pertes de paquets dans les réseaux IP, bruits de fond non stationnaires dans les réseaux mobiles).

Afin de satisfaire leurs clients et de leur offrir la meilleure QoS possible, les opérateurs de télécommunications se doivent de contrôler la qualité perçue par les utilisateurs de leurs services, et doivent pour cela l'évaluer. Les méthodes subjectives, faisant appel à des participants humains qui testent un système dans différentes conditions réelles d'utilisation définies par l'expérimentateur, restent la meilleure solution pour évaluer la qualité perçue par les utilisateurs. Bien que ces méthodes subjectives sont, par définition, le seul moyen d'atteindre le jugement des utilisateurs, les opérateurs de télécommunications cherchent à éviter le recours à de telles méthodes, du fait du coût et du temps qu'elles demandent. Ainsi, des méthodes objectives plus poussées que les mesures objectives simples telles que le rapport signal-à-bruit (RSB) et l'erreur quadratique moyenne (EQM) ont été développées. Elles sont construites afin d'être corrélées avec les résultats de tests subjectifs et ainsi constituent un moyen de substitution aux méthodes subjectives. Or, parmi tous les modèles objectifs existants, aucun ne peut prédire efficacement la qualité perçue dans le contexte le plus couramment expérimenté par l'utilisateur de services téléphoniques : la conversation. En effet, la quasi-totalité des modèles existants s'intéresse au contexte d'écoute (*i.e.* situation où le sujet écoute un message vocal, sans parler), qui peut être dégradé par la distorsion due au codage de la parole, le bruit présent dans le signal, la perte d'information, le niveau du signal ou encore le bruit ambiant autour du sujet. Ces modèles ne prennent cependant pas en compte d'autres dégradations rencontrées en contexte de conversation, comme l'écho ou le délai.

L'objectif général de cette thèse est double. Tout d'abord, il s'agit de proposer un modèle objectif de la qualité de conversation, basé sur une analyse des signaux échangés durant la communication testée. Afin de construire un modèle objectif, il est indispensable de disposer de données subjectives. Dans la littérature, la majorité des tests subjectifs concerne le contexte d'écoute et les tests subjectifs en contexte de conversation sont peu nombreux. Un autre objectif de cette thèse consiste donc à concevoir et mettre en œuvre plusieurs tests sub-

jectifs afin d'étudier l'impact des dégradations rencontrées dans le contexte de conversation sur la qualité vocale perçue.

La conversation peut être décrite, du point de vue d'un interlocuteur, comme une alternance des rôles d'auditeur et de locuteur introduisant de l'interaction entre les interlocuteurs. Partant de cette description de la conversation, la méthode proposée dans la thèse repose sur l'hypothèse que la qualité de conversation peut être décomposée selon trois dimensions : la qualité d'écoute, la qualité de locution et la qualité d'interaction. Le modèle est divisé en deux parties : la « partie intégration » combine les notes de qualités d'écoute, de locution et d'interaction pour estimer une note de qualité de conversation, et la « partie mesure » fournit les notes objectives de qualité à la partie intégration en se basant sur les modèles existants de qualité vocale dans les différents contextes. Ces deux parties sont différenciées pour obtenir un modèle fonctionnant pour plusieurs applications selon les modèles utilisés dans la partie mesure, la partie intégration restant commune à toutes les applications. Afin de vérifier l'hypothèse de décomposition de la qualité de conversation et de construire la partie intégration du modèle, quatre tests subjectifs, étudiant différentes conditions de dégradation, sont mis en œuvre. Une nouvelle méthodologie de test subjective est proposée pour évaluer au cours d'un même test les qualités de conversation, d'écoute et de locution, dans les mêmes conditions de dégradation. La qualité d'interaction est difficile à évaluer puisqu'elle ne fait l'objet d'aucune méthodologie de test normalisée à l'Union Internationale des Télécommunications (UIT) : elle est représentée ici par la valeur du délai présent dans la communication testée. L'hypothèse de décomposition de la note de qualité de conversation en trois dimensions (note de qualité d'écoute, note de qualité de locution, valeur du délai représentant la qualité d'interaction) est vérifiée sur chacun de ces tests. Les résultats de tests subjectifs sont utilisés pour déterminer la partie intégration du modèle objectif, *i.e.* la combinaison des trois dimensions à partir des notes subjectives d'écoute, des notes subjectives de locution et des valeurs de délai.

Le premier chapitre de ce mémoire décrit tout d'abord la qualité vocale dans les télécommunications et le besoin de son évaluation créé par l'évolution des systèmes de téléphonie. Après avoir présenté les méthodes d'évaluation subjective de la qualité vocale, un état de l'art des modèles objectifs existants est proposé. Les avancées effectuées dans ce domaine et les limites des méthodes actuelles sont notamment exposées. Le manque d'un outil objectif dans le contexte de conversation est mis en exergue.

La problématique posée par l'élaboration d'un tel modèle objectif est présentée dans le chapitre 2. Après avoir fixé les objectifs à atteindre par ce modèle, la méthode proposée dans la thèse pour construire un modèle objectif de la qualité vocale de conversation est décrite.

La description des quatre tests subjectifs mis en œuvre pendant la thèse, leur analyse et la construction de la partie intégration du modèle à partir des résultats subjectifs font l'objet du chapitre 3.

Pour fonctionner, la partie mesure du modèle objectif a besoin d'un modèle de la qualité d'écoute et d'un modèle de la qualité de locution. Dans la littérature et à l'UIT, de nombreux modèles de la qualité d'écoute sont disponibles (*e.g.* [UIT-T Rec. P.862 2001], [UIT-T Rec. P.563 2004] et [UIT-T Rec. P.564 2006]), mais un seul modèle proposé dans [Appel et Beerends 2002] existe dans le contexte de locution. La première partie du chapitre 4 concerne l'implémentation et l'optimisation de ce modèle objectif de locution sur les notes subjectives recueillies lors d'un test de locution mis en œuvre pendant la thèse. La seconde partie de ce chapitre propose un outil de traitement des signaux, basé sur une détection d'activité vocale, permettant l'utilisation du modèle objectif de conversation sur plusieurs types de signaux (de test ou réels).

Le modèle proposé est appliqué aux signaux de parole enregistrés pendant les quatre tests subjectifs dans différentes conditions de dégradation. Les résultats obtenus et les performances du modèle sont présentés et discutés dans le chapitre 5. Comme le chapitre 1 met en évidence

un impact du délai sur la qualité de conversation variable en fonction de l'interactivité de la communication testée, le chapitre 5 propose une étude, préliminaire, de l'interactivité effectuée à partir des signaux de conversation enregistrés durant les tests subjectifs.

Enfin, la conclusion reprend les points essentiels développés dans les chapitres précédents et ouvre sur différentes perspectives.

Chapitre 1

État de l'art

Introduction

Le langage parlé est l'une des fonctions essentielles qui permettent aux humains de communiquer, *i.e.* de transmettre un message d'un émetteur à un récepteur. Pour que le message soit correctement compris par le récepteur, il est nécessaire que la qualité du signal de parole transmis soit correcte. Si ceci est valable lors d'une communication face à face, cela l'est d'autant plus lors d'une communication téléphonique au cours de laquelle le signal de parole est la seule source d'information disponible pour les deux interlocuteurs. Il est donc primordial que les systèmes de communication garantissent une qualité de parole transmise suffisante pour leurs utilisateurs. Or, les réseaux et systèmes de télécommunications ont connu une (r)évolution très rapide depuis deux décennies, aboutissant à une qualité de parole non toujours garantie en comparaison avec la téléphonie classique. Pour les opérateurs de télécommunications, la qualité vocale est donc devenue un enjeu majeur.

Dans la première partie du présent chapitre, nous décrirons tout d'abord la qualité de la parole transmise par un système de télécommunications, que nous dénommerons « qualité vocale » dans la suite de la thèse. Ensuite, nous présenterons l'évolution des systèmes de télécommunications depuis l'invention du téléphone jusqu'à aujourd'hui et leur impact sur la qualité vocale.

Dans les deuxième et troisième parties de ce chapitre, les méthodes d'évaluation subjective et objective de la qualité vocale des systèmes de télécommunications seront décrites, respectivement.

1.1 Qualité vocale dans les télécommunications

L'objectif est de donner une description de la qualité vocale rencontrée dans les systèmes de télécommunications, afin d'en appréhender les caractéristiques et de comprendre comment elle est jugée.

1.1.1 Considérations sur la qualité vocale

1.1.1.1 Subjective par nature

De façon générale, la qualité dépend de la personne qui la juge. La qualité vocale est donc une notion complexe à définir du fait de sa forte subjectivité. Elle dépend de l'interprétation que fait chacun d'un événement sonore donné. Cette subjectivité entre en jeu tout d'abord dans la perception de l'événement sonore, puis dans la description de cet événement sonore. Möller [Möller 2000b], se basant sur les travaux de Blauert [Blauert 1997], donne une représentation schématique de l'auditeur lors d'une expérience auditive, reproduite dans la figure 1.1. Quand un événement sonore se produit, le système auditif humain analyse le signal tant du point

de vue de son contenu que de sa forme. Ainsi, si l'événement sonore correspond à de la parole, le contenu (*i.e.* l'information sémantique) et la forme (*i.e.* le signal acoustique) sont analysés. Bien qu'ici la qualité vocale se rapporte à la qualité de la forme du signal de parole (*i.e.* le signal acoustique), l'interprétation de la qualité vocale est influencée par le contenu du signal acoustique, dans une mesure qui dépend de chaque individu (facteurs individuels). La subjectivité entre ensuite en jeu dans la description de l'événement sonore, donc dans le jugement de la qualité vocale de celui-ci. Ce jugement dépend des attentes et de l'expérience passée de chacun, qui constituent la référence interne à laquelle chaque nouvel événement sonore est comparé. Ainsi, Jekosch [Jekosch 2000] décrit la qualité vocale comme le résultat d'un processus de perception et de jugement, durant lequel le sujet établit une relation entre ce qui est perçu (*i.e.* l'événement sonore) et ce qui est désiré ou attendu (*i.e.* la référence interne). De ce fait, la qualité vocale n'existe pas dans l'absolu, mais est attribuée par l'auditeur.

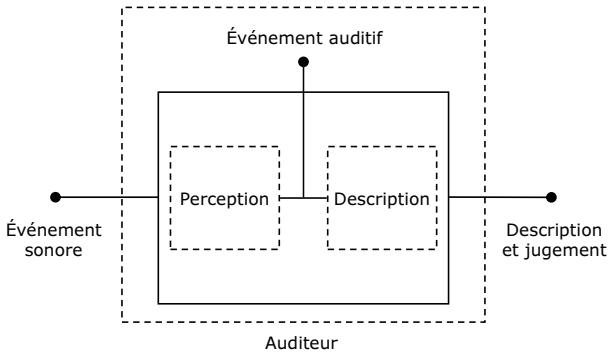


Figure 1.1 : *Représentation schématique de l'auditeur lors d'une expérience auditive (d'après [Blauert 1997])*

1.1.1.2 Critères de qualité

La qualité vocale est un phénomène multidimensionnel [Preminger et van Tasell 1995]. En effet, elle peut être évaluée selon différents critères de qualité. Les deux principaux critères sont la sonie et l'intelligibilité (*i.e.* le niveau et la compréhensibilité du signal de parole, respectivement), qui permettent à l'auditeur d'entendre et de comprendre le message du locuteur. D'autres critères de qualité ont ensuite été étudiés, tels que l'agrément (*i.e.* la satisfaction globale de l'usager par rapport au service utilisé), l'effort d'écoute, l'impression globale, la fidélité ou le naturel de la voix. Ainsi la qualité vocale peut être étudiée vis-à-vis d'une multitude de critères de qualité, tous l'influençant de façon plus ou moins importante, et pouvant, de surcroît, être dépendants les uns des autres. Ces méthodes d'évaluation de la qualité vocale sont dites « analytiques » [IEEE 1969], considérant la qualité vocale comme un phénomène multidimensionnel. Or, d'après Jekosch [Jekosch 1993], il semble très complexe, voire irréaliste, d'explorer toutes les dimensions de la qualité vocale. Pour des questions de simplicité, la qualité vocale sera donc souvent représentée par un scalaire. Ce sont les méthodes dites « utilitaires » [IEEE 1969], considérant la qualité vocale comme un phénomène unidimensionnel. Si les méthodes analytiques présentent un intérêt pour comprendre comment le jugement de la qualité vocale est construit par le sujet, les méthodes utilitaires sont actuellement les plus courantes, en particulier pour la modélisation de la qualité vocale. Il est en effet moins complexe de modéliser un score scalaire que plusieurs.

1.1.1.3 Contexte

La perception de la qualité vocale dépend du contexte dans lequel est placée la personne qui juge [Gros 2001]. Il existe trois contextes dans lesquels une personne juge la qualité vocale : le contexte d'écoute, le contexte de locution et le contexte de conversation. Nous présenterons

dans ce paragraphe chacun des trois contextes, ainsi que les dégradations qui affectent chacun d'entre eux plus particulièrement.

Contexte d'écoute Comme son nom l'indique, le contexte d'écoute correspond à la situation où le sujet écoute un message vocal, sans parler. Dans la vie courante, les utilisateurs de systèmes de télécommunications peuvent être placés dans ce contexte lorsque, par exemple, ils consultent leur répondeur téléphonique. Un tel contexte peut être dégradé par la distorsion due au codage de la parole, le bruit présent dans le signal, la perte d'information, le niveau du signal ou encore le bruit ambiant autour du sujet. Ces différentes dégradations diminuent la qualité vocale en affectant l'intelligibilité, le naturel de la voix ou la sonie, rendant difficile la compréhensibilité du message vocal par le sujet.

Contexte de locution Dans le contexte de locution, le sujet parle, sans recevoir de message vocal d'un interlocuteur. Un utilisateur de systèmes de télécommunications le rencontre par exemple lorsqu'il laisse un message vocal sur le répondeur téléphonique d'un correspondant. Les dégradations affectant ce contexte sont essentiellement relatives à la propre voix de l'utilisateur, *i.e.* l'effet local et l'écho, et au bruit ambiant. L'effet local désigne les sons qui sont captés par le microphone du combiné et sont transmis à l'écouteur du même combiné, avec un délai faible (quelques millisecondes) [ETSI ETR 250 1996]. La gêne causée par l'effet local est fonction de son niveau et de sa distorsion. L'écho est produit soit par un couplage acoustique soit par un couplage électrique. La perception de l'écho dépend principalement de deux paramètres : le délai et l'atténuation de l'écho [UIT-T Rec. G.131 1996]. Les réseaux de télécommunications incluent généralement des annuleurs d'écho, estimant le signal d'écho pour le retirer du signal reçu. Si les dégradations rencontrées dans le contexte d'écoute affectent de façon évidente la qualité vocale en diminuant la compréhensibilité du message reçu, celles rencontrées dans le contexte de locution la dégradent de façon moins évidente, mais peuvent être gênantes pour le locuteur. Quand nous parlons, nous percevons notre propre signal de parole (rétroaction), transmis de notre bouche à notre oreille à la fois par l'air et par la conduction osseuse. Ce signal de rétroaction nous sert pour adapter notre volume, notre timbre, et pour contrôler notre articulation [Jones et Munhall 2002]. Ainsi, l'effet local en lui-même n'est pas gênant et est même utile au locuteur. Par contre, une distorsion de l'effet local va avoir pour effet de distordre le signal de rétroaction et rendre plus difficile la production de la parole par le locuteur. L'écho va perturber le locuteur en lui retournant le signal qu'il vient de prononcer atténué et retardé. En présence de bruit, le signal de rétroaction va être dégradé par le bruit, le locuteur va donc augmenter le volume de sa voix pour compenser cette perte d'information. Cet effet a été décrit par Lombard en 1911 [Lombard 1911] et est désormais connu sous le nom d'« effet Lombard ». Ces différentes dégradations vont obliger le locuteur à faire des efforts pour articuler et ainsi se faire comprendre de son correspondant.

Contexte de conversation Dans la vie quotidienne, nous sommes assez peu souvent placés, durant une communication téléphonique, dans un contexte d'écoute pur ou dans un contexte de locution pur, mais plus souvent dans un contexte de conversation. Dans [Richards 1973], Richards est le premier à proposer une étude et une description de la conversation (face à face et via un système de télécommunications), d'un point de vue théorique et mathématique. Lors d'une conversation (face à face ou via un système de télécommunications), les participants échangent des informations tour à tour. Ils adoptent ainsi alternativement les rôles de locuteur et d'auditeur, cette alternance de rôles introduisant de l'interaction entre les participants. Richards insiste de plus sur le fait que ces rôles ne sont pas toujours mutuellement exclusifs. Il arrive en effet régulièrement dans une conversation que les participants adoptent un rôle de locuteur en même temps, *i.e.* les participants parlent simultanément (double parole), ou au contraire qu'ils ne soient ni auditeurs ni locuteurs en même temps, *i.e.* qu'aucun des participants ne parle (silence mutuel). Partant de ce constat, Richards propose un modèle à 4

états de la conversation entre deux interlocuteurs, du point de vue d'un participant, reproduit dans la figure 1.2.

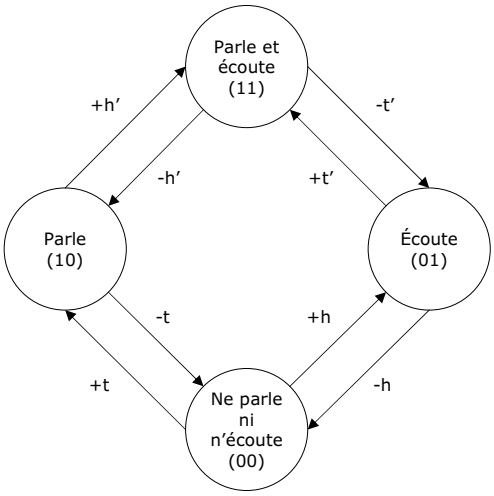


Figure 1.2 : *États et événements dans une conversation tels que perçus par un participant (d'après [Richards 1973])*

Les quatre états sont les suivants :

00 le participant ne parle ni n'entend son interlocuteur parler,

01 le participant écoute son interlocuteur parler, mais ne parle pas,

10 le participant parle, mais n'entend pas son interlocuteur,

11 le participant parle en écoutant son interlocuteur parler.

La transition entre les quatre états est commandée par les huit événements conversationnels suivants :

+t le participant commence à parler alors qu'il n'entend pas son interlocuteur parler,

+t' le participant commence à parler alors qu'il entend son interlocuteur parler,

-t le participant cesse de parler alors qu'il n'entend pas son interlocuteur parler,

-t' le participant cesse de parler alors qu'il entend son interlocuteur parler,

+h le participant commence à entendre son interlocuteur parler alors qu'il ne parle pas lui-même,

+h' le participant commence à entendre son interlocuteur parler alors qu'il parle lui-même,

-h le participant cesse d'entendre son interlocuteur parler alors qu'il ne parle pas lui-même,

-h' le participant cesse d'entendre son interlocuteur parler alors qu'il parle lui-même.

D'après cette description, la conversation, du point de vue d'un participant, est composée de périodes d'écoute et de périodes de locution, qui alternent en fonction de l'interaction avec l'interlocuteur. Vis-à-vis de la qualité vocale, le contexte de conversation est donc affecté par les dégradations rencontrées dans le contexte d'écoute et celles rencontrées dans le contexte de locution. À ces dégradations (présentées précédemment), s'ajoutent les dégradations affectant l'interaction de la conversation, *i.e.* le délai et la dégradation due aux périodes de double parole. Le délai va avoir pour effet de diminuer l'interaction de la conversation en augmentant les périodes de double parole et de silence mutuel, comme l'illustre la figure 1.3. Il peut se produire une dégradation de la qualité lors des périodes de double parole en présence d'écho, car l'annulation d'écho est moins efficace dans les conditions de double parole [UIT-T Rec. G.131 1996]. Ainsi, cela risque de générer une mauvaise annulation d'écho, qui peut créer une gêne considérable.

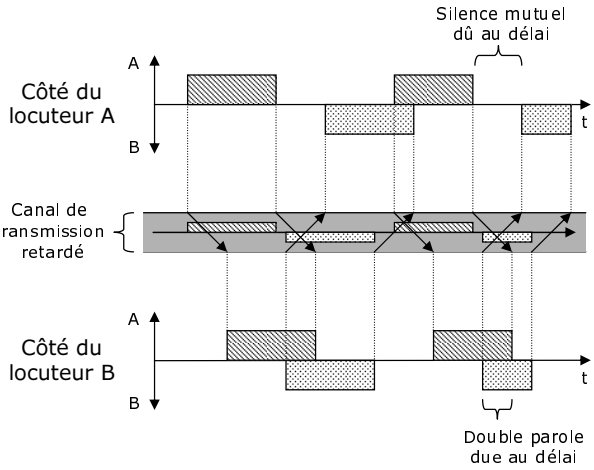


Figure 1.3 : *Impact du délai sur la conversation (d'après [Hammer* et al. *2005])*

1.1.1.4 Synthèse

Pour résumer, le jugement de la qualité vocale est influencé par de nombreux paramètres qui peuvent dépendre : (i) de facteurs individuels à la personne qui la juge (expérience passée, attentes et humeur de chacun), (ii) du contenu du signal de parole lui-même, (iii) des critères de qualité explorés, et (iv) de facteurs extérieurs à l'individu (contexte et environnement). Les différents paramètres influençant la qualité vocale sont résumés dans la figure 1.4.

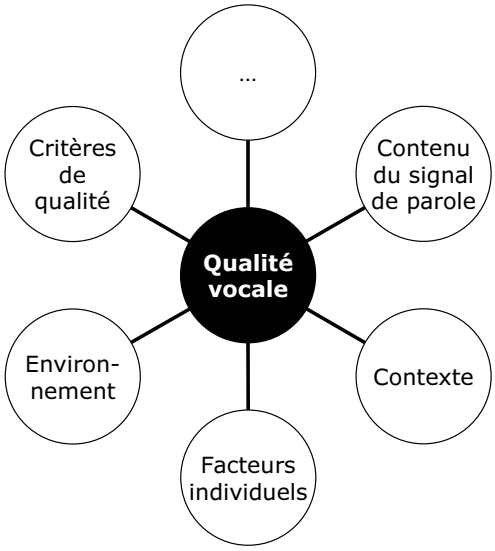


Figure 1.4 : *Différents paramètres influençant le jugement de la qualité vocale*

1.1.2 Évolution des systèmes de télécommunications

Dominé par le téléphone classique depuis son invention en 1876 par Graham Bell, le monde des télécommunications a considérablement évolué en quelques décennies. Cette révolution des systèmes de télécommunications peut être attribuée à plusieurs ruptures, telles que la numérisation des réseaux, la miniaturisation des circuits et l'émergence du haut débit, accélérée par la dérégulation du marché et l'ouverture à la concurrence entre opérateurs. Les différentes technologies répondent à une demande des clients et leur fournissent de nouveaux services. Cependant, elles introduisent par la même occasion des dégradations additionnelles à celles rencontrées avec le téléphone classique. Nous présenterons dans ce paragraphe les principaux

systèmes de télécommunications existants et les différentes dégradations que chacun d'eux engendre.

1.1.2.1 Téléphonie classique

Le téléphone classique utilise la technique de commutation de circuits à multiplexage temporel, mise en œuvre dans le réseau téléphonique commuté (RTC). Les commutateurs mettent en relation tous les abonnés en empruntant les lignes du réseau public. Au sein d'une même entreprise, les communications internes sont traitées par un autocommutateur privé (PABX, *Private Automatic Branch eXchange*). Aujourd'hui, la voix est transmise sur le RTC de manière analogique sur une paire de fils de cuivre entre le poste de l'usager et le plus proche central téléphonique (portion du réseau appelée « boucle locale » ou « réseau local »). La voix est ensuite transmise entre les deux centraux téléphoniques de manière numérique. La transformation du signal, analogique vers numérique et inversement, est assurée par des « codecs » (COdeur/DECodeur), qui fonctionnent suivant le schéma classique de l'échantillonnage et de la quantification. Le téléphone classique fonctionne en bande étroite (300 - 3400 Hz), alors que la bande passante de l'oreille est située entre 20 Hz et 20 kHz. Cette réduction de bande passante constitue une première dégradation de la communication par téléphone comparée à la communication face à face. Ensuite des dégradations telles que l'écho (acoustique et/ou électrique), la distorsion de l'effet local ou le bruit (ambiant et/ou de circuit) peuvent affecter la communication.

1.1.2.2 Systèmes numériques

La voix peut aussi être transmise en numérique de bout en bout de la communication. La transformation du signal, analogique vers numérique et inversement, est alors assurée par le codec situé dans le poste de l'usager. Cette technologie a été appelée RNIS (Réseau Numérique à Intégration de Services) par l'Union Internationale des Télécommunications (UIT). Le RNIS peut servir à transmettre tout type de donnée numérique, *i.e.* la voix et également des données informatiques à un débit de 64 kbits par seconde. Les dégradations additionnelles par rapport au RTC sont principalement celles induites par le codage à débit réduit de la parole, *i.e.* délais de traitement, distorsion du signal de parole.

1.1.2.3 Systèmes mobiles

Depuis le début des années 90, la téléphonie fixe est fortement concurrencée par la téléphonie mobile. Le mobile a connu une croissance fulgurante : en France, il y avait déjà en 1997 près de 6 millions d'abonnés et en 2003 plus de 41 millions, soit une augmentation de 583% sur 6 ans. Jusqu'à présent la téléphonie mobile fonctionne essentiellement avec la technologie GSM (*Global System for Mobile communications*) standardisée par l'ETSI (*European Telecommunications Standards Institute*) en 1990. Les principales dégradations identifiées pour le RTC et/ou le RNIS (écho, bruit, dégradation due au codage de la parole, bande étroite) sont également présentes avec la téléphonie mobile. Cependant, la mobilité de l'utilisateur ajoute une variabilité des dégradations dans le temps, tels que des bruits non stationnaires (*e.g.* voiture qui passe dans la rue), et un délai plus long dû aux traitements numériques. La téléphonie mobile est actuellement en pleine convergence avec Internet : c'est la troisième génération de téléphones mobiles (3G), notamment avec la technologie UMTS (*Universal Mobile Telecommunications System*). Elle offre une bande passante plus large et un débit plus élevé pour l'échange de données et pour de nouvelles applications telles que la visiophonie ou la télévision.

1.1.2.4 Réseaux en mode paquet

Depuis le milieu des années 90, le grand public a accès à Internet, qui est un ensemble de réseaux interconnectés, fonctionnant par commutation de paquets avec le protocole IP (*Internet Protocol*) et permettant à ses utilisateurs de s'échanger des informations (sons, images, vidéos et données). Ces informations ont d'abord été transmises à bas débit via des modems de faible puissance, ensuite remplacés par la technologie haut débit ou DSL (*Digital Subscriber Line*). L'un des intérêts d'Internet est qu'il permet de téléphoner via le réseau IP, grâce à la « VoIP » (*Voice over IP*). La voix est numérisée, comprimée et découpée sous forme de paquets IP par l'émetteur, le récepteur effectuant les opérations inverses pour reconstituer la parole. Une des particularités du réseau IP par rapport au RTC réside dans le mode de transport des données. En effet, les paquets IP sont acheminés dans le réseau indépendamment les uns des autres, ce qui ne garantit pas leur arrivée dans le bon ordre au niveau du récepteur. Ce sont les routeurs qui assurent l'acheminement de chaque paquet IP à travers le réseau en empruntant le chemin le plus court. Il arrive cependant que des paquets soient en retard ou perdus. Ceci se traduit au niveau du récepteur par un délai variable en fonction du retard de chaque paquet, appelé « gigue », ou par des paquets perdus, *i.e.* par des silences dans le signal de parole reçu, appelés « pertes de paquets ». La gigue, les pertes de paquets et le délai de traitement sont donc les trois dégradations majeures induites par la VoIP, en plus des dégradations classiques identifiées pour le RTC et/ou le RNIS (écho, bruit, dégradation due au codage de la parole).

1.1.2.5 Synthèse

Si la qualité vocale est assurée avec le téléphone classique, l'émergence et le succès commercial de nouvelles technologies telles que la téléphonie mobile ou la VoIP, qui introduisent de nouvelles dégradations, obligent les opérateurs de réseaux à évaluer de manière fiable la qualité vocale de leurs systèmes de télécommunications.

1.2 Évaluation subjective de la qualité vocale

Comme il a été vu dans la section 1.1.1, le jugement de la qualité vocale est avant tout subjectif. La meilleure façon d'évaluer la qualité vocale est donc de faire appel à des utilisateurs et de les interroger, sous forme de sondages ou de tests en laboratoire. Ce sont les méthodes d'évaluation subjective de la qualité.

1.2.1 Choix des sujets

D'après la Recommandation P.800 de l'UIT [UIT-T Rec. P.800 1996] qui présente les différentes méthodes d'évaluation subjective de la qualité vocale :

> « *Les sujets participant aux essais de conversation sont choisis au hasard parmi la population normale utilisant le téléphone, les conditions étant :*
> - *qu'ils n'aient pas participé directement à des travaux d'évaluation de la qualité des circuits téléphoniques ou à des travaux connexes comme le codage de la voix ;*
> - *qu'ils n'aient pas participé à des essais subjectifs au cours des six mois précédents au moins, ni à des essais de conversation depuis un an au moins.*
>
> *Si la population disponible est trop limitée on en tiendra compte pour tirer des conclusions des résultats. On n'équilibrera pas le nombre des sujets masculins et féminins, à moins que le montage de l'expérience ne l'exige.* »

Certains tests pourront exiger d'avoir recours à des sujets dits « experts » (*e.g.* spécialistes du traitement de la parole), mais la plupart des tests feront appel à des sujets dits « naïfs », c'est-à-dire recrutés dans la population.

Les notes des participants pour une condition de test donnée sont moyennées pour obtenir une note moyenne d'opinion dénommée « note MOS » (*Mean Opinion Score*). Le fait de moyenner les notes individuelles permet de diminuer l'effet subjectif sur l'évaluation de la qualité vocale mis en évidence dans le paragraphe 1.1.1.

1.2.2 Méthodes normalisées

L'évaluation subjective des systèmes de télécommunications consiste généralement en un test dans un laboratoire. L'UIT a normalisé plusieurs méthodes d'évaluation subjective de la qualité vocale et différents types d'essai subjectif, décrits dans la Recommandation P.800 [UIT-T Rec. P.800 1996]. Le choix parmi ces méthodes dépend de plusieurs facteurs, le principal étant le type de dégradation à tester. Comme il a été vu dans le paragraphe 1.1.1.3, les dégradations susceptibles d'affecter une communication téléphonique sont de trois types [Richards 1973] :

- type A : les dégradations entraînant une difficulté d'écoute quand la communication est unidirectionnelle et qu'aucune assistance n'est donnée par le locuteur, *i.e.* en contexte d'écoute.
- type B : les dégradations entraînant une difficulté pour parler, *i.e.* en contexte de locution.
- type C : les dégradations entraînant une difficulté pour converser (dégradations associées avec l'alternance des rôles de locuteur et d'auditeur de chacun des participants), *i.e.* en contexte de conversation.

À chacun de ces types de dégradations correspond un type d'essai subjectif normalisé, décrit dans les Recommandations P.800 [UIT-T Rec. P.800 1996] et P.831 [UIT-T Rec. P.831 1998] de l'UIT. Selon le type de dégradation ou de système à évaluer, l'un ou l'autre des tests suivants sera choisi.

1.2.2.1 Essais d'opinion d'écoute

Les tests d'écoute consistent à placer les participants en situation d'écoute et à leur diffuser des séquences audio correspondant à différentes conditions de dégradation. Les séquences audio ont été préalablement enregistrées par plusieurs locuteurs différents et tous les participants écoutent les mêmes enregistrements. Les conditions testées concernent les dégradations affectant la qualité d'écoute, comme la distorsion de la parole due au codage, le bruit pour l'auditeur et les pertes de paquets. La notation s'effectue selon l'une des méthodes définies par l'UIT-T dans la Recommandation P.800 [UIT-T Rec. P.800 1996]. La plus utilisée est la méthode d'évaluation par catégories absolues (*Absolute Category Rating*, ACR) avec les catégories : 5 = Excellente, 4 = Bonne, 3 = Passable, 2 = Médiocre, 1 = Mauvaise. Référence peut aussi être faite à la méthode d'évaluation par catégories de dégradation (*Degradation Category Rating*, DCR) avec les catégories : 5 = Dégradation inaudible, 4 = Dégradation audible mais pas gênante, 3 = Dégradation un peu gênante, 2 = Dégradation gênante, 1 = Dégradation très gênante, et à la méthode d'évaluation par catégories de comparaison (*Comparison Category Rating*, CCR) avec les catégories : 3 = Bien meilleure, 2 = Meilleure, 1 = Légèrement meilleure, 0 = À peu près équivalente, -1 = Un peu moins bonne, -2 = Moins bonne, -3 = Beaucoup moins bonne. Il est recommandé par l'UIT-T que chaque test comprenne des conditions de référence (*i.e.* sans aucune dégradation) afin que les sujets aient une qualité d'écoute de référence. Plusieurs questions peuvent être posées aux participants permettant ainsi d'évaluer différents critères de la qualité vocale. Les échelles d'appréciation les plus fréquemment employées dans un tel contexte sont la qualité d'écoute, l'effort d'écoute et le niveau sonore préféré, données dans les tableaux 1.1, 1.2 et 1.3.

Qualité de la parole	Note
Excellente	5
Bonne	4
Passable	3
Médiocre	2
Mauvaise	1

Tableau 1.1 : *Échelle de qualité d'écoute (méthode ACR)*

Effort nécessaire pour comprendre le sens des phrases	Note
Détente absolue, aucun effort	5
Attention nécessaire, pas d'effort appréciable	4
Effort modéré	3
Effort considérable	2
Incompréhensible en dépit de tous les efforts possibles	1

Tableau 1.2 : *Échelle des efforts d'écoute (méthode ACR)*

Niveau sonore préféré	Note
Beaucoup plus fort que préféré	5
Plus fort que préféré	4
Selon préférence	3
Plus faible que préféré	2
Bien plus faible que préféré	1

Tableau 1.3 : *Échelle de niveau sonore préféré (méthode ACR)*

1.2.2.2 Tests de parole et d'écoute

Dans un test de parole et d'écoute, les participants sont placés dans le contexte de locution. Ils doivent donc parler dans le microphone du système à tester et écouter simultanément ce qui arrive du haut-parleur. Des dégradations affectant la qualité de locution, telles que l'écho, la distorsion de l'effet local et le bruit pour le locuteur, peuvent ainsi être testées. De même que pour les tests d'écoute, les participants notent les conditions testées selon l'une des méthodes définies par les Recommandations P.800 [UIT-T Rec. P.800 1996] et P.831 [UIT-T Rec. P.831 1998]. Chaque test doit également comprendre des conditions de référence (*i.e.* sans aucune dégradation) afin que les sujets aient une qualité de locution de référence. Les questions posées dans un tel contexte concernent en général la qualité globale, la dégradation due à l'écho et la dégradation due au bruit. La Recommandation P.831 [UIT-T Rec. P.831 1998] introduit une échelle de notation spécifique à l'évaluation de la dégradation due à l'écho, basée sur la méthode DCR et donnée dans le tableau 1.4. La dégradation due au bruit peut être testée selon la même échelle que la dégradation due à l'écho, en modifiant l'intitulé de la question.

Dégradation due à l'écho	Note
Imperceptible	5
Perceptible, mais non gênante	4
Légèrement gênante	3
Gênante	2
Très gênante	1

Tableau 1.4 : *Échelle de dégradation due à l'écho (méthode DCR)*

1.2.2.3 Essais d'opinion de conversation

Les tests de conversation sont conçus pour évaluer la qualité dans la situation la plus réaliste. Deux participants sont installés chacun dans une salle et dialoguent via le système de télécommunications étudié. Les conditions testées dans ces tests concernent les dégradations des deux contextes précédents (écoute et locution), ainsi que les dégradations affectant spécifiquement l'interaction de la conversation, comme le délai et la double parole. De même que pour les tests d'écoute, il est recommandé par l'UIT-T que chaque test de conversation

comprenne des conditions de référence (*i.e.* sans aucune dégradation) afin que les sujets aient une qualité de conversation de référence. Les conditions testées peuvent être les mêmes pour les deux participants (test symétrique) ou différentes (test asymétrique). Le but étant de reproduire une communication téléphonique réaliste, des prétextes de conversation (sous la forme de dessin à décrire ou de jeu de rôle) sont généralement fournis aux participants. Ainsi, des scénarios de conversation (*Short Conversation Test*, SCT) ont été créés [Möller 1997a]. Leurs thèmes sont par exemple une commande de pizza ou un achat de billet d'avion. Chacun des participants note ensuite la qualité de la conversation qu'il vient d'expérimenter selon l'une des méthodes définies par les Recommandations P.800 [UIT-T Rec. P.800 1996] et P.831 [UIT-T Rec. P.831 1998]. Généralement lors de tels tests, il est demandé aux participants d'évaluer la qualité globale, la dégradation due à l'écho, la dégradation due au bruit et l'effort d'interruption.

Contrairement aux tests d'écoute qui nécessitent simplement l'enregistrement de séquences audio dans différentes conditions de dégradation et la diffusion de ces enregistrements aux participants du test, les tests de conversation imposent de concevoir un montage qui fonctionne en duplex intégral et qui dégrade la qualité de parole en temps réel. De plus, pour simuler fidèlement une conversation téléphonique réelle, il est important que chacune des conversations du test ait une durée réaliste, c'est-à-dire de quelques minutes. S'il peut être intéressant de mettre en relation les notes de qualité attribuées par les participants avec les signaux de parole correspondants, l'enregistrement de ces signaux devra également être effectué en temps réel et intégré au montage de test. De fait, les tests de conversation sont coûteux et ne permettent pas, en un temps donné, d'étudier autant de conditions que les autres tests, ce qui les rend plus rares dans la littérature.

1.2.2.4 Synthèse

Au cours de ces différents tests, chaque sujet donne une note pour chaque condition testée et pour chaque question posée. Pour chaque condition et chaque question, une moyenne des notes de tous les sujets est effectuée. La note ainsi obtenue est appelée note MOS et représente le jugement moyen des sujets. Cette moyenne permet à l'expérimentateur de contrôler la principale source de variabilité d'un test subjectif : le sujet lui-même. Les autres sources de variabilité peuvent être contrôlées en prenant de nombreuses précautions, afin d'obtenir des résultats de test fiables et exploitables. Les sources de variabilité d'un test subjectif, en général, sont les suivantes :

A contenu de la séquence audio ou de la conversation (variation du contenu phonétique, des temps de parole, etc.),

B variations entre locuteurs (niveau de parole),

C variations entre auditeurs (niveau d'écoute),

D genre des sujets (*i.e.* hauteur de la voix),

E type de sujet (expert ou naïf),

F degré de connaissance de la voix de l'interlocuteur,

G langue,

H effet d'ordre (ordre de présentation des conditions, des prétextes, etc.).

Certaines sources de variabilité (A, C, D, F et H) peuvent être contrôlées à condition de prendre les précautions suivantes :

I dans les tests d'écoute et de locution, le contenu de la séquence audio écoutée par l'auditeur et de la phrase prononcée par le locuteur est contrôlé par l'expérimentateur,

II le niveau d'écoute peut être imposé,

III si besoin, les sujets peuvent être choisis de telle sorte qu'il y ait autant d'hommes que de femmes,

IV si besoin, les sujets peuvent être choisis de telle sorte qu'ils se connaissent et ainsi connaissent la voix de leur interlocuteur,

V l'ordre de présentation des conditions et des séquences audio / phrases à prononcer / prétextes de conversation est différent pour tous les sujets, ce qui permet d'éviter, en moyennant, l'effet d'ordre. L'ordre de présentation peut être déterminé par exemple avec un carré gréco-latin. Un carré gréco-latin d'ordre 8 est donné dans le tableau 1.5. Les lettres indiquent l'ordre des conditions et les chiffres l'ordre des séquences audio / phrases à prononcer / prétextes de conversation.

A1	B2	C3	D4	E5	F6	G7	H8
G7	F6	E5	H8	B2	D4	C3	A1
C3	H8	B2	G7	D4	A1	F6	E5
E5	D4	H8	F6	G7	C3	A1	B2
H8	C3	A1	B2	F6	G7	E5	D4
B2	G7	D4	C3	A1	E5	H8	F6
D4	E5	F6	A1	C3	H8	B2	G7
F6	A1	G7	E5	H8	B2	D4	C3

Tableau 1.5 : *Exemple de carré gréco-latin d'ordre 8. Les lettres indiquent l'ordre des conditions et les chiffres l'ordre des séquences audio / phrases à prononcer / prétextes de conversation*

1.2.3 Effets subjectifs des différentes dégradations sur la qualité vocale

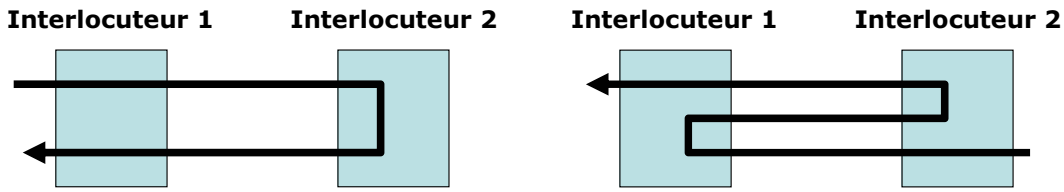
Les dégradations et les phénomènes les plus importants à prendre en compte sont les suivants :

- écho,
- délai,
- distorsion de la parole due au codage,
- bruit,
- perte de paquets,
- distorsion de l'effet local,
- variation dans le temps des dégradations,
- effet des périodes de double parole,
- effet des dispositifs de traitement du signal (annuleurs d'écho, détecteurs d'activité vocale, générateurs de bruit de confort, réducteurs de bruit, etc.).

1.2.3.1 Échos

Quand des réflexions de signal se produisent en association avec des délais notables, le locuteur et l'auditeur peuvent percevoir de l'écho.

L'écho pour le locuteur (*cf.* figure 1.5(a)) se produit lorsqu'une certaine partie de son signal vocal est renvoyée en retour avec un retard suffisant (en général supérieur à 25 ms) pour que le signal puisse être distingué d'un effet local normal. Il dépend de deux paramètres : le temps de propagation, caractérisé par le temps de propagation moyen dans un sens du trajet d'écho, noté T, et le niveau du signal vocal réfléchi vers le locuteur, caractérisé par l'équivalent en sonie du trajet d'écho pour le locuteur, noté TELR (*Talker Echo Loudness Rating*). L'écho pour l'auditeur survient quand un signal doublement réfléchi arrive du côté de l'auditeur, avec un retard par rapport au signal original (*cf.* figure 1.5(b)). Il peut être caractérisé par l'affaiblissement pondéré du trajet des courants d'écho, noté WEPL (*Weighted Echo Path Loss*) et le temps de propagation aller-retour, noté Tr, d'un circuit à 4 fils faisant partie de la chaîne de connexion.



(a) Écho pour le locuteur (b) Écho pour l'auditeur

Figure 1.5 : *Échos pour le locuteur et pour l'auditeur, perçus par l'interlocuteur 1*

Basiquement, l'impact sur la qualité de parole de l'écho, en particulier de l'écho pour le locuteur, augmente quand le volume du signal d'écho augmente (*i.e.* quand TELR diminue) et quand le délai du trajet d'écho augmente (*i.e.* quand T augmente) [ETSI ETR 250 1996]. Ceci est confirmé par les différents tests subjectifs qui ont été entrepris sur l'écho pour le locuteur. Dans [Möller 1997b], un test de conversation est présenté pour de larges gammes de valeurs de T et de TELR. Les résultats sont reproduits dans la figure 1.6. Un autre test de conversation est présenté dans [Osaka *et al.* 1992] sur l'écho pour le locuteur avec également de larges gammes de valeurs pour Tr et TELR. Les résultats sont donnés dans la figure 1.7, avec une échelle MOS comprise entre 0 et 3.

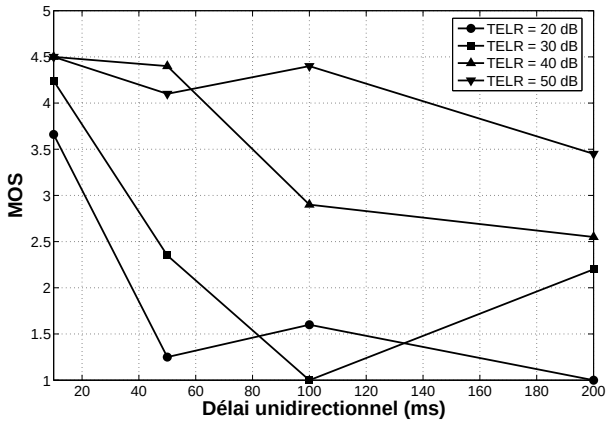


Figure 1.6 : *Résultats d'un test de conversation présenté dans [Möller 1997b] sur l'écho pour le locuteur*

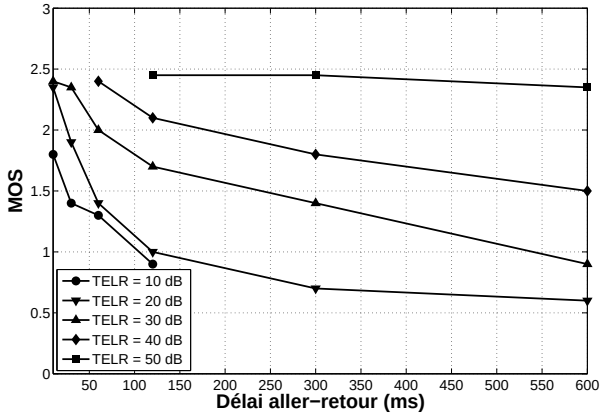


Figure 1.7 : *Résultats d'un test de conversation présenté dans [Osaka* et al. *1992] sur l'écho pour le locuteur*

Tous ces tests amènent à la même conclusion, à savoir que la qualité diminue quand le délai du trajet d'écho T augmente et quand l'affaiblissement de l'écho TELR diminue.

En se basant sur cette constatation et ce type de test subjectif, les limites de l'écho pour le locuteur ont été définies et sont représentées par une courbe limite de TELR en fonction de T. Cette courbe est fournie dans la Recommandation G.131 [UIT-T Rec. G.131 1996] et reproduite dans la figure 1.8.

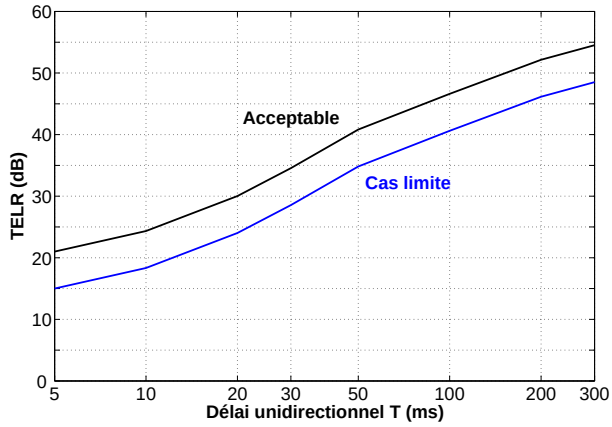


Figure 1.8 : *Courbes de tolérance à l'écho pour le locuteur [UIT-T Rec. G.131 1996]*

De même que l'écho pour le locuteur, la gêne due à l'écho pour l'auditeur augmente quand le délai Tr augmente et quand l'affaiblissement WEPL diminue. Cependant, si l'écho pour le locuteur est contrôlé, l'écho pour l'auditeur ne posera pas problème.

1.2.3.2 Délai

Le délai bout en bout d'une communication téléphonique est la somme du temps nécessaire pour le traitement de la parole à l'émission, pour la transmission des échantillons de parole (respectivement, des paquets de parole) sur le réseau téléphonique classique (resp., sur le réseau IP) et pour le traitement de la parole à la réception. Le temps de transmission, dans le cas de communication sur IP, est lui-même divisé en deux parties : l'une constante (délai constant dû au réseau) et l'autre variable dans le temps (gigue).

Le délai a comme principal effet de réduire l'interactivité entre le locuteur et l'auditeur (*cf.* paragraphe 1.1.1.3), et joue aussi un rôle sur l'écho. Bien que l'impact du délai soit connu, le seuil au-delà duquel celui-ci devient gênant pour la conversation n'est pas clairement identifié. Dans [Möller 2000b], l'auteur souligne en particulier la subjectivité de la perception du délai : lors d'une conversation, chacun attend une réponse de son interlocuteur dans une certaine fenêtre temporelle, dont la durée varie en fonction des individus et du type de conversation. De plus, quand le temps de réponse ne correspond pas à cette fenêtre temporelle, le tort n'est pas attribué au système mais plutôt à l'interlocuteur. Dans [Cermak 2002], une discussion est faite sur l'effet des dégradations introduites par la VoIP sur la qualité de la parole et en particulier sur le délai. Il y est constaté que la littérature sur l'effet du délai est plus nombreuse mais aussi moins consistante que celle sur les pertes de paquets et la gigue. Les résultats des études sur les effets du délai menées depuis une trentaine d'années se sont accumulés mais n'ont pas convergé vers une asymptote claire. En effet, contrairement à la limite de 400 ms donnée dans [UIT-T Rec. G.114 2003] et obtenue grâce à une notation MOS de la qualité globale, il semble que d'autres mesures telles que le seuil de détection du délai ou l'efficacité de la conversation donnent d'autres limites du délai supérieures à 400 ms. En particulier, plusieurs études récentes [Dvorak et James 2004, ETSI 3GPP TR 26.935 2004, Raake 2004b] ont conclu que la perception du délai chez les utilisateurs pourrait avoir évolué au fil des ans, notamment du fait des nouvelles technologies (mobile, IP) qui ont habitué les utilisateurs à des délais plus longs qu'en téléphonie classique. Cette évolution peut être observée grâce à deux tests de conversation effectués à 10 années d'intervalle par les laboratoires Bellcore en

1992 [Möller 1997b] et IKA en 2002 [Möller et Raake 2002]. Les résultats de ces deux tests sont reproduits dans la figure 1.9.

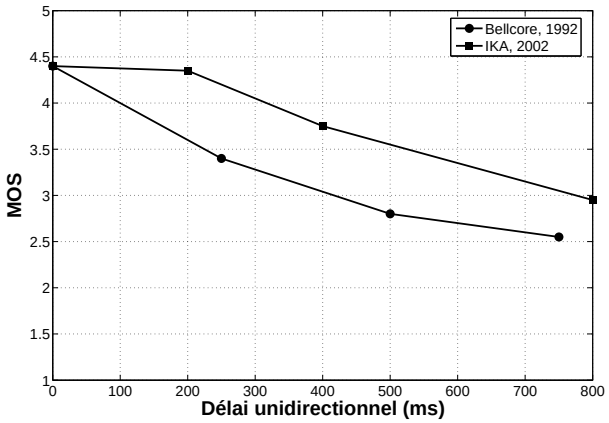


Figure 1.9 : *Résultats des deux tests de conversation présentés dans [Möller 1997b, Möller et Raake 2002] sur le délai absolu*

Le test de conversation présenté dans [Kitawaki et Itoh 1991] montre que selon le type de tâche de conversation (*i.e.* en fonction du niveau d'interactivité de la conversation) l'effet du délai n'est pas le même sur la qualité. Quatre des six différentes tâches étudiées sont les suivantes :

- tâche 1 : lire à tour de rôle des nombres aléatoires aussi vite que possible,
- tâche 2 : vérifier à tour de rôle des nombres aléatoires aussi vite que possible,
- tâche 4 : vérifier à tour de rôle des noms de villes aussi vite que possible,
- tâche 6 : conversation libre.

La vitesse de commutation de la conversation diminue de la tâche 1 à la tâche 6. Ceci est confirmé par les résultats du test, présentés dans la figure 1.10 (avec une échelle MOS comprise entre 0 et 4), la note MOS diminue quand le délai aller-retour augmente et quand la vitesse de commutation de la conversation augmente (*i.e.* de la tâche 6 vers la tâche 1). Ainsi, l'effet du délai sur le jugement des utilisateurs est dépendant du type de communication qu'ils effectuent.

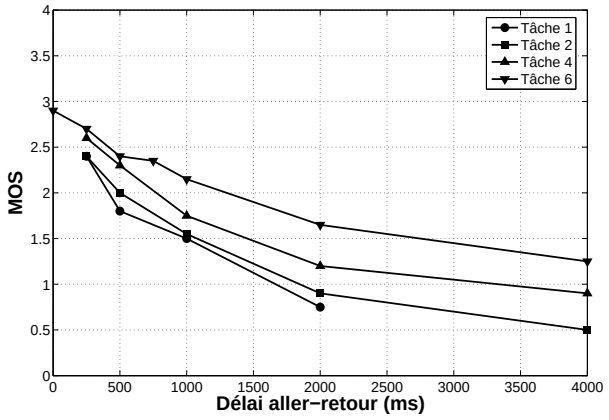


Figure 1.10 : *Résultats du test de conversation présenté dans [Kitawaki et Itoh 1991] sur l'effet du délai absolu en fonction de la tâche de conversation*

1.2.3.3 Distorsion de la parole due au codage

Les algorithmes de codage de la parole peuvent être classés en 3 catégories [Mohamed 2003].

Codage de formes d'onde (*waveform coding*) : son but est de reproduire la forme d'onde originale aussi fidèlement que possible et il est utilisé à des débits binaires plutôt élevés. Le signal encodé est alors de très bonne qualité. Les codecs de type PCM (*Pulse Code Modulation*), DPCM (*Differential Pulse Code Modulation*) ou ADPCM (*Adaptive Differential Pulse Code Modulation*) appartiennent à cette catégorie.

Codage paramétrique (*parametric coding*) : il opère à des débits binaires très faibles et utilise un modèle de production de la parole pour transmettre seulement les paramètres importants d'un point de vue perceptuel. Ces codeurs sont aussi appelés vocoders (voice coders). La qualité du signal résultant est dégradée mais le signal reste intelligible.

Codage hybride (*hybrid coding*) : il combine les caractéristiques des codeurs par forme d'onde et des vocoders, permettant d'obtenir une bonne qualité à des débits binaires modérés. Les codecs de type MPE (*Multi-Pulse Excitation*), RPE (*Regular Pulse Excitation*) ou CELP (*Codebook Excited Linear Prediction*) appartiennent à cette catégorie.

Les caractéristiques des principaux codecs vocaux bande étroite normalisés sont résumées dans le tableau 1.6.

Standard	Type de codage	Débit binaire (kb/s)
UIT-T G.711	Forme d'onde (PCM)	64
UIT-T G.723.1	Hybride (ACELP / MP-MLQ)	5.3 / 6.3
UIT-T G.726	Forme d'onde (ADPCM)	16, 24, 32 et 40
UIT-T G.728	Hybride (LD-CELP)	16
UIT-T G.729	Hybride (CSA-CELP)	8
IS-54	Hybride (VSELP)	7.95
GSM-FR	Hybride (LTP-RPE)	13
GSM-HR	Hybride (VSELP)	5.6
GSM-EFR	Hybride (ACELP)	12.2

Tableau 1.6 : *Caractéristiques des principaux codeurs vocaux en bande étroite*

Il existe de nombreux tests subjectifs sur le codage [Coverdale et Cheung 1997, Barriac 1997, Cheung 1998, Möller 1998, Möller 2000a, Möller 2001], que ce soit avec un seul codec ou des cascades de plusieurs codecs. La contribution [Möller 1998] à elle seule fait le bilan de nombreux tests sur les différents codecs. Une moyenne des notes MOS de chaque codec seul ou en cascade avec lui-même est donnée dans la figure 1.11.

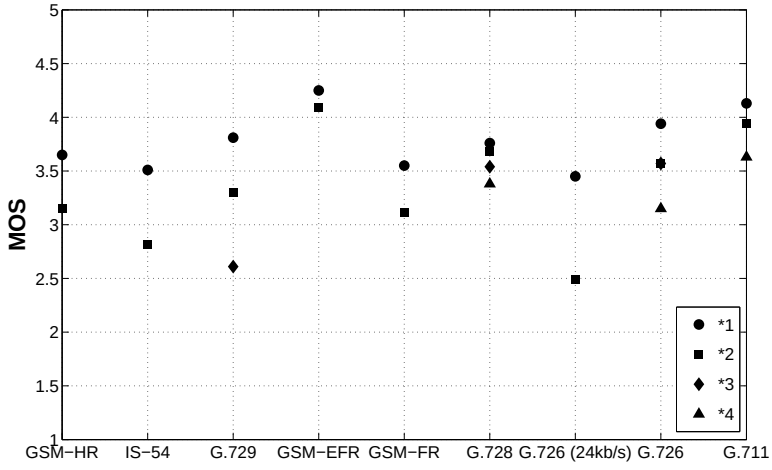


Figure 1.11 : *Moyennes des notes MOS de plusieurs tests subjectifs recueillis dans [Möller 1998] pour les principaux codecs vocaux bande étroite en cascade avec eux-mêmes*

La figure 1.11 confirme tout d'abord le fait que la note MOS associée à n codecs identiques mis en cascade diminue quand n augmente, quel que soit le codec considéré. Ensuite, la qualité de la parole dépend non seulement du débit binaire du codec considéré mais aussi du type

d'algorithme de codage utilisé dans ce codec. Ainsi, de manière générale, le codec G.711 utilisant un algorithme de codage par formes d'onde est mieux jugé que les autres codecs sauf le codec GSM-EFR. Proportionnellement, pour un débit beaucoup moins élevé que celui de G.711, les codecs utilisant un algorithme de codage hybride (*e.g.* GSM-EFR, GSM-HR, G.729) ne sont pas beaucoup moins bien notés.

1.2.3.4 Bruits

Bien que le bruit soit moins un problème pour les systèmes numériques en comparaison aux systèmes analogiques, il influence encore la perception de la qualité des systèmes par les usagers [ETSI ETR 250 1996]. Il existe deux sources de bruit principales dans une connexion téléphonique : le bruit de circuit et le bruit de salle (à l'émission et à la réception).

Bruit de circuit Le bruit de circuit peut comprendre un bruit blanc de circuit et un bruit d'intermodulation introduit par les systèmes de transmission ainsi que d'autres types de perturbations tels que le bruit impulsif. La satisfaction de l'usager dépend de la puissance, de la distribution fréquentielle et de la distribution d'amplitude du bruit. De manière générale, la satisfaction diminue de façon monotone lorsque la puissance du bruit augmente [UIT-T Rec. P.11 1993].

Bruit de salle (à l'émission et à la réception) Le bruit de salle correspond au bruit de fond dans lequel fonctionne l'appareil téléphonique. Il peut affecter la communication de plusieurs manières [Möller 2000b] :

- du côté émission[1] : le bruit de salle atteint l'oreille engagée dans la communication à travers l'espace existant entre le combiné et l'oreille, et aussi via le trajet électrique de l'effet local. L'autre oreille (oreille « libre ») est beaucoup plus affectée par le bruit, mais il est considéré, bien que cela ne soit pas toujours vérifié pour les niveaux de bruit de salle élevés ou pour des bruits de salle ayant une certaine signification pour l'auditeur, que les mécanismes auditifs cérébraux « déconnectent » les sons atteignant l'oreille libre. Des études sur différents types de bruits de salle sont présentées dans la suite du paragraphe.
- du côté réception[2] : le bruit de salle a un impact sur le signal de parole transmis, auquel il se superpose. Une partie de ce bruit est compensée par l'effet Lombard, présenté dans le paragraphe 1.1.1.3. Il a été montré que le locuteur compense de cette façon la moitié de la réduction du rapport signal-à-bruit introduite, mesurée en décibels.

L'effet perturbant du bruit de salle est caractérisé par sa valeur en dB(A), qui est une moyenne de puissance pondérée en fréquence. Les nombreux tests subjectifs relatifs au bruit de salle s'intéressent principalement à son niveau, à son type, et à la relation entre bruit de salle à l'émission et bruit de salle à la réception. Dans [Möller 1997b], un test est présenté sur l'influence du niveau du bruit de salle à l'émission sur la qualité. Les résultats correspondants sont donnés dans la figure 1.12. Comme espéré, plus le niveau du bruit de salle à l'émission augmente, plus la qualité diminue.

Des tests subjectifs d'écoute sont rapportés dans [Jekosch et Klaus 1997] sur l'influence des bruits de salle à l'émission et à la réception, tant du point de vue du niveau que du type des bruits de salle. Les conditions des quatre tests sont données dans le tableau 1.7. Les niveaux des bruits de salle se situent dans l'intervalle 40-80 dB(A).

Les différents résultats présentés dans [Jekosch et Klaus 1997] montrent que quels que soient les types de bruits de salle à l'émission et à la réception :

[1] Nous appellerons « côté émission » le côté du système où le bruit de salle est diffusé.

[2] Nous appellerons « côté réception » le côté du système où le bruit de salle est reçu après transmission par le système.

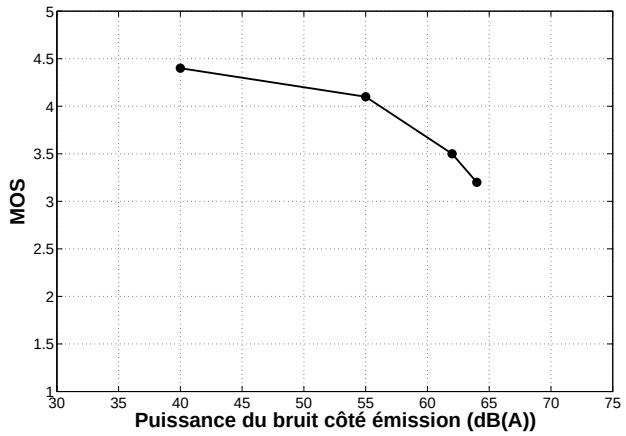


Figure 1.12 : *Résultats d'un test présenté dans [Möller 1997b] sur l'influence du bruit de salle à l'émission sur la qualité*

Test		Bruit de salle à l'émission	Bruit de salle à la réception
A	a	bruit blanc	bruit blanc
	b	bruit de gare	bruit de gare
B	c	bruit blanc	bruit de gare
	d	bruit de gare	bruit blanc
C	e	bruit de télévision	musique de fond
	f	pleurs de bébé	marteau-piqueur
D	g	sonnerie d'une cloche	bruit de gare
	h	marteau-piqueur	musique de fond

Tableau 1.7 : *Conditions des tests présentés dans [Jekosch et Klaus 1997] sur l'influence des bruits de salle sur la qualité*

- pour des valeurs faibles du niveau du bruit à la réception (*i.e.* entre 40 et 60 dB(A)), la qualité diminue quand le niveau du bruit à l'émission augmente,
- pour des valeurs élevées du niveau bruit à la réception (*i.e.* entre 60 et 80 dB(A)), la qualité est quasiment stable autour d'une valeur,
- d'une manière générale, pour une valeur du niveau du bruit à l'émission donnée, la qualité diminue quand le niveau du bruit à la réception augmente.

Les résultats montrent aussi qu'il n'y a pas de différence significative dans les jugements selon les types de bruit de salle, sauf dans les cas où le niveau du bruit à l'émission est supérieur à 70 dB(A), alors il existe des différences qui peuvent être d'un point sur l'échelle MOS. Les résultats sont les meilleurs quand le bruit de salle du côté réception est du type musique, alors qu'un bruit du type bruit de gare donne les pires résultats (pires que le jugement d'un bruit de type marteau-piqueur).

1.2.3.5 Pertes de paquets

Les pertes de paquets surviennent dans le cas d'applications du type voix sur IP, transmettant la parole sous forme de paquets. Ceux-ci sont acheminés indépendamment les uns des autres dans le réseau Internet jusqu'au destinataire, par des chemins qui peuvent être différents. Des paquets peuvent être perdus au cours du trajet soit parce qu'ils ont emprunté une route sans issue, soit parce qu'un routeur ne les a volontairement pas transmis afin de décongestionner le réseau. De plus, les paquets de parole qui arrivent avec un délai trop long sont considérés comme perdus afin de ne pas trop retarder la communication. L'impact des pertes de paquets sur la qualité de la parole se manifeste par des coupures et/ou des craquements dans le signal reçu, pouvant rendre, dans les cas extrêmes, la parole inintelligible pour l'auditeur. Les pertes de paquets sont caractérisées *a priori* par :

- le taux de pertes de paquets (exprimé en %),

- le type de pertes de paquets (aléatoires ou en rafale),
- la taille des paquets perdus,
- la localisation des paquets perdus dans la communication,
- la présence ou non d'un algorithme de compensation des pertes de paquets (PLC, *Packet Loss Concealment*)[3] à la réception.

Leur effet sur la qualité dépend aussi fortement du type de codec utilisé, puisque chaque codec a une robustesse différente face aux pertes de paquets.

De nombreux tests subjectifs sont disponibles concernant l'effet des pertes de paquets sur la qualité, en fonction des différents paramètres énumérés précédemment. La plupart d'entre eux porte sur des taux de perte de paquets compris entre 0 et 20%, qui semble être la limite supérieure acceptable. En particulier, des tests subjectifs ont été effectués pour estimer l'effet des pertes de paquets (aléatoires ou en rafale) sur la qualité en fonction du codec utilisé et de la longueur des paquets [Karlsson 2001, BharrathSingh et Britt 2001, Raake 2001].

Taille des paquets Les résultats des tests présentés dans [Karlsson 2001] sont donnés dans les figures 1.13(a) à 1.13(c), pour les codecs G.711 (sans PLC) (longueur des paquets = 30 ms et 60 ms), G.723.1 (avec PLC) (longueur des paquets = 30 ms et 60 ms), G.729 (avec PLC) (longueur des paquets = 10 ms, 20 ms, 30 ms et 60 ms), respectivement, dans le cas de pertes de paquets aléatoires.

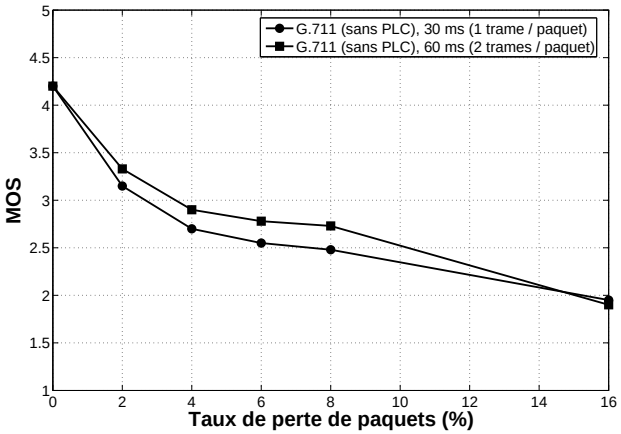
Ces résultats montrent d'une part que le codec G.711 étudié dans la figure 1.13(a) n'incluant pas d'algorithme de compensation des pertes de paquets (PLC) est moins bien évalué que les deux autres codecs en présence de pertes de paquets aléatoires, pour une longueur de paquets donnée. En ce qui concerne l'influence de la taille des paquets sur le jugement, elle semble dans l'ensemble très faible, puisque pour un codec et un taux de pertes aléatoires donnés, le jugement est quasiment identique quelle que soit la longueur des paquets.

Type de pertes de paquets Quelques tests subjectifs ont également été effectués avec des pertes de paquets en rafale [Karlsson 2001, Raake 2001]. Dans [Karlsson 2001], il a été démontré que :

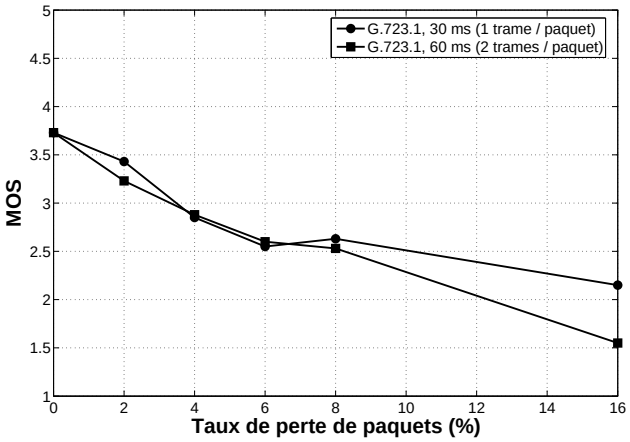
- la qualité de la parole pour des pertes de paquets en rafale dépend principalement du taux de pertes global pour tout le stimulus mais aussi de la densité des pertes de paquets dans la rafale elle-même,
- la qualité ne semble pas dépendre du nombre de trames de parole par paquet (*i.e.* de la longueur des paquets) pour un taux de pertes global donné, comme pour les pertes de paquets aléatoires,
- l'impact perceptuel des pertes de paquets en rafale dépend aussi du codec utilisé.

L'influence des deux types de pertes de paquets sur le jugement peut ainsi être comparée pour un codec donné. Dans la figure 1.14(a), les résultats d'un test présenté dans [Raake 2001] sur l'effet des pertes de paquets en rafale sur la qualité pour le codec G.711 (avec PLC) sont comparés à ceux sur l'effet des pertes de paquets aléatoires sur la qualité pour le codec G.711 (avec PLC). Dans la figure 1.14(b), les résultats d'un test présenté dans [Raake 2001] sur l'effet des pertes de paquets en rafale sur la qualité pour le codec G.729 sont comparés à ceux sur l'effet des pertes de paquets aléatoires sur la qualité pour le codec G.729. Les résultats des deux tests subjectifs présentés dans les figures 1.14(a) et 1.14(b) montrent que la différence d'influence sur la qualité entre les pertes de paquets aléatoires et en rafale dépend du codec considéré. Ainsi, pour le codec G.711 les pertes de paquets en rafale sont moins bien jugées que les pertes de paquets aléatoires alors que pour le codec G.729 les deux types de pertes de paquets ont quasiment le même impact sur la qualité. Cependant, le peu de tests subjectifs disponibles à ce jour sur l'impact des pertes de paquets en rafale semble indiquer que, à un

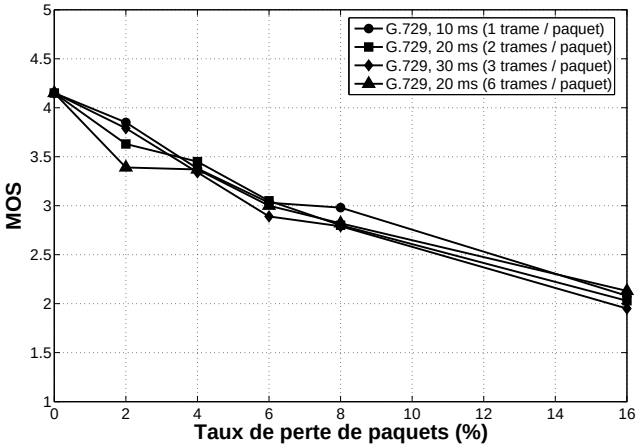
[3] Tous les codecs à bas débit contiennent un algorithme de PLC. Le codec G.711 a un algorithme de PLC optionnel, mais généralement activé.



(a) Codec G.711 (sans PLC). Longueur d'un paquet = 30 ms (1 trame / paquet) et 60 ms (2 trames / paquet)



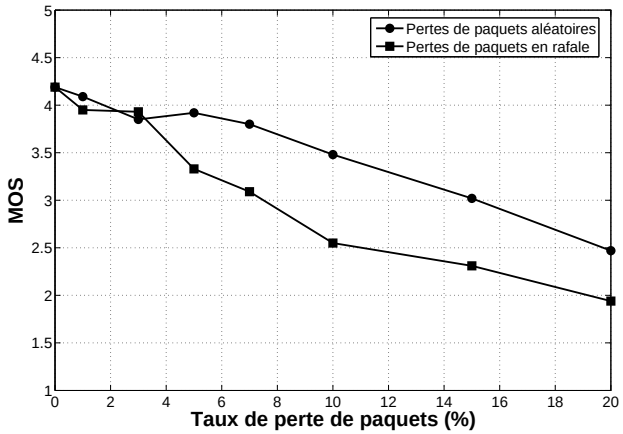
(b) Codec G.723.1 (avec PLC). Longueur d'un paquet = 30 ms (1 trame / paquet) et 60 ms (2 trames / paquet)



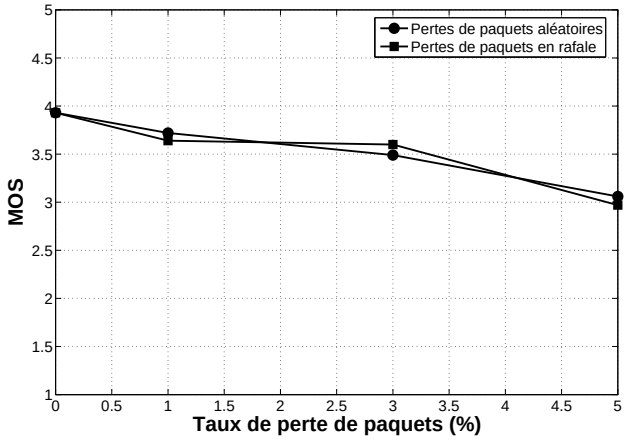
(c) Codec G.729 (avec PLC). Longueur d'un paquet = 10 ms (1 trame / paquet), 20 ms (2 trames / paquet), 30 ms (3 trames / paquet) et 60 ms (6 trames / paquet)

Figure 1.13 : *Résultats d'un test présenté dans [Karlsson 2001] sur l'effet des pertes de paquets aléatoires sur la qualité pour les codecs G.711 (sans PLC), G.723.1 (avec PLC) et G.729 (avec PLC)*

taux de pertes de paquets global donné, les pertes de paquets en rafale sont généralement moins bien notées que les pertes de paquets aléatoires [Karlsson 2001].



(a) Codec G.711 (avec PLC). Longueur d'un paquet = 10 ms



(b) Codec G.729. Longueur d'un paquet = 10 ms

Figure 1.14 : *Résultats d'un test présenté dans [Raake 2001] sur l'effet des pertes de paquets aléatoires et en rafale sur la qualité pour les codecs G.711 (avec PLC) et G.729*

1.2.3.6 Distorsion de l'effet local

L'effet local est défini comme étant la reproduction, par l'écouteur d'un appareil téléphonique, des sons captés par le microphone du même appareil et transmis de l'un à l'autre via plusieurs trajets mécaniques, acoustiques et électriques [ETSI ETR 250 1996].

Le trajet de l'effet local pour le locuteur doit idéalement être identique au trajet de transmission qui existe de la bouche à la propre oreille du locuteur dans une conversation face à face. Ainsi, l'affaiblissement du trajet d'effet local doit se situer dans un certain intervalle afin d'être dans une situation de locution confortable. Un effet local pour le locuteur trop important amène le locuteur à baisser le volume de sa voix, de telle sorte que l'auditeur à l'autre bout peut l'interpréter comme un affaiblissement trop important dans la connexion.

D'après des tests subjectifs de locution rapportés dans le supplément 11 aux Recommandations de la série P de l'UIT [UIT-T Supp. 11 (Séries P) 1995], l'intervalle préféré pour l'affaiblissement d'effet local est de 7 à 12 dB.

1.2.3.7 Variations dans le temps des dégradations

Les nouveaux systèmes de télécommunications ont non seulement introduit de nouvelles dégradations qui n'étaient pas rencontrées avec la téléphonie classique, mais aussi des dégradations variables dans le temps. Ainsi, le contexte de la voix sur IP, du fait qu'il repose sur une transmission de type paquet, introduit de nouvelles dégradations, qui sont par nature variables dans le temps : les pertes de paquets et la gigue. Le contexte des communications

mobiles a quant à lui introduit des bruits non stationnaires (*e.g.* bruits de rue) et des baisses de qualité en fonction de l'éloignement des antennes du réseau.

La qualité perçue varie alors avec la localisation de la dégradation dans le temps. Ce phénomène est appelé « effet de récence » [Gros 2001]. En effet, il a été montré que les dégradations se produisant à la fin d'une communication ont plus de poids sur le jugement global que les dégradations se produisant au début [Clark 2001].

Afin de rendre la gigue la plus petite possible et d'envoyer au décodeur un flot de paquets le plus régulier possible, un buffer de gigue est souvent implémenté au niveau de la réception. Il a pour rôle de recueillir les paquets arrivants et d'ajouter autant de délai de buffer possible pour compenser la gigue. Il envoie ensuite les paquets dans le bon ordre. En gardant le paquet reçu pendant un temps fixe avant de l'envoyer, il transforme le délai variable en délai fixe. Ainsi l'effet de la gigue se traduit pour l'utilisateur :

- si aucun buffer de gigue n'est utilisé : par un délai variable qui s'ajoute au délai fixe de la transmission,
- si un buffer de gigue est utilisé : par un délai fixe, propre au buffer, et par des pertes de paquets supplémentaires dues au fait que le buffer de gigue élimine les paquets arrivés trop tard pour être transmis de façon fluide.

Une augmentation de la gigue, dans le cas où un buffer de gigue est utilisé, résulte soit en une augmentation du délai soit en une augmentation des pertes de paquets, en fonction du buffer utilisé à la réception.

1.2.3.8 Double parole

La double parole n'est pas une dégradation en soi, mais une situation qui peut se produire en contexte de conversation, quand les deux interlocuteurs parlent en même temps. Cette situation particulière a alors une incidence sur la perception de certaines dégradations par les participants à la conversation. Ainsi la perception de l'écho, de l'affaiblissement en sonie, des dégradations introduites par les systèmes de traitement du signal, etc. s'en trouvera changée. Peu de tests sont disponibles sur la double parole.

1.2.3.9 Dispositifs de traitement du signal

De nombreux dispositifs de traitement du signal peuvent intervenir dans un système de communication, tels que des annuleurs d'écho, des détecteurs d'activité vocale / générateurs de bruit de confort ou des réducteurs de bruit. Leurs effets sur la qualité sont difficilement évaluables puisqu'ils ont pour but principal de l'améliorer et ne l'affectent que lorsqu'ils ne fonctionnent pas correctement. Cependant quelques effets peuvent d'ores et déjà être donnés qualitativement.

Les annuleurs d'écho : ils doivent estimer le signal d'écho et sont souvent conçus pour le trouver dans une certaine fenêtre temporelle et un certain intervalle d'amplitude. Si le signal d'écho ne correspond pas à ces critères, l'annuleur d'écho peut contribuer à la distorsion du signal en ne retirant pas l'écho ou en convergeant vers une mauvaise estimation de l'écho. Un autre point intéressant de mauvaise performance des annuleurs d'écho est la condition de double parole. Dans cette situation, l'annuleur d'écho peut diverger et la parole interrompante est alors distordue (*cf.* paragraphe 1.2.3.8).

Les détecteurs d'activité vocale / générateurs de bruit de confort : un détecteur d'activité vocale est un système placé à l'émission qui sert à discriminer les instants de silence et de parole dans les signaux de conversation et supprime la paquetisation des signaux vocaux pendant les périodes de silence qui représentent environ 50% de la durée d'une conversation. Cela permet d'économiser une quantité importante de bande passante. Le détecteur supprime les parties du signal inférieures à un certain seuil du rapport signal-à-bruit. Le risque est alors qu'une partie du signal correspondant à de la

parole soit lui aussi supprimé, introduisant des distorsions dans le signal, des coupures en début ou fin de mots (« clipping ») et une perte d'intelligibilité. Des générateurs de bruit de confort peuvent y être ajoutés du côté réception afin de remplacer les périodes de silence par un bruit, qui permet d'éviter que l'auditeur ait l'impression que la communication a été coupée. La qualité d'un tel dispositif est alors déterminée par la ressemblance entre le bruit de confort et le bruit de fond naturel.

Les réducteurs de bruit : ils ont pour but d'estimer le bruit de fond et de le réduire voire de le supprimer. En cas de mauvaise estimation du bruit, le signal de parole peut être endommagé (niveau de parole réduit, modification spectrale, etc.).

Peu de tests subjectifs sont cependant disponibles sur ce type de dispositifs, le lecteur pourra se référer en particulier à l'étude effectuée à France Télécom par Nicolas Le Faucheur sur la qualité des réducteurs de bruit [Le Faucheur 2004].

1.2.3.10 Synthèse

Ce bilan montre l'étendue des dégradations rencontrées dans les différents systèmes de télécommunications. Certaines données subjectives sont manquantes ou incomplètes, concernant par exemple le délai, la double parole, les dispositifs de traitement du signal ou encore la combinaison des dégradations. De plus, la majorité des tests disponibles dans la littérature sont des tests d'écoute : il y a peu de tests de locution et de conversation. Une revue complète des tests subjectifs recensés pour rédiger ce bilan est disponible dans [Barriac et Guéguin 2006].

1.2.4 Limites de l'évaluation subjective

Les tests subjectifs sont indispensables pour l'évaluation de la qualité vocale, puisqu'ils représentent le jugement humain de la qualité vocale. Cependant, les tests subjectifs nécessitent de mobiliser beaucoup de moyens (temps, personnes et argent). Les méthodes objectives se présentent comme une alternative aux méthodes subjectives et permettent d'automatiser l'évaluation de la qualité vocale. Néanmoins, elles doivent présenter une forte corrélation avec les résultats des tests subjectifs, qui représentent le jugement des utilisateurs. Qui dit modélisation objective, dit donc données subjectives pour « alimenter » le modèle.

1.3 Modèles objectifs de la qualité vocale

L'évaluation objective de la qualité vocale s'est tout d'abord effectuée avec des outils simples de traitement du signal tels que le rapport signal-à-bruit (RSB), le rapport signal-à-bruit segmental (RSBseg), l'erreur quadratique moyenne (EQM) pour les mesures temporelles, et la distance cepstrale, la distance spectrale pour les mesures fréquentielles. Ces mesures objectives simples ne se corrèlent pas bien avec les données subjectives [Barnwell 1980], puisque la qualité vocale est affectée par des dégradations complexes, qui, de plus, peuvent être masquées ou amplifiées par la présence d'autres dégradations. Le recours à des méthodes objectives plus élaborées a donc été nécessaire. Désormais, de nombreux modèles objectifs de la qualité vocale existent. Nous présenterons dans ce paragraphe les différents types de modèles objectifs et les principaux modèles de la qualité vocale existants.

Les modèles objectifs de la qualité vocale peuvent être classés selon :

- le fait qu'ils se basent sur des mesures physiques du système (paramétriques) ou sur les signaux,
- le besoin qu'ils ont d'avoir accès aux informations des deux côtés du système (bout en bout ou avec référence) ou d'un seul côté seulement (mono-extrémité ou sans référence),
- le contexte dans lequel ils fonctionnent (écoute, locution ou conversation).

Il est de plus important de distinguer deux méthodes de mesure du système testé :

– les méthodes de mesure intrusives, qui sont utilisées dans les modèles avec référence, et qui consistent à faire une communication « test » avec le système étudié, en envoyant un signal de référence et en recueillant le signal dégradé à la sortie du système testé,
– les méthodes de mesure non intrusives, qui n'ont besoin que du signal dégradé (modèle sans référence) et qui utilisent les signaux échangés dans les réseaux réels sans perturber les communications.

La figure 1.15 classe les différents modèles de la qualité vocale existants en fonction de ces trois critères.

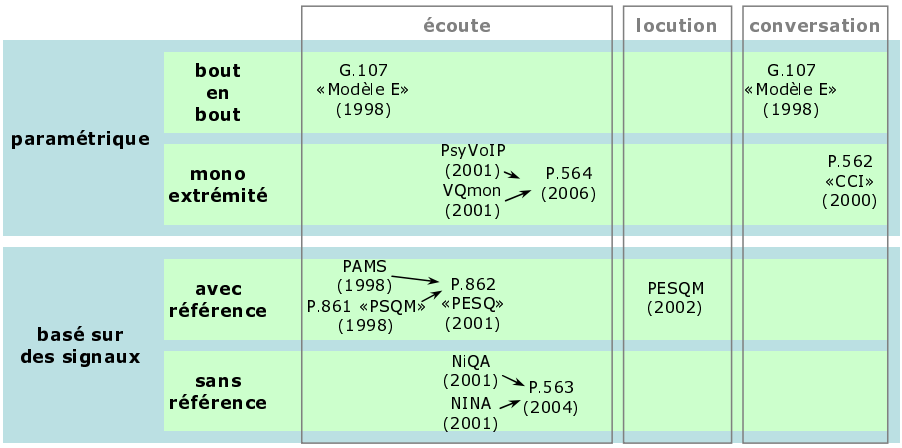


Figure 1.15 : *Classification des modèles objectifs de la qualité vocale existants. Le sigle P.xxx ou G.xxx désigne la recommandation correspondante à l'UIT-T, le nom usuel du modèle est entre guillemets.*

1.3.1 Modèles paramétriques

Les modèles paramétriques utilisent des mesures physiques du système à évaluer pour estimer une note de qualité vocale.

1.3.1.1 Modèle E

Parmi les modèles paramétriques, le modèle E est le plus utilisé. Il a été développé par l'ETSI [ETSI ETR 250 1996] comme un outil bout en bout pour les concepteurs de réseaux et normalisé en 1998 à l'UIT-T dans la Recommandation G.107 [UIT-T Rec. G.107 2005]. En pratique, il est utilisé par les planificateurs de réseaux pour répartir les « éléments de qualité » (annuleurs d'écho, codecs, etc.) le long de la chaîne de transmission afin d'atteindre la qualité optimale pour l'utilisateur [Möller et Raake 2002]. De nombreux tests subjectifs ont été nécessaires à son optimisation. La qualité de transmission de bout en bout est exprimée en sortie du modèle E par un scalaire, appelé « facteur d'évaluation de transmission » et noté R, compris entre 0 et 100, qui varie directement avec la qualité globale conversationnelle

$$R = R_o - I_s - I_d - I_{e,eff} + A \tag{1.1}$$

où R_o représente le rapport signal-à-bruit basique, incluant les sources de bruit telles que le bruit de circuit et le bruit de salle. Le facteur I_s est une combinaison de toutes les dégradations qui se produisent plus ou moins simultanément avec le signal vocal. Le facteur I_d représente les dégradations causées par le délai et l'écho, et le facteur de dégradation due à l'équipement effectif $I_{e,eff}$ représente les dégradations causées par les codecs bas-débit et inclut aussi la dégradation due aux pertes de paquets aléatoires et en rafale. Le facteur d'avantage A permet au modèle E de prendre en compte le fait que les clients peuvent accepter une certaine baisse de qualité en échange d'un avantage d'accès (par exemple, la mobilité ou les connexions dans les régions difficiles à atteindre). Les facteurs intervenant dans le calcul du facteur R sont

estimés à partir de mesures physiques des deux côtés du système à évaluer telles que le délai, l'écho, l'atténuation, le bruit de salle, etc. Ce facteur R peut être transformé pour obtenir une estimation des réactions des utilisateurs sous différentes formes, comme les notes MOS, le pourcentage de connexions au moins bonnes (GoB, *Good or Better*) ou le pourcentage de connexions au mieux médiocres (PoW, *Poor or Worse*).

Cependant, il est bien précisé dans la Recommandation G.107 [UIT-T Rec. G.107 2005] que :

> « *de telles estimations sont seulement faites dans le but de la planification de transmission et non pour la prédiction de l'opinion des utilisateurs* ».

1.3.1.2 Modèle CCI

L'équivalent du modèle E en mono-extrémité est le modèle appelé CCI (*Call Clarity Index*), décrit dans la Recommandation P.562 de l'UIT-T [UIT-T Rec. P.562 2004]. Il permet d'évaluer la qualité de conversation à partir de mesures du système (*e.g.* niveau de parole, niveau de bruit, atténuation de l'écho) effectuées par des sondes non intrusives appelées les INMDs (*In-service Non-intrusive Measurement Devices*), décrites dans la Recommandation P.561 [UIT-T Rec. P.561 2002]. Ce modèle permet d'interpréter les mesures faites par les INMDs pour prédire la qualité de conversation, telle que perçue par chaque utilisateur de la communication, en faisant des hypothèses sur le réseau et sur les utilisateurs de chaque extrémité. Son principe général est donné dans la figure 1.16.

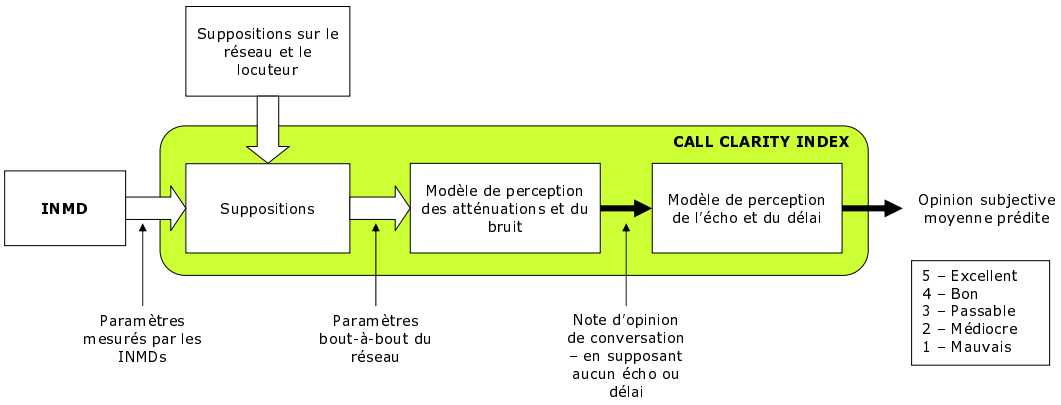


Figure 1.16 : *Principe de fonctionnement du modèle non intrusif CCI (*Call Clarity Index*) [UIT-T Rec. P.562 2004]*

1.3.1.3 Modèle P.564

Un autre modèle mono-extrémité de la qualité d'écoute est décrit dans la Recommandation P.564 [UIT-T Rec. P.564 2006]. Il fixe les objectifs de performances qui doivent être atteints par des modèles tels que PsyVoIP [Rix *et al.* 2001] et VQmon [Clark 2001]. Le but de ce type de modèle est de se baser sur les informations des paquets IP sans utiliser les données vocales contenues dans le flot IP, qui sont longues à désencapsuler des paquets. Ce modèle peut être utilisé dans la surveillance en temps réel de la qualité des réseaux IP. Il estime des paramètres intermédiaires de qualité (taux de perte de paquets, type de perte de paquets et gigue) à partir des informations contenues dans l'en-tête du protocole RTP (*Real-Time Protocol*). La note de qualité d'écoute est ensuite estimée à partir de ces paramètres.

1.3.1.4 Avantages et limites

L'avantage des modèles paramétriques est leur rapidité, ils peuvent donc être facilement embarqués dans des éléments du réseau et les terminaux. Cependant, ils n'atteignent pas les

mêmes performances que les modèles basés sur des signaux.

1.3.2 Modèles basés sur les signaux avec référence

Les modèles basés sur les signaux avec référence envoient un signal connu (référence) à travers le système à tester, recueillent le signal en sortie du système (signal généralement appelé « signal dégradé »), et comparent ces deux signaux afin d'en déduire une note objective de qualité, qui doit être bien corrélée avec la note subjective.

Parmi les modèles avec référence, les plus utilisés sont les modèles basés sur une comparaison des transformations internes propres à l'oreille humaine, appelés modèles perceptuels. Cette méthode consiste à transformer la représentation physique d'un signal (mesurée en décibels, secondes, Hertz) en une représentation psychoacoustique (mesurée en sones, secondes, barks) et est basée sur les travaux de Zwicker et Feldtkeller sur la psychoacoustique [Zwicker et Feldtkeller 1981], présentés dans l'annexe A. En particulier, il a été montré que la perception du bruit dû au codage change en fonction de la distribution spectrale du bruit par rapport à celle du signal de parole. Le bruit peut ainsi être « caché » sous le spectre du signal en exploitant les propriétés de masquage de l'oreille humaine. Schroeder et *al.* ont utilisé cette propriété de masquage auditif pour améliorer la qualité des codeurs de parole [Schroeder *et al.* 1979]. Pour cela, ils ont introduit une mesure objective de la dégradation du signal de parole par un bruit de quantification en mesurant le rapport entre la sonie relative du bruit en présence du signal de parole et la sonie du signal de parole.

Les modèles perceptuels constituent dorénavant l'approche dominante pour les modèles avec référence [Rix 2004]. Les premiers modèles perceptuels, fonctionnant tous en contexte d'écoute, ont été développés dans les années 90 pour évaluer la qualité des codecs audio. Deux types de modèles coexistaient : les modèles à base de décomposition du signal audio en trames temporelles et de FFT (*e.g.* [Brandenburg 1987, Beerends et Stemerdink 1992, Paillard *et al.* 1992, Colomes *et al.* 1995]) et les modèles à base de bancs de filtres auditifs (*e.g.* [Thiede et Kabot 1996, Sporer 1997]). Leur réunion a abouti au modèle PEAQ (*Perceptual Evaluation of Audio Quality*), normalisé à l'UIT-R dans la Recommandation BS.1387 [UIT-R Rec. BS.1387 1998] et fonctionnant avec les deux versions : FFT et bancs de filtres auditifs.

Partant de leur modèle dans le domaine audio [Beerends et Stemerdink 1992], Beerends et Stemerdink ont développé un outil similaire pour le domaine de la parole, appelé PSQM (*Perceptual Speech Quality Measure*) [Beerends et Stemerdink 1994] et anciennement normalisé par l'UIT-T sous le nom P.861 [UIT-T Rec. P.861 1998]. Le modèle PSQM se base sur une comparaison des représentations internes des signaux de référence et dégradé, préalablement alignés temporellement par intercorrélation. En parallèle de PSQM, Rix et *al.* ont développé le modèle perceptuel appelé PAMS (*Perceptual Analysis Measurement System*) [Rix *et al.* 1999]. Ce modèle est basé, contrairement au modèle PSQM, sur des bancs de filtres auditifs et a l'avantage, par rapport à PSQM, d'intégrer un algorithme d'alignement temporel des signaux de référence et dégradé plus robuste aux délais variables rencontrés en VoIP, ce qui en étend le champ d'application.

La mise en commun des avantages des deux modèles (modèle psychoacoustique de PSQM et alignement temporel de PAMS) a permis de créer le modèle PESQ (*Perceptual Evaluation of Speech Quality*), normalisé à l'UIT-T sous le nom P.862 [UIT-T Rec. P.862 2001] et qui a remplacé PSQM en tant que norme. Le modèle PESQ utilise donc le même principe que le modèle PSQM au niveau psychoacoustique, présenté dans le paragraphe suivant.

1.3.2.1 Transformation par représentation interne

Cette transformation des signaux par représentation interne est composée de trois phases (*cf.* figures 1.17 et 1.18).

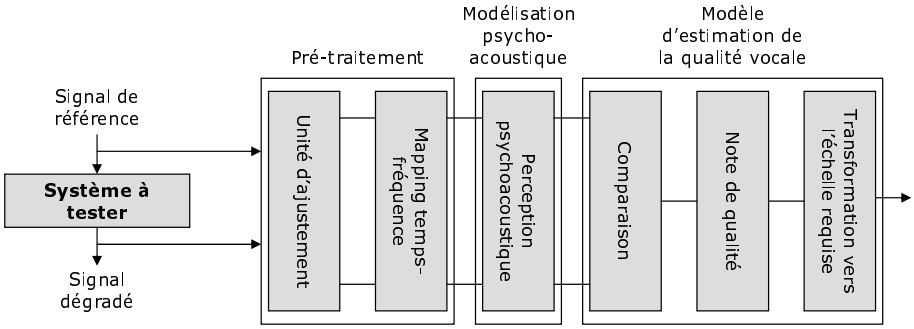


Figure 1.17 : *Principe de fonctionnement des modèles basés sur une comparaison des transformations internes d'après [ETSI Guide 201 377-1 2002]*

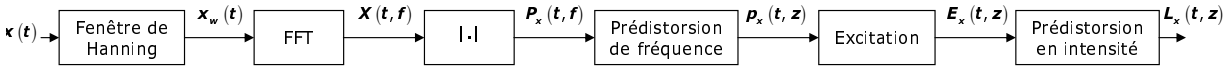


Figure 1.18 : *Transformation par représentation interne*

Pré-traitement La transformation en représentation interne est effectuée sur les deux signaux de référence $x(t)$ et dégradé $y(t)$.

- *Unité d'ajustement* : il est nécessaire de faire un étalonnage entre le niveau d'écoute et la sonie comprimée (*cf.* annexe A). Pour cela, une sinusoïde de référence avec une fréquence de 1 kHz et une amplitude de 40 dB SPL (c'est-à-dire -64 dBov, et une amplitude zéro à crête de 29.54)[4] est générée. La valeur maximale de la représentation de puissance fondamentale contenue dans la sinusoïde d'étalonnage est échelonnée à 10^4 (c'est-à-dire, si la valeur $max_z(p_x(t,z)) = 1$ pour un niveau acoustique de 0 dB, cette valeur $max_z(p_x(t,z)) = 10000$ pour un niveau acoustique de 40 dB). Ce facteur d'échelonnement en puissance est calculé comme suit

$$S_p = \frac{10000}{\max_z(p_x(t,z))} \tag{1.2}$$

lorsque la valeur $p_x(t,z)$ est calculée pour la sinusoïde de référence.
Le deuxième étalonnage fixe à la valeur de 1 sone la sonie comprimée de la sinusoïde de référence. Ce facteur est calculé comme suit

$$S_l = \frac{1}{L_x(t,z)} \tag{1.3}$$

lorsque la valeur $L_x(t,z)$ est calculée pour la sinusoïde de référence.
Il est aussi nécessaire d'échelonner les données d'entrée à un niveau de conversation active de -26 dBov ou au niveau acoustique d'écoute de 79 dB SPL [UIT-T Rec. P.56 1993].

- *Mapping temps-fréquence* : le signal $x(t)$ dans la trame i est pondéré par une fenêtre de Hanning

$$x_{w_i}[n] = x_i[n]w[n] \tag{1.4}$$

où $w[n] = 0.5\left(1 - \cos\left(\frac{2\pi i}{N_f}\right)\right)$ pour $0 \leq i \leq N_f - 1$.

La densité spectrale de puissance du signal $x_{w_i}[n]$, représentée par $P_{x_i}[k]$ est calculée au moyen de la FFT

[4] Dans un fichier vocal codé sur 16 bits, le niveau 0 dBov est représenté par une composante constante de 32767 ($= 2^{16}/2 - 1$). Une onde sinusoïdale ayant une amplitude zéro à crête de 32767 aura donc un niveau efficace de -3.01 dBov, correspondant à un niveau acoustique d'environ 101 dB.

$$P_{x_i}[k] = (\Re(X_i[k]))^2 + (\Im(X_i[k]))^2 . \tag{1.5}$$

où $X_i[k] = FFT(x_{w_i}[n])$.

Modélisation psychoacoustique La modélisation psychoacoustique des deux signaux de référence $x(t)$ et dégradé $y(t)$ est faite d'après le processus suivant.

- *Prédistorsion fréquentielle* : la prédistorsion permet de passer de l'échelle en Hertz à l'échelle des bandes critiques, pour obtenir une représentation de la densité de puissance fondamentale échantillonnée trame par trame. L'indice de fréquence k, exprimé en Hertz, est transformé en indice tonal j dans le domaine des bandes critiques, au moyen d'une prédistorsion de l'échelle des fréquences. L'échelle des bandes critiques est d'abord subdivisée en bandes d'intervalle égal et, pour chaque bande, une valeur (échantillon) de densité de puissance fondamentale est calculée à partir des échantillons de densité de puissance spectrale dans la bande correspondante sur l'échelle en Hertz. La densité de puissance fondamentale échantillonnée, $P_{x_i}[j]$ pour la bande j dans la trame i est donnée par la formule suivante

$$P_{x_i}[j] = S_p \frac{\Delta f_j}{\Delta z} \frac{1}{I_l[j] - I_f[j] + 1} \sum_{k=I_f[j]}^{I_l[j]} P_{x_i}[k] \tag{1.6}$$

où $I_f[j]$ est l'indice du premier échantillon et $I_l[j]$ celui du dernier sur l'échelle hertzienne pour la bande j, Δf_j étant la largeur de la bande j en Hertz, Δz étant la largeur de chaque sous-bande dans le domaine des bandes critiques, et S_p étant le facteur d'échelonnement en puissance, calculé précédemment.
La puissance totale du signal dans une trame i notée P_{x_i} est calculée comme suit

$$P_{x_i} = \sum_{j=1}^{N_b} P_{x_i}[j] \tag{1.7}$$

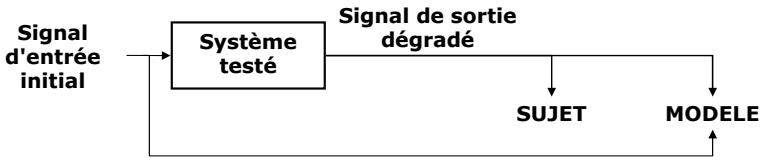
où N_b est le nombre total de bandes critiques.

- *Prédistorsion d'intensité* : après le calcul de la densité de puissance fondamentale échantillonnée, l'échelle des intensités est transformée en échelle des sonies de façon à obtenir une fonction de densité de sonie comprimée et échantillonnée. La densité de sonie comprimée et échantillonnée, $L_{x_i}[j]$, est calculée à partir de la densité de puissance fondamentale $P_{x_i}[j]$, au moyen d'une fonction de compression donnée comme suit par Zwicker [Zwicker et Feldtkeller 1981]

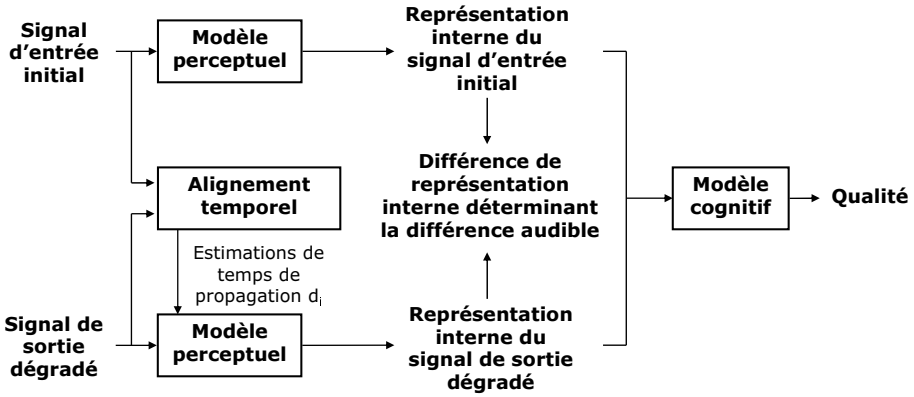
$$L_{x_i}[j] = S_l \left(\frac{P_0[j]}{0.5}\right)^{\gamma} \left(\left(0.5 + 0.5\frac{P_{x_i}[j]}{P_0[j]}\right)^{\gamma} - 1\right) \tag{1.8}$$

où $P_0[j]$ est le seuil d'audition interne et S_l est le facteur d'étalonnage en sonie fondamentale. Les valeurs négatives de $L_{x_i}[j]$ sont mises à zéro. La valeur optimale de γ est trouvée lors de l'optimisation des bases de données construites par différentes expériences d'évaluation de la qualité de la parole.
La valeur instantanée (totale) de la sonie comprimée L_{x_i} (exprimée en sones comprimés) est calculée par sommation de la densité comprimée et échantillonnée $L_{x_i}[j]$

$$L_{x_i} = \sum_{j=1}^{N_b} L_{x_i}[j]\Delta z. \tag{1.9}$$



(a) Le sujet juge le signal dégradé, en comparaison avec sa référence de qualité interne



(b) Le modèle PESQ transforme le signal dégradé et le signal de référence correspondant en représentations internes, puis utilise leur différence pour calculer une note de qualité d'écoute

Figure 1.19 : *Principe de fonctionnement du modèle PESQ d'après [UIT-T Rec. P.862 2001]*

Modèle d'estimation de la qualité de parole Les sonies comprimées des signaux de référence $x(t)$ et dégradé $y(t)$ sont ensuite comparées pour déterminer une note de qualité, enfin transformée sur l'échelle adéquate.

1.3.2.2 Modèle PESQ

Son principe de fonctionnement est présenté dans la figure 1.19 et détaillé dans l'annexe B. Lors d'un test subjectif d'écoute, le sujet juge le signal de sortie dégradé par le système à tester, sans avoir accès au signal de référence correspondant à la phrase prononcée. La comparaison s'effectue avec les conditions de référence (*i.e.* sans dégradation) présentées durant le test, ainsi qu'avec la référence propre à chaque sujet. Le modèle PESQ a accès pour chaque condition testée aux signaux de référence et dégradé, qu'il transforme et compare pour obtenir une note de qualité d'écoute.

Le modèle PESQ permet d'évaluer la qualité d'écoute dans de nombreuses conditions de dégradation (perte de paquets, distorsion due au codage, bruit ambiant du côté émission, variation du délai), aboutissant à une corrélation de 0.935 à la fois pour les 22 tests subjectifs ayant servi à l'élaboration du modèle et pour les 8 tests subjectifs de validation du modèle.

Une extension de PESQ au domaine acoustique (avec prise en compte des terminaux) en bande élargie (de 50 à 7000 Hz, au lieu de 300 à 3400 Hz en bande étroite) est en cours d'étude à l'UIT-T sous le nom provisoire P.OLQA (*Objective Listening Quality Assessment*) [Berger 2005]. Une version bande élargie de PESQ a par ailleurs été adoptée en 2005 dans la Recommandation P.862.2 [UIT-T Rec. P.862.2 2005].

1.3.2.3 Modèle PESQM

Parmi les modèles objectifs présentés dans ce chapitre, le modèle perceptuel PESQM (*Perceptual Echo and Sidetone Quality Measure*) [Appel et Beerends 2002] est le seul à évaluer la qualité dans le contexte de locution d'un système de communications potentiellement affecté par de l'écho et/ou une distorsion de l'effet local. Ce modèle fonctionne sur le même principe que le modèle PESQ en comparant un signal dégradé avec le signal de référence correspondant.

Dans le contexte de locution, le signal de référence est le signal prononcé par le participant dans le microphone et le signal dégradé est le signal retourné par le système dans le haut-parleur du même participant, pouvant donc contenir de l'écho et/ou un effet local distordu. La difficulté d'évaluer la qualité de locution réside tout d'abord dans la définition du signal de référence à utiliser en entrée du modèle [Appel et Beerends 2002]. En effet en situation de locution, chacun utilise sa propre voix comme référence, alors que lors d'un test d'écoute les signaux de référence et dégradé sont les mêmes pour tous les participants.

Deux approches sont proposées par les auteurs. La première approche consiste à enregistrer au niveau *électrique* le signal d'entrée du système, considéré comme le signal de référence, et le signal retourné par le système. L'avantage de cette approche est sa simplicité d'un point de vue pratique, l'inconvénient est qu'elle ne permet pas de prendre en compte les dégradations acoustiques qui auraient pu gêner le participant lors du jugement (caractéristiques acoustiques du combiné, de la salle, etc.). La seconde approche consiste à utiliser un « simulateur de tête et torse » (HATS, *Head And Torso Simulator*) [UIT-T Rec. P.58 1996] pour simuler les caractéristiques d'un locuteur. Un signal de parole enregistré auparavant est appliqué à la bouche du HATS, ce qui permet d'obtenir une voix naturelle. En plaçant un combiné téléphonique sur le HATS, les signaux à son oreille peuvent être enregistrés au niveau *acoustique*, contenant ainsi les signaux d'effet local et d'écho. L'avantage de cette approche est que les signaux de référence et dégradé sont proches de la réalité, l'inconvénient est qu'elle nécessite l'utilisation d'un HATS.

Une fois l'approche choisie, les signaux de référence et dégradé sont envoyés en entrée du modèle, transformés en représentations internes et comparés. La qualité de locution est essentiellement affectée par l'écho et la distorsion de l'effet local. Or ceux-ci peuvent être masqués par le bruit venant de l'autre côté de la communication. Ainsi, en contexte de locution, la présence d'un bruit peut améliorer la qualité (si celui-ci n'est pas trop fort pour être gênant). Le modèle PESQM intègre donc un module d'estimation du bruit, combiné à un module de suppression du bruit qui utilise un seuil en deçà duquel le bruit n'est pas gênant (et même bénéfique).

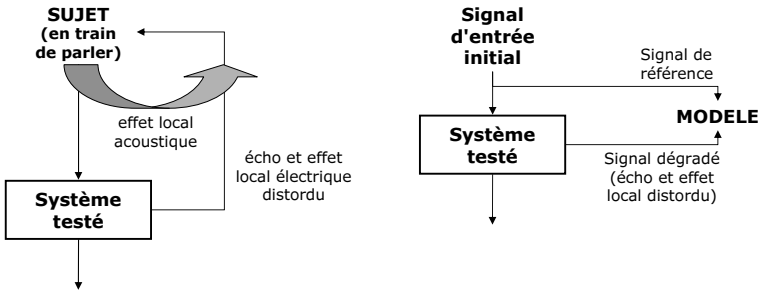
Le principe de fonctionnement du modèle PESQM est présenté dans la figure 1.20 et détaillé dans l'annexe C. Ce modèle est à la fois un modèle avec référence puisqu'il compare le signal dégradé au signal de référence et un modèle mono-extrémité puisqu'il ne nécessite d'avoir accès aux informations que d'un seul côté du système. Sa mise en œuvre pratique et son optimisation seront abordées dans le chapitre 4.

1.3.2.4 Avantages et limites

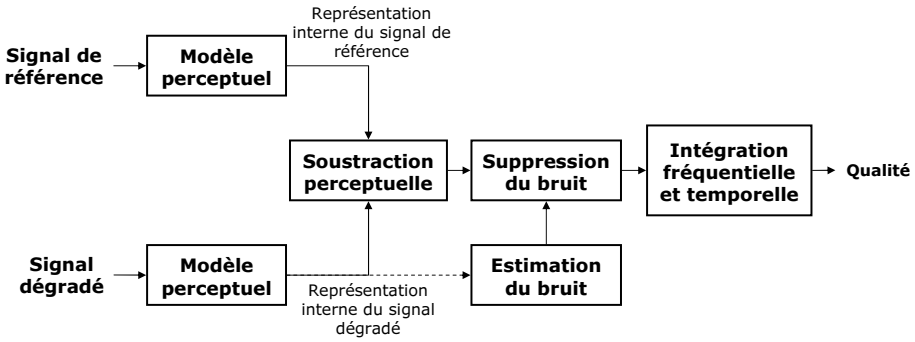
L'avantage des modèles basés sur les signaux avec référence est leur performance à estimer les notes subjectives de qualité telle qu'elle est perçue par les utilisateurs. Cependant, ils nécessitent des signaux de test (envoi d'une référence dans le système) et ne peuvent donc pas être utilisés sur des communications réelles sans les perturber.

1.3.3 Modèles basés sur les signaux sans référence

Les méthodes mono-extrémité permettent l'analyse des signaux sans référence connue. L'équivalent de PESQ en mono-extrémité a été normalisé par l'UIT-T sous le nom P.563 [UIT-T Rec. P.563 2004], à partir des modèles NiQA (*Non-intrusive speech Quality Assessment*) [Rix et Gray 2001] et NINA (*Non Intrusive Network Assessment*) [Juric 2001]. Le modèle P.563 permet d'évaluer la qualité d'écoute dans de nombreuses conditions de dégradation (distorsion due aux annuleurs d'écho ou aux systèmes de réduction de bruit, perte de paquets, distorsion due au codage et bruit ambiant du côté émission), aboutissant à une corrélation proche de 0.89 avec les données subjectives. Son cadre d'application est limité à des mesures moyennes, il ne fonctionne donc pas pour des mesures individuelles du fait de sa variabilité. Le principe de ce modèle, décrit dans la figure 1.21, est de détecter les trames de parole dans le signal



(a) Le sujet en train de parler perçoit sa propre voix, l'effet local acoustique et l'écho et/ou l'effet local électrique éventuellement retournés par le système testé



(b) Le modèle PESQM transforme le signal de référence et le signal dégradé (différents selon l'approche choisie) en représentations internes puis les compare pour calculer une note de qualité

Figure 1.20 : *Principe de fonctionnement du modèle PESQM d'après [van Vugt 2005]*

dégradé et d'en extraire un ensemble de paramètres permettant de faire une analyse du conduit vocal et du caractère non naturel de la voix, une analyse des bruits additionnels intenses, une analyse des interruptions, silences et écrêtages temporels. La note de qualité vocale finale est calculée en faisant une combinaison linéaire des différents résultats de l'évaluation de la qualité intermédiaire avec certaines caractéristiques additionnelles du signal.

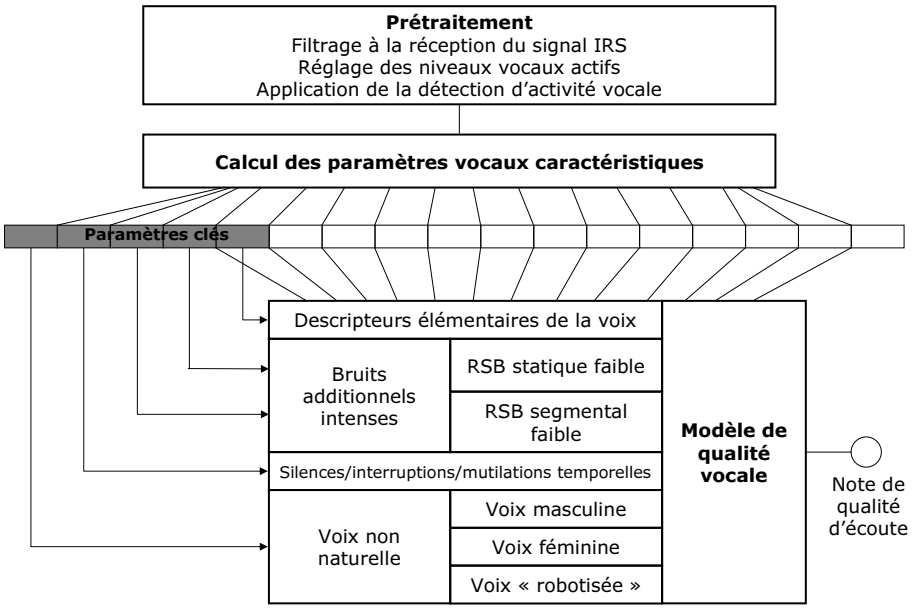


Figure 1.21 : *Principe de fonctionnement du modèle mono-extrémité basé sur les signaux [UIT-T Rec. P.563 2004]*

1.3.4 Évaluation des mesures objectives de la qualité vocale

Une évaluation des performances des modèles objectifs est nécessaire. Les notes objectives de qualité qu'ils fournissent doivent être en adéquation avec les notes subjectives correspondantes. Traditionnellement, les performances des modèles objectifs de la qualité vocale sont évaluées grâce au calcul du coefficient de corrélation linéaire (dit coefficient de Pearson, noté r). Le coefficient de Pearson entre notes subjectives et objectives est formulé par

$$r = \frac{\sum \left[\left(X - \bar{X} \right) \left(Y - \bar{Y} \right) \right]}{\sqrt{\sum \left(X - \bar{X} \right)^2 \sum \left(Y - \bar{Y} \right)^2}} \tag{1.10}$$

où X est le vecteur des notes subjectives, Y le vecteur des notes objectives correspondantes, $\bar{X}$ la moyenne des notes subjectives et $\bar{Y}$ la moyenne des notes objectives. Par définition, le coefficient de corrélation linéaire est une mesure de la relation linéaire entre deux variables et de plus suppose que les deux variables observent une distribution gaussienne.

Cependant, l'absence de relation linéaire (*i.e.* $r = 0$) ne signifie pas l'absence de toute relation entre les deux variables étudiées. Il peut donc être intéressant de calculer en supplément le coefficient de corrélation de rang (dit coefficient de Spearman, noté r_s), qui étudie la relation non linéaire monotone entre deux variables. Il fonctionne sur l'étude de la différence des rangs entre les observations des deux variables X et Y

$$r_s = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} \tag{1.11}$$

où D est la différence entre les rangs des valeurs correspondantes de X et de Y, et N le nombre de paires de valeurs.

Les coefficients de corrélation indiquent une tendance des deux variables à varier de manière corrélée. Cependant ils ne permettent pas la comparaison entre différentes bases de données, puisque le calcul des coefficients de corrélation ne prend pas en compte la dispersion des données [Yang 1999]. Il est donc nécessaire d'évaluer l'erreur entre les notes objectives et les notes subjectives. Pour cela, en règle générale, l'erreur absolue moyenne, notée EAM, entre les notes subjectives et objectives est calculée telle que

$$EAM = \frac{1}{N} \sum |X - Y| \tag{1.12}$$

où N est le nombre de paires de valeurs. L'avantage de l'erreur absolue moyenne est de s'exprimer dans la même unité que les variables X et Y, à savoir ici en MOS.

Conclusion

Dans ce chapitre, nous avons présenté la nécessité d'évaluer la qualité vocale dans les télécommunications. Les méthodes subjectives, bien qu'étant le seul moyen d'atteindre le jugement humain de la qualité vocale, sont contraignantes à mettre en œuvre et coûteuses. Des méthodes objectives ont donc été développées comme alternative, accessible au plus grand nombre, aux méthodes subjectives d'évaluation de la qualité vocale.

L'état de l'art des méthodes objectives d'évaluation de la qualité de la parole a présenté les différentes approches utilisées et l'étendue des modèles objectifs existants. Les modèles paramétriques ont été conçus pour être facilement utilisés lors de la planification des réseaux, mais ne donnent pas une estimation de la qualité vocale telle qu'elle est perçue par les utilisateurs. Les modèles basés sur la comparaison des représentations internes des signaux de référence et dégradé modélisent la perception auditive humaine, aboutissant ainsi aux meilleures performances. Les modèles basés sur l'analyse des signaux sans référence estiment la qualité vocale uniquement à partir du signal dégradé et peuvent ainsi être utilisés dans le trafic réel tout en fournissant des performances honorables pour prédire la qualité vocale.

Parmi ces modèles, la plupart s'intéresse à la qualité vocale en contexte d'écoute, et l'un d'entre eux évalue la qualité vocale en contexte de locution. Cependant, il est très rare dans la réalité qu'un usager soit placé soit uniquement en contexte d'écoute, soit uniquement en contexte de locution, si ce n'est par exemple lors de communications avec un répondeur téléphonique. Ainsi, il existe peu de modèles objectifs de la qualité vocale dans le contexte le plus courant pour les usagers, à savoir le contexte de conversation. Les seuls modèles évaluant la qualité de conversation sont des modèles paramétriques (modèles E et CCI), qui ne sont pas efficaces pour estimer la qualité telle qu'elle est perçue par l'utilisateur. Il manque donc un modèle basé sur l'analyse des signaux (avec ou sans référence) capable de fonctionner dans le contexte de conversation et de prédire avec fiabilité la qualité subjective de conversation.

Chapitre 2

Problématique, objectifs et méthode proposée

Introduction

Le chapitre précédent a mis en évidence le manque d'un outil performant d'évaluation objective de la qualité vocale dans le contexte le plus courant pour les usagers : la conversation. Dans ce chapitre, nous présenterons tout d'abord la problématique posée par l'élaboration d'un tel modèle objectif non paramétrique de la qualité vocale en contexte de conversation. Après avoir fixé les objectifs à atteindre par ce modèle, nous décrirons la méthode proposée dans la thèse pour construire un modèle objectif de la qualité vocale de conversation.

2.1 Problématique

Il a été vu dans le chapitre 1 que les modèles paramétriques E et CCI ne sont pas suffisamment efficaces pour estimer la qualité de conversation. Il manque donc un modèle non paramétrique de la qualité vocale, telle qu'elle est perçue par l'utilisateur final, dans le contexte de conversation. Un tel modèle peut être implémenté :

- dans des sondes sans intrusion pour la supervision de la qualité de transmission dans les réseaux (essentiellement en technologie IP),
- dans des sondes avec intrusion,
- dans des éléments de réseau sous forme d'agent logiciel, avec la même finalité que précédemment mais sans nécessité d'outil de mesure externe,
- au niveau des installations des clients (réseaux locaux, voire même terminaux) sous forme d'agent logiciel, pour la supervision de la qualité de service basée sur une capture d'informations directement au niveau du client et ce de façon répartie dans le réseau.

L'UIT-T s'intéresse depuis janvier 2005 à la modélisation de la qualité de conversation à la Question 20 de la Commission d'Études 12, en vue de la normalisation d'un modèle objectif nommé provisoirement P.CQO (*Conversational Quality Objective*). Son but final, décrit dans le document [Pomy 2005], est de normaliser un modèle objectif de la qualité de conversation qui prédit l'impact des dégradations du réseau sur la qualité conversationnelle ressentie par l'utilisateur final. Dans un premier temps, elle fixe les objectifs à atteindre par les modèles proposés par les différents contributeurs, puis sélectionnera dans un second temps le meilleur modèle ou la meilleure combinaison de modèles. À l'heure actuelle, les objectifs fixés sont les suivants :

- P.CQO modélise une connexion électrique/électrique, *i.e.* il ne prend en compte que les dégradations au niveau électrique et pas celles au niveau acoustique, telles que celles introduites par les terminaux,

- P.CQO est limité aux applications bande étroite, *i.e.* bande téléphonique entre 300 et 3400 Hz,
- P.CQO prédit la qualité de conversation appel par appel, sur la base de quelques minutes de communication.

2.2 Objectifs

Nous avons choisi de conformer le modèle objectif proposé dans cette thèse aux différents objectifs et contraintes posés par la Question 20. Nous nous sommes fixés comme objectif supplémentaire de construire un modèle non paramétrique, s'appuyant sur l'analyse des signaux échangés durant la communication testée. Le dernier choix concerne le type de mesure souhaité en entrée du modèle (intrusive ou non intrusive), qui n'a pas encore été défini par la Question 20. Les deux types de mesure présentent des avantages et des inconvénients selon le type d'application visé :

- les mesures intrusives envoient un signal de test dans le système à évaluer et ainsi perturbent le réseau réel. Elles sont mieux adaptées pour évaluer les paramètres directement liés à la perception de qualité par les utilisateurs et donnent des corrélations élevées avec les résultats de tests subjectifs,
- les mesures non intrusives utilisent les signaux échangés dans le trafic réel sans le perturber. Elles sont mieux adaptées pour évaluer la qualité de service d'un réseau, mais il est plus difficile d'obtenir avec de telles mesures des corrélations élevées avec les résultats de tests subjectifs.

Un bilan des avantages et inconvénients respectifs des deux types de mesure est fourni dans le tableau 2.1, tiré de [ETSI Guide EG 201 377-3 2003]. Mesures intrusives et non intrusives sont complémentaires et peuvent être utilisées dans des applications différentes. Il nous semble donc intéressant de proposer un modèle qui puisse fonctionner à la fois comme mesure intrusive et comme mesure non intrusive.

	Mesures intrusives	**Mesures non intrusives**
Mieux adaptées pour l'évaluation	Des paramètres directement liés à la perception de la qualité vocale par les utilisateurs	De la qualité de service du réseau
Corrélation avec l'évaluation subjective	Élevée	Moyenne
Utilisation de signaux de test	Obligatoire	Non obligatoire
Perturbation de la charge du réseau	Oui	Non
Interfaces	Interface au niveau de l'utilisateur final (sondes), électrique ou acoustique	Nœuds du réseau (sondes) ou dans les terminaux (agents logiciels)
Utilisateurs finaux surveillés par une sonde	Une sonde par ligne testée	Une carte INMD par unité de transmission
Mesure de l'écho possible	Oui (difficile pour des délais courts)	Oui
Mesure du délai unidirectionnel possible	Seulement si les sondes sont synchronisées (*e.g.* par GPS)	Seulement s'il y a un écho détectable

Tableau 2.1 : *Comparaison des avantages et inconvénients des mesures intrusives et non intrusives*

Pour résumer, le modèle objectif proposé dans le cadre de cette thèse devra atteindre les objectifs suivants :

- être non paramétrique, *i.e.* basé sur l'analyse des signaux échangés pendant la communication testée,
- fonctionner pour une connexion électrique/électrique et bande étroite,
- évaluer la qualité de conversation appel par appel,
- utiliser des mesures intrusives ou des mesures non intrusives, selon l'application visée.

Afin de construire un modèle répondant à ces différents objectifs, il est indispensable de disposer de données subjectives. Comme il a été vu dans le chapitre 1, la majorité des tests

subjectifs disponibles dans la littérature concerne le contexte d'écoute, les tests subjectifs dans le contexte de conversation sont peu nombreux. Un autre objectif de cette thèse consistera donc à concevoir et effectuer plusieurs tests subjectifs afin étudier l'impact des dégradations rencontrées dans le contexte de conversation sur la qualité vocale perçue.

2.3 Méthode proposée

Pour atteindre ces objectifs, il est tout d'abord important de comprendre ce qu'est la qualité de conversation, et donc plus largement ce qu'est la conversation. Il a été vu dans le chapitre 1 que la conversation pouvait être décrite, du point de vue d'un interlocuteur, comme une alternance des rôles d'auditeur et de locuteur introduisant de l'interaction entre les interlocuteurs, ces rôles n'étant pas exclusifs. Au niveau de la qualité vocale, le contexte de conversation est par conséquent affecté par les dégradations rencontrées dans le contexte d'écoute et celles rencontrées dans le contexte de locution, auxquelles s'ajoutent les dégradations affectant l'interaction de la conversation (délai et dégradation due aux périodes de double parole). La qualité de conversation pourrait ainsi être décomposée selon trois dimensions : la qualité d'écoute, la qualité de locution et la qualité d'interaction. Cette idée, qui était une hypothèse au début de l'étude et qui demandait d'être confirmée par des tests subjectifs, est à la base de la méthode proposée dans la thèse. Le modèle est divisé en deux parties, nommées ici « partie intégration » et « partie mesure » pour reprendre la terminologie proposée par la Question 20 de l'UIT-T [Pomy 2006]. Dans notre approche :

- la partie intégration combine les notes de qualité d'écoute, de locution et d'interaction pour estimer une note de qualité de conversation,
- la partie mesure fournit les notes objectives de qualité à la partie intégration en se basant sur les modèles existants de qualité vocale dans les différents contextes.

Différencier ces deux parties permet d'obtenir un modèle fonctionnant pour plusieurs applications en fonction des modèles utilisés dans la partie mesure, la partie intégration restant commune à toutes les applications.

2.3.1 Partie intégration

Pendant la construction du modèle, la combinaison des notes de qualité d'écoute, de locution et d'interaction utilisée dans la partie intégration est déterminée grâce aux notes subjectives de qualité récoltées lors de tests subjectifs effectués dans différentes conditions de dégradation. *A priori*, cette combinaison de trois composantes (qualité d'écoute, qualité de locution et qualité d'interaction) va dépendre du type de dégradation présent dans la communication testée, la qualité en contexte de conversation pouvant être plus ou moins influencée par chacune des trois composantes, en fonction des dégradations considérées. Ainsi, quand seule une dégradation de la qualité d'écoute est présente, la note de qualité de conversation sera essentiellement corrélée avec la note de qualité d'écoute, et ne dépendra pas (ou peu) de la note de qualité de locution et de la note de qualité d'interaction. L'approche proposée tient compte de cette influence du type de dégradation sur la combinaison des trois composantes en introduisant un système de décision, qui pondère l'influence des trois composantes sur la note de qualité de conversation.

Des tests subjectifs sont nécessaires pour déterminer cette relation, en fonction des dégradations considérées. Il a été vu dans le chapitre 1 qu'il existait des méthodes normalisées d'évaluation subjective de la qualité vocale dans les contextes d'écoute, de locution et de conversation. Cependant, il n'existe pas de méthode normalisée pour évaluer la qualité vocale d'interaction. La qualité vocale d'interaction est essentiellement dégradée par le délai présent dans la communication, qui va, comme l'illustre la figure 1.3 du chapitre 1, augmenter les périodes de silence et de double parole de la conversation. Nous choisissons donc de considérer, dans notre modèle, la valeur du délai comme un indicateur de la qualité vocale d'interaction.

La note de qualité de conversation est ainsi estimée par une combinaison des notes de qualité d'écoute et de locution et de la valeur du délai présent dans la communication testée.

L'approche proposée comporte les étapes suivantes :

1. Construction du modèle : les notes subjectives de qualité de locution, de qualité d'écoute et de qualité de conversation sont obtenues lors de tests subjectifs dans différentes conditions de dégradation. À partir de ces notes subjectives et de la valeur du délai, une relation F est déterminée entre les trois composantes (note subjective de qualité de locution, note subjective de qualité d'écoute et délai) afin d'estimer la note subjective de qualité de conversation. L'influence du type de dégradation sur la combinaison des trois composantes est prise en compte en considérant une relation F_i pour chaque dégradation i :

$$\widehat{MOS}_{conversation} = F_i(MOS_{locution}, MOS_{écoute}, délai) \tag{2.1}$$

   La construction du modèle et la détermination de l'ensemble des relations F_i à partir de résultats de tests subjectifs est décrite en détail dans le chapitre 3.

2. Application du modèle à une communication testée :
   (a) détermination, par la partie mesure présentée dans le paragraphe 2.3.2, des notes objectives d'écoute et de locution, en appliquant les modèles objectifs correspondants soit aux signaux de la communication testée, soit aux paramètres physiques du système testé,
   (b) détection des dégradations observées pour la communication testée, en particulier du délai présent dans le système testé,
   (c) calcul de la note de qualité de conversation estimée à partir des notes objectives d'écoute et de locution et de la mesure du délai grâce à la combinaison choisie dans la base de donnée des relations F_i en fonction des dégradations détectées à l'étape 2b.

2.3.2 Partie mesure

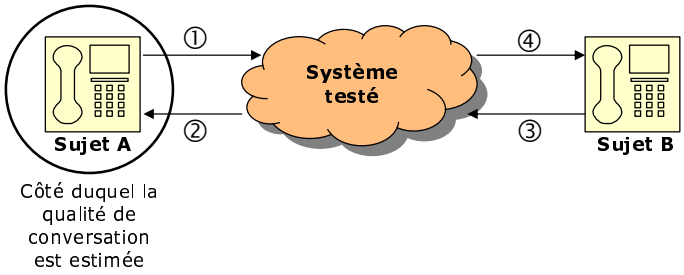
Le chapitre 1 a fait le bilan des modèles objectifs existants dans les contextes d'écoute et de locution (*cf.* figure 1.15). Dans le contexte d'écoute, les modèles normalisés suivants sont disponibles :

- P.862 (PESQ) [UIT-T Rec. P.862 2001], pour une mesure intrusive basée sur les signaux,
- P.563 [UIT-T Rec. P.563 2004], pour une mesure non intrusive basée sur les signaux,
- P.564 [UIT-T Rec. P.564 2006], pour une mesure non intrusive paramétrique.

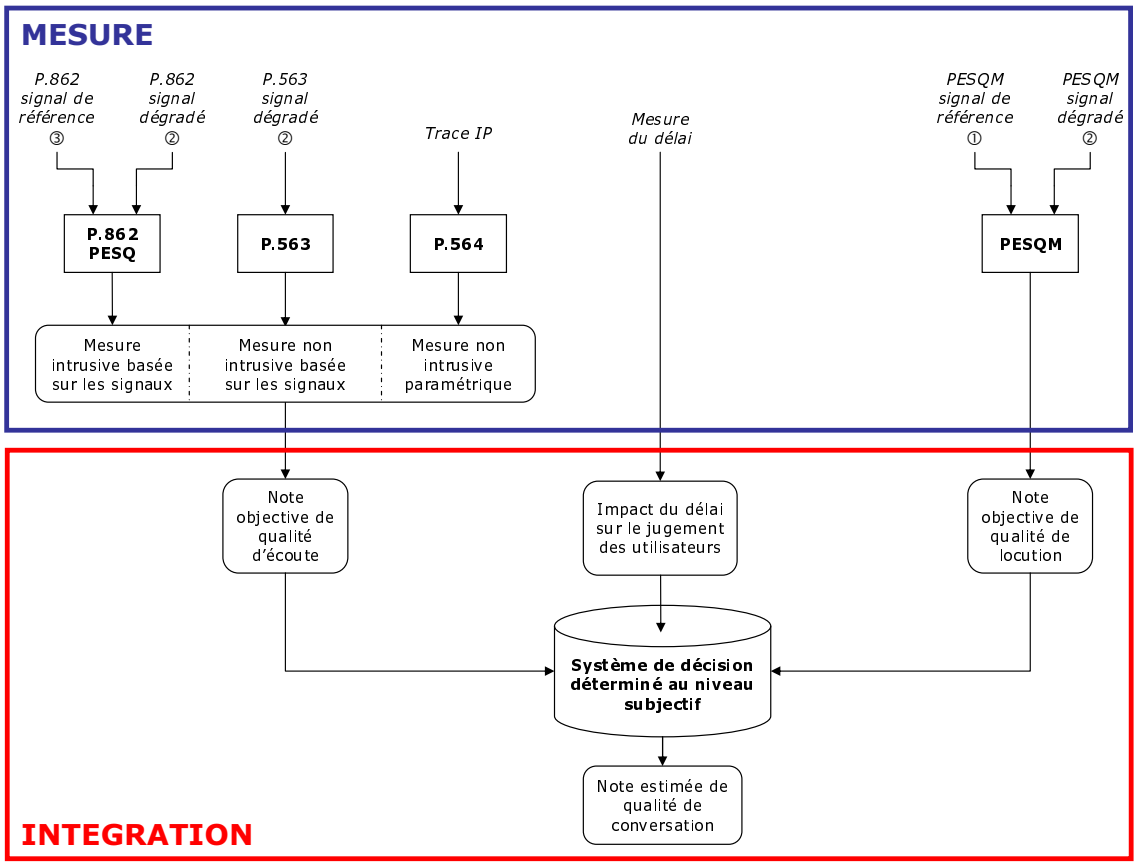
Dans le contexte de locution, il n'existe aucun modèle normalisé par l'UIT-T, seul le modèle PESQM [Appel et Beerends 2002], intrusif et basé sur les signaux, est disponible. Le délai peut être mesuré de différentes façons selon l'application visée :

- en mesure intrusive basée sur les signaux : grâce au modèle PESQ, qui permet la mesure du retard entre ses signaux de référence et dégradé, si ceux-ci sont synchrones.
- en mesure intrusive ou non intrusive basée sur les signaux : grâce à la mesure du délai de l'écho, s'il est présent.
- en mesure non intrusive pour les communications IP : grâce aux informations contenues dans la trame RTCP ou RTCP-XR sur les délais des paquets transmis.
- en mesure non intrusive : grâce à deux sondes synchronisées (*e.g.* par GPS) placées chacune d'un côté de la communication.

Le modèle de conversation va pouvoir utiliser ces différents modèles dans la partie mesure pour fournir à la partie intégration les notes objectives de qualité d'écoute et de locution et



(a) Communication testée



(b) Modèle objectif

Figure 2.1 : *Méthode proposée pour l'évaluation objective de la qualité conversationnelle*

la valeur du délai. Le schéma de la méthode proposée est donné dans la figure 2.1(b) pour l'application à la communication présentée dans la figure 2.1(a). Les outils nécessaires à la partie mesure du modèle objectif seront présentés dans le chapitre 4 et les résultats obtenus dans différentes conditions de dégradation dans le chapitre 5.

Conclusion

Dans ce chapitre, nous avons décrit la problématique posée par l'élaboration d'un modèle objectif non paramétrique de la qualité vocale en contexte de conversation. Nous avons ensuite fixé les objectifs à atteindre par un tel modèle, en se basant sur la Question 20 de l'UIT-T. Nous avons proposé un modèle répondant à ces objectifs, à savoir un modèle non paramétrique (*i.e.* basé sur l'analyse des signaux échangés pendant la communication testée), fonctionnant pour une connexion électrique/électrique et bande étroite, évaluant la qualité de conversation appel par appel, et utilisant des mesures intrusives ou non intrusives, selon l'application

visée. Le modèle proposé présente l'avantage d'être indépendant des modèles utilisés dans la partie mesure, il peut ainsi évoluer en même temps que les modèles d'écoute et de locution s'améliorent. Ceci vaut en particulier dans le contexte de locution, dans lequel le modèle PESQM est pour l'instant le seul modèle (non normalisé) disponible, fonctionnant uniquement en mesure intrusive.

Dans le chapitre suivant, la construction du modèle à l'aide de nouveaux tests subjectifs est présentée.

Chapitre 3

Construction du modèle

Introduction

La méthode proposée dans le chapitre 2 consiste à essayer de combiner les qualités d'écoute et de locution, et le délai pour estimer la qualité de conversation. Cette combinaison de trois composantes est déterminée pendant la construction du modèle grâce aux notes subjectives de qualité. Pour tester une communication, la combinaison à utiliser est choisie par la partie intégration du modèle en fonction des dégradations présentes et est appliquée aux notes objectives fournies par les modèles de la partie mesure.

Dans ce chapitre, nous déterminons la combinaison des trois composantes à partir des résultats de plusieurs tests subjectifs. Cette étude nécessite de disposer des notes subjectives de qualité vocale dans les trois contextes (écoute, locution, conversation), obtenues dans les mêmes conditions de dégradation et de test. Or, dans la littérature, il n'existe pas de tels résultats de tests subjectifs : de nouveaux tests subjectifs ont donc été nécessaires. Ils ont pour but de vérifier la faisabilité de cette approche et de construire le modèle dans différentes conditions de dégradation. Ce chapitre décrit tout d'abord la méthodologie de test proposée pour réaliser les tests subjectifs nécessaires à l'étude, puis les quatre tests subjectifs mis en œuvre dans le cadre de la thèse (conditions testées et montage de test, résultats, analyse statistique des résultats, analyse des notes individuelles).

3.1 Méthodologie de test proposée

3.1.1 Protocole

Ce protocole de test subjectif doit permettre de déterminer les qualités d'écoute, de locution et de conversation d'une liaison vocale. Il s'agit d'un test de conversation, impliquant des participants non experts (A et B) placés dans deux pièces insonorisées séparées et communiquant grâce au système testé. Le protocole de test est divisé en trois phases successives :

1. Locution du participant A, écoute du participant B : détermination de la qualité de locution du côté A et de la qualité d'écoute du côté B.
2. Locution du participant B, écoute du participant A : détermination de la qualité de locution du côté B et de la qualité d'écoute du côté A.
3. Conversation libre : détermination de la qualité de conversation globale.

Ce protocole permet ainsi d'obtenir des notes pour les qualités d'écoute, de locution et de conversation des deux côtés du lien téléphonique. Dans les deux premières phases, le contenu de la communication est contrôlé afin de maîtriser les temps de parole des participants, ainsi que le contenu sémantique de la communication. Un texte ou une série de phrases à lire, différant pour chaque condition testée, est fourni à chacun des participants. Pendant la troisième phase, les participants ont une conversation libre en se basant sur les scénarios de conversation

proposés dans [Möller 1997a]. Ces scénarios consistent en des jeux de rôles sur un thème (*e.g.* agence de voyage, pizza, etc.) et ont été conçus pour aboutir à des conversations structurées, naturelles, durant environ 2-3 minutes. Un exemple de scénario est donné dans l'annexe E.

À la fin de chaque phase, les deux participants jugent la qualité, selon des critères choisis auparavant en fonction des conditions testées. Dans les tests subjectifs présentés ici, les critères de qualité explorés sont la gêne due à l'écho, la voix, le niveau vocal, la possibilité d'interruption, la qualité pendant les périodes de double parole, la gêne due au bruit, la gêne due aux défauts et la qualité globale, en fonction des dégradations testées. Les questions posées lors des tests de la thèse sont fournies dans l'annexe E. Cette évaluation se fait selon deux méthodes, la méthode ACR normalisée (pour les jugements de qualité) et la méthode DCR sans référence (pour les jugements de gêne).

3.1.2 Déroulement des tests

Avant le début du test, les instructions sont données aux participants pour leur expliquer ce qui est attendu d'eux pendant le test.

Plusieurs conditions d'apprentissage sont effectuées au début de chaque session de test pour que les participants :

- se familiarisent avec l'enchaînement des différentes phases du test et avec les scénarios de conversation,
- maîtrisent le processus de notation,
- prennent connaissance de l'étendue des niveaux de dégradation qu'ils vont rencontrer ultérieurement dans le test, en leur présentant la condition de référence (sans dégradation) et plusieurs conditions avec dégradations.

La notation se fait à la fin de chaque phase. Un logiciel de notation a été développé dans le cadre de la thèse pour accélérer et automatiser la récolte et l'analyse des notes attribuées par les participants. Il présente à chacun des participants les textes/phrases à lire, les scénarios de conversation tels que ceux présentés dans l'annexe E, les instructions et les questions au fur et à mesure du test. Le logiciel gère l'ordre de présentation des conditions et des textes/phrases/scénarios de conversation, grâce à un carré gréco-latin tel que celui présenté dans le tableau 1.5 du chapitre 1.

3.1.3 Choix des conditions de test

La contrainte principale d'un test de conversation est sa durée, qui en règle générale n'excède pas 2 heures 30 minutes (incluant une pause de 15 minutes au milieu et une phase d'apprentissage de 15 minutes environ au début du test), afin d'éviter une fatigue des participants. Cette contrainte de temps agit directement sur le nombre de conditions testées et donc sur le choix de ces conditions. L'autre contrainte forte est de retenir des conditions réalistes et à des niveaux rencontrés dans les réseaux réels. Ces deux contraintes doivent constamment être gardées en mémoire dans la conception de tests de conversation.

3.1.4 Montage expérimental et enregistrement

Pour chaque test, une fois les conditions de dégradation choisies, le montage expérimental est mis en œuvre pour permettre la communication en duplex intégral et pour dégrader la qualité vocale en temps réel. Ce montage est conçu de telle sorte que la communication semble la plus naturelle possible aux participants du test, avec une numérotation et une sonnerie au début de la communication.

Le modèle objectif de qualité conversationnelle tel qu'il a été proposé dans le chapitre 2 utilise en entrée des modèles objectifs de qualités d'écoute et de locution basés sur les signaux, tels que PESQ et PESQM. Pour une communication donnée, le modèle a donc besoin des quatre signaux de parole, émission et réception de chaque côté du système (*cf.*

figure 2.1(a)). Les communications effectuées lors des tests subjectifs seront ainsi enregistrées afin d'être utilisées pour l'évaluation des performances du modèle proposé au niveau objectif, abordée dans le chapitre 5. Cet enregistrement doit être effectué en temps réel au cours de chaque communication, sans en perturber le bon déroulement, et être le plus proche possible de ce que les participants ont entendu. Idéalement les quatre enregistrements doivent être synchronisés.

3.1.5 Analyse des résultats subjectifs et rejet des sujets aberrants

Pour chaque test, une analyse statistique des résultats est effectuée en fonction des différents critères étudiés et des questions posées aux participants durant le test. Seules les notes relatives au critère de qualité globale seront utilisées pour la construction du modèle objectif, puisqu'il doit estimer la qualité de la communication sans se focaliser sur une dégradation en particulier. L'étude des autres critères permet d'analyser la façon dont les sujets ont jugé la qualité et de vérifier la cohérence de ce jugement vis-à-vis des dégradations testées. Les outils statistiques nécessaires à l'analyse sont présentés dans l'annexe D. Un seuil de significativité égal à 0.05 sera utilisé systématiquement et un astérisque signalera un niveau significatif ($p < 0.05$).

Avant d'analyser les notes moyennes calculées sur l'ensemble des sujets, il est important d'étudier les notes individuelles attribuées par chacun des participants afin d'écarter les sujets dont les notes sont aberrantes. Le choix d'un critère de rejet est difficile, dans la mesure où l'ensemble des sujets est très hétérogène comme l'est l'ensemble des utilisateurs des services téléphoniques. Ici, le critère retenu consiste à rejeter les sujets qui ont présenté des difficultés pendant le test (compréhension du déroulement du test, utilisation du logiciel de notation, etc.), ainsi que les sujets dont les notes ne varient pas ou peu quand les dégradations augmentent.

3.2 Tests subjectifs réalisés

Quatre tests subjectifs ont été mis en œuvre conformément à la méthodologie décrite dans la section 3.1. Le premier test étudiait le délai et l'écho, le deuxième les pertes de paquets et le bruit transmis, le troisième le bruit, et le quatrième le délai, l'écho et les pertes de paquets.

3.2.1 Test 1 : délai et écho

3.2.1.1 Objectifs

L'objectif premier de ce test est de valider la méthodologie de test, et en particulier de mieux connaître le temps nécessaire à chaque condition, ainsi que la réaction des participants vis-à-vis du déroulement du test et de l'enchaînement des phases. Le second objectif est d'étudier l'effet du délai et de l'écho sur la qualité de conversation et sur la relation de celle-ci avec les qualités d'écoute, de locution et d'interaction. Ces deux dégradations sont étudiées car :

- elles peuvent avoir séparément et de façon combinée un impact sur la qualité de conversation et certaines des trois composantes : le délai seul a un impact sur la qualité d'interaction et de conversation, l'écho combiné au délai a un impact sur la qualité de locution et sur la qualité de conversation,
- comme cela a été souligné dans le chapitre 1, le délai a un statut particulier parmi les dégradations rencontrées en contexte de conversation : il semble nécessaire d'en analyser l'impact.

3.2.1.2 Conditions et facteurs expérimentaux

Ce test étant le premier, il était difficile d'estimer *a priori* la durée nécessaire à chaque condition, celle-ci pouvant être modulée par la réaction des participants vis-à-vis de cette nouvelle méthodologie. Huit conditions ont été retenues, afin de s'assurer du déroulement du test en moins de 2 heures, celles-ci sont fournies dans le tableau 3.1. Les niveaux de délai ont été choisis pour balayer une gamme réaliste de valeurs, en-dessous et au-dessus du seuil de 400 ms, considéré comme critique dans la Recommandation G.114 [UIT-T Rec. G.114 2003].

Condition	Délai par sens entre A et B (ms)	Niveau d'atténuation de l'écho du côté A (dB)
1	0	Pas d'écho électrique
2	0	25
3	200	Pas d'écho électrique
4	200	25
5	400	Pas d'écho électrique
6	400	25
7	600	Pas d'écho électrique
8	600	25

Tableau 3.1 : *Conditions - Test 1 sur le délai et l'écho*

Seize paires de sujets non experts (20 femmes et 12 hommes), âgés de 20 à 60 ans, ont participé au test. *A priori* otologiquement sains, ils ont tous été recrutés en dehors de France Télécom et rémunérés. La mise en œuvre du test a été faite à l'aide d'un carré gréco-latin. Un test asymétrique a été choisi pour étudier l'effet du délai combiné à l'écho du côté A et l'effet du délai seul du côté B. Cependant, un écho acoustique non prévu et non contrôlé, atténué d'environ 50 dB, a gêné les sujets B. Leurs données n'ont donc volontairement pas été exploitées et ne seront pas présentées ici.

Les sujets s'entraînaient au début du test avec deux conditions d'apprentissage, les aidant à comprendre le déroulement du test et à ancrer leur jugement. Il s'agissait des conditions 1 (sans dégradation) et 8 (délai à 600 ms et écho atténué à 25 dB).

Le montage du test, présenté dans la figure 3.1, relie deux combinés RNIS. L'autocommutateur (PABX) gère les indicatifs correspondant chacun à une condition de test. Le délai dans chaque sens est généré à l'aide d'une ligne à retard de 200 ms, dans laquelle le signal passe le nombre de fois correspondant à la condition testée. L'écho est créé en atténuant le signal transitant de A vers B et en le réinjectant dans le sens B vers A. Deux magnétophones numériques DAT (*Digital Audio Tape*) enregistrent chacun un côté de la communication. Les deux DAT ne sont pas synchronisés. Ainsi, pour ce test, quatre signaux sont disponibles par phase, par condition et par groupe, soit un total de 1536 signaux (format wav, mono, 48 kHz). Les facteurs expérimentaux du test sont fournis dans le tableau 3.2.

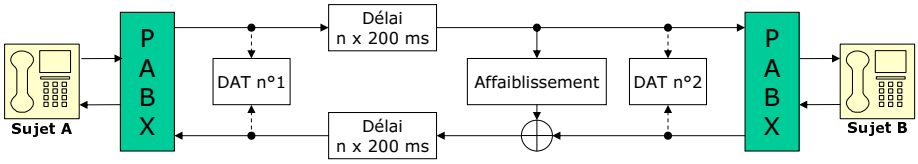


Figure 3.1 : *Schéma du dispositif - Test 1 sur le délai et l'écho*

Les questions posées aux sujets sont les suivantes (*cf.* annexe E) pour chaque contexte :

- locution : gêne due à l'écho, qualité globale.
- écoute : voix de l'interlocuteur, niveau vocal, qualité globale.
- conversation : voix de l'interlocuteur, niveau vocal, gêne due à l'écho, possibilité d'interruption, qualité en période de double parole, qualité globale.

Facteurs	
Terminaux	combinés téléphoniques RNIS
Codec	G.711
Enregistrement	DAT, 48 kHz, 16 bits
Nombre de paires de sujets	16 paires de sujets non experts
Mise en œuvre du test	carré gréco-latin
Méthodes de notation	ACR et DCR
Langue	français

Tableau 3.2 : *Facteurs expérimentaux - Test 1 sur le délai et l'écho*

3.2.1.3 Analyse des résultats

Les données d'un sujet ont été écartées, car elles étaient constantes quels que soient la condition et le contexte. Les résultats présentés dans la suite correspondent donc à la moyenne des données de quinze sujets.

Les effets de trois facteurs ont été analysés : le contexte dans lequel se trouve le sujet, l'écho et le délai. Le facteur Contexte peut prendre trois niveaux (locution, écoute, conversation). Le facteur Délai présente quatre niveaux (0, 200, 400, 600 ms). Le facteur Écho a deux niveaux (pas d'écho électrique, écho électrique). Les questions posées aux sujets portent sur six critères de qualité (qualité globale, gêne due à l'écho, voix, niveau sonore, possibilité d'interruption, qualité en double parole). L'analyse menée ici consiste à étudier l'effet des trois facteurs (Contexte, Délai, Écho) sur le jugement de qualité des sujets, pour chacun des six critères.

Corrélations Nous avons tout d'abord étudié la corrélation qui existait entre ces différents critères pour chacun des trois contextes (locution, écoute, conversation).

En contexte de locution les critères évalués sont la qualité globale et la gêne due à l'écho. La matrice de corrélation obtenue entre ces deux critères est donnée dans le tableau 3.3. La qualité globale est corrélée avec l'écho : le jugement des sujets sur la qualité globale dépend fortement de leur jugement de l'écho.

	Écho	**Qualité globale**
Écho	1	0.838
Qualité globale	0.838	1

Tableau 3.3 : *Corrélation entre critères en contexte de locution - Test 1 sur le délai et l'écho*

En contexte d'écoute, les critères évalués sont la qualité globale, le niveau sonore et la voix. Une transformation a été nécessaire afin de convertir l'échelle du niveau sonore (*cf.* tableau E.1 de l'annexe E) en une échelle correspondant aux autres échelles (de 1 à 5 par préférence croissante). La transformation est effectuée d'après l'équation suivante

$$\text{Note transformée} = -2|\text{Note originale} - 3| + 5. \tag{3.1}$$

La matrice de corrélation obtenue entre ces trois critères est donnée dans le tableau 3.4. La corrélation la plus importante est celle qui existe entre les critères de la voix et de la qualité globale. Le niveau sonore est très peu corrélé aux autres critères. Les sujets semblent avoir basé leur jugement de la qualité globale en situation d'écoute en grande partie sur la qualité de la voix de leur interlocuteur. Cependant, le seul critère de la voix n'explique pas tout le jugement de qualité globale en situation d'écoute. D'autres critères, que nous n'avons pas évalués dans ce test, ont pu intervenir dans le jugement des sujets.

	Niveau	**Voix**	**Qualité globale**
Niveau	1	0.115	0.047
Voix	0.115	1	0.631
Qualité globale	0.047	0.631	1

Tableau 3.4 : *Corrélation entre critères en contexte d'écoute - Test 1 sur le délai et l'écho*

En contexte de conversation, les critères évalués sont la qualité globale, l'écho, l'interruption, la double parole, le niveau sonore et la voix. La transformation des notes obtenues pour le niveau sonore est également effectuée selon l'équation 3.1. La matrice de corrélation entre ces six critères est donnée dans le tableau 3.5. Le jugement de la qualité globale semble avoir été influencé essentiellement par l'écho et la double parole.

	Écho	Interruption	Double parole	Niveau	Voix	Qualité globale
Écho	1	0.086	0.333	0.075	0.113	0.665
Interruption	0.086	1	0.321	0.243	0.218	0.114
Double parole	0.333	0.321	1	0.034	0.193	0.493
Niveau	0.075	0.243	0.034	1	0.096	0.081
Voix	0.113	0.218	0.193	0.096	1	0.128
Qualité globale	0.665	0.114	0.493	0.081	0.128	1

Tableau 3.5 : *Corrélation entre critères en contexte de conversation - Test 1 sur le délai et l'écho*

Il est aussi intéressant d'étudier la corrélation entre les trois contextes pour le critère de qualité globale, présentée dans le tableau 3.6. Les contextes de locution et d'écoute sont peu corrélés. Le contexte de conversation est majoritairement corrélé au contexte de locution, ce qui est logique étant données les dégradations testées.

	Conversation	Locution	Écoute
Conversation	1	0.684	0.256
Locution	0.684	1	0.126
Écoute	0.256	0.126	1

Tableau 3.6 : *Corrélation entre contextes pour le critère de qualité globale - Test 1 sur le délai et l'écho*

Qualité globale La figure 3.2 montre les notes MOS de qualité globale et les intervalles de confiance à 95% correspondants, en absence d'écho (graphe de gauche) et en présence d'écho (graphe de droite), en fonction du contexte et du délai unidirectionnel subi par les sujets.

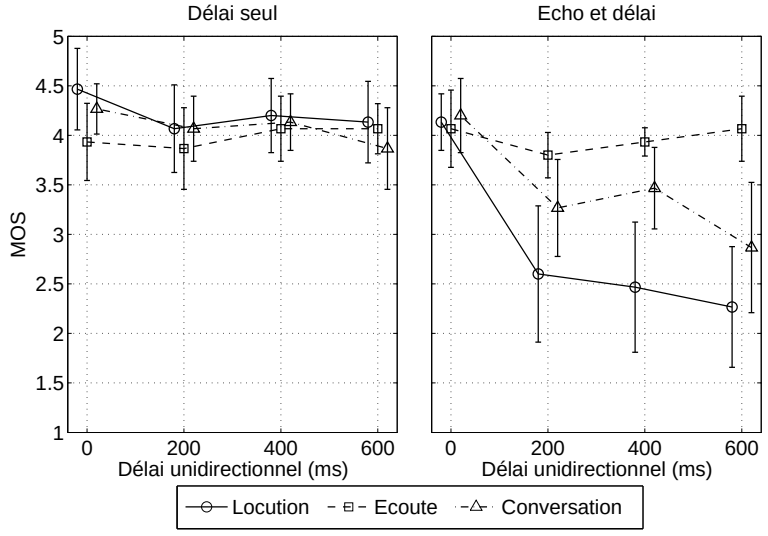


Figure 3.2 : *Notes MOS du critère de qualité globale et intervalles de confiance à 95% correspondants du test 1 sur le délai et l'écho*

Il semble qu'en absence d'écho, le jugement global moyen soit peu influencé par le contexte et par la valeur du délai. L'écho, dès lors qu'il a un délai non nul, a un effet très important sur le jugement global moyen, sauf en situation d'écoute. Dans le cas avec écho, la note moyenne de qualité globale dépend du contexte dans lequel se trouve le sujet.

Facteur	SC	dl	CM	F	P>F
Contexte	11.27	2	5.64	10.31	0.000*
Délai	22.10	3	7.37	11.36	0.000*
Écho	40.00	1	40.00	26.25	0.000*
Contexte×Délai	13.82	6	2.30	7.27	0.000*
Contexte×Écho	26.72	2	13.36	18.99	0.000*
Délai×Écho	10.38	3	3.46	11.21	0.000*
Contexte×Délai×Écho	4.51	6	0.75	2.48	0.029*

Tableau 3.7 : *ANOVA pour le critère de qualité globale et pour les trois contextes - Test 1 sur le délai et l'écho*

Une ANOVA (*ANalysis Of VAriance*, *cf.* annexe D) conduite sur les notes obtenues pour le critère de qualité globale, dont les résultats sont présentés dans le tableau 3.7, confirme les effets significatifs de chacun des trois facteurs Contexte, Délai et Écho.

L'effet de l'écho est effectivement très significatif ($p < 0.05$). De plus, cet effet dépend du contexte dans lequel se trouve le sujet, comme l'indique l'interaction significative entre les deux facteurs (Écho et Contexte). Dans le contexte d'écoute, la note moyenne globale reste stable entre le cas sans écho et le cas avec écho, ce qui s'explique aisément puisqu'en contexte d'écoute l'auditeur n'expérimente jamais son écho. En revanche, dans le contexte de locution, la note moyenne globale chute, en moyenne sur les délais non nuls (200, 400 et 600 ms), de 1.68 MOS entre le cas sans écho et le cas avec écho. Dans le contexte de conversation, la note moyenne globale chute, en moyenne sur les délais non nuls, de 0.82 MOS entre le cas sans écho et le cas avec écho. Ainsi, en mettant à part le contexte d'écoute qui n'est pas affecté par l'écho, plus le contexte est interactif, moins l'effet de l'écho sur le jugement global moyen est important. Ceci peut s'expliquer par le fait qu'en contexte de locution, les sujets sont plus attentifs à la qualité et à son jugement, alors qu'en contexte de conversation (donc d'interaction) les ressources attentionnelles nécessaires à la conduite de la discussion sont autant d'attention en moins consacrée au jugement de la qualité.

L'ANOVA révèle un effet très significatif du délai. Cet effet du délai dépend de la présence ou non d'écho (interaction significative entre les deux facteurs). En effet, en l'absence d'écho, l'augmentation du délai n'entraîne pas de chute significative de la note moyenne de qualité globale, quel que soit le contexte considéré. En présence d'écho, l'effet du délai dépend du contexte dans lequel se trouve le sujet (interaction significative entre les deux facteurs), ceci étant essentiellement dû au contexte d'écoute, qui n'est pas affecté par l'écho et pour lequel la note moyenne globale reste relativement stable (autour de 4 MOS) lorsque le délai augmente. En contexte de conversation et sans écho, une légère inflexion de la qualité moyenne entre 400 et 600 ms est observée.

Gêne due à l'écho La figure 3.3 représente les notes moyennes de gêne due à l'écho et les intervalles de confiance à 95% correspondants, en absence d'écho (graphe de gauche) et en présence d'écho (graphe de droite), en fonction du contexte et du délai unidirectionnel subi par les sujets.

En absence d'écho, le jugement moyen de la gêne due à l'écho semble constant quels que soient le contexte et la valeur du délai. En toute logique, l'écho, dès lors qu'il a un délai non nul, a un effet très important sur le jugement moyen de la gêne due à l'écho. Dans le cas avec écho, la note moyenne dépend du contexte dans lequel se trouve le sujet, ainsi que de la valeur du délai.

Une ANOVA réalisée sur les notes obtenues pour la gêne due à l'écho pour les deux contextes (locution, conversation), dont les résultats sont présentés dans le tableau 3.8, confirme les effets significatifs des trois facteurs Contexte, Délai et Écho.

L'effet de l'écho est très significatif et dépend du délai (interaction significative entre les deux facteurs), puisque, comme pour le critère de qualité globale, un écho avec un délai nul n'est pas perçu et est assimilé à une absence d'écho. Comme pour le critère de qualité globale,

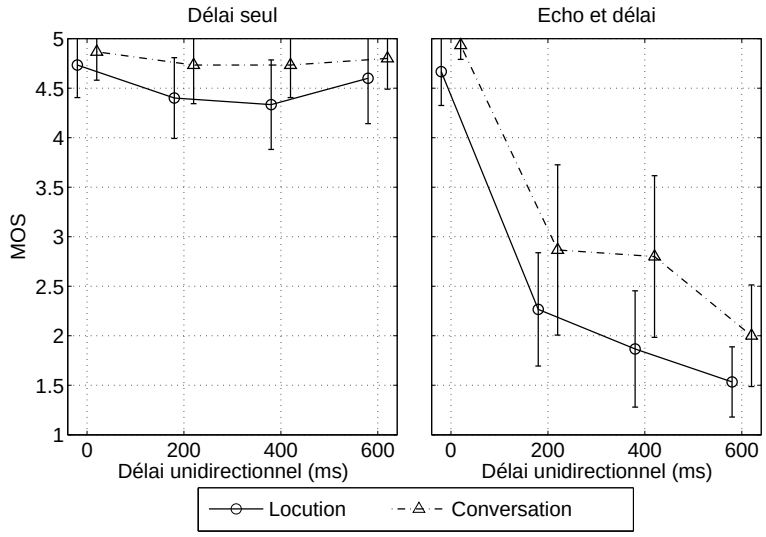


Figure 3.3 : *Notes MOS du critère de gêne due à l'écho et intervalles de confiance à 95% correspondants - Test 1 sur le délai et l'écho*

Facteur	SC	dl	CM	F	P>F
Contexte	10.4	1	10.4	27.3	0.000*
Délai	90.2	3	30.1	58.5	0.000*
Écho	190.8	1	190.8	129.2	0.000*
Contexte×Délai	1.8	3	0.6	1.4	0.271
Contexte×Écho	1.3	1	1.3	2.3	0.150
Délai×Écho	70.8	3	23.6	37.2	0.000*
Contexte×Délai×Écho	0.3	3	0.1	0.2	0.865

Tableau 3.8 : *ANOVA pour le critère de gêne due à l'écho pour les deux contextes (Locution, Conversation) - Test 1 sur le délai et l'écho*

Facteur	SC	dl	CM	F	P>F
Délai	2.73	3	0.91	1.843	0.154
Écho	0.3	1	0.3	0.808	0.384
Délai×Écho	0.7	2	0.233	0.620	0.606

Tableau 3.9 : *ANOVA pour le critère de possibilité d'interruption pour le contexte de conversation - Test 1 sur le délai et l'écho*

moins la situation est interactive, plus l'écho a d'effet sur le jugement de qualité. En absence d'écho, le délai n'a pas d'effet significatif sur le jugement, quel que soit le contexte considéré, puisque la question porte sur la gêne due à l'écho. En présence d'écho, l'essentiel de l'effet du délai est dû à la chute observée entre l'écho avec un délai nul (assimilé à une absence d'écho) et l'écho avec un délai non nul (200, 400 et 600 ms). L'effet du contexte est aussi significatif : l'écho gêne plus les sujets en contexte de locution qu'en contexte de conversation, confirmant ce qui avait déjà été constaté pour le critère de la qualité globale.

Interruption La figure 3.4 montre les notes moyennes de possibilité d'interruption et les intervalles de confiance à 95% correspondants, en absence d'écho (graphe de gauche) et en présence d'écho (graphe de droite), en fonction du délai unidirectionnel subi par les sujets, dans le contexte de conversation.

Le jugement moyen de la possibilité d'interruption est peu influencé par la valeur du délai. Une ANOVA est réalisée sur les notes obtenues pour le critère d'interruption, en considérant les facteurs Délai et Écho. Le tableau 3.9 présente les résultats de l'ANOVA.

L'ANOVA ne révèle aucun effet significatif du délai ni de l'écho. Ainsi, les sujets n'ont pas eu de difficulté particulière à interrompre leur interlocuteur, quels que soient le délai

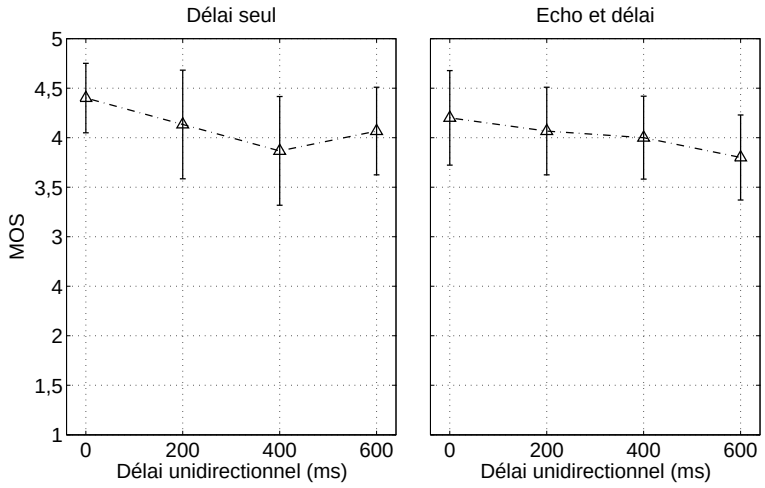


Figure 3.4 : *Notes MOS du critère de possibilité d'interruption et intervalles de confiance à 95% correspondants - Test 1 sur le délai et l'écho*

et l'atténuation de l'écho. Une fois de plus, le délai n'a pas perturbé les sujets, bien que la question porte en particulier sur l'interactivité de la communication, qui est essentiellement dégradée par le délai.

Double parole La figure 3.5 représente les notes moyennes pour le critère de double parole et les intervalles de confiance à 95% correspondants, en absence d'écho (graphe de gauche) et en présence d'écho (graphe de droite), en fonction du délai unidirectionnel subi par les sujets, pour le contexte de conversation.

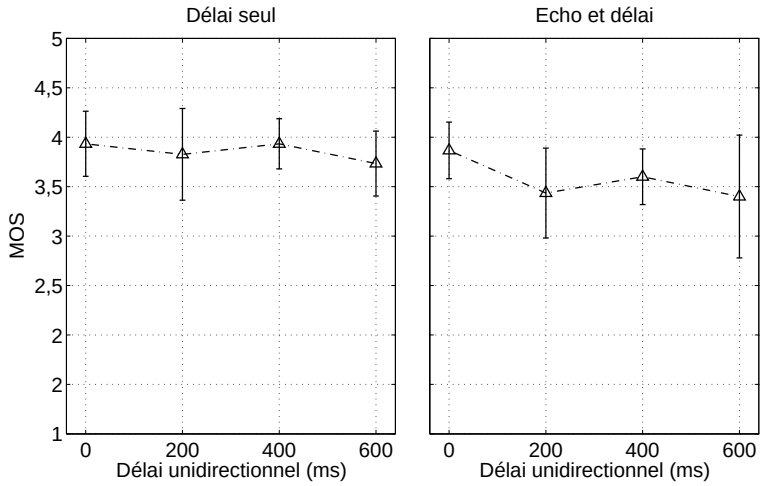


Figure 3.5 : *Notes MOS du critère de double parole et intervalles de confiance à 95% correspondants - Test 1 sur le délai et l'écho*

En absence et en présence d'écho le jugement moyen de la double parole semble peu influencé par la valeur du délai. Une ANOVA est conduite sur les notes de double parole obtenues, en considérant les facteurs Délai et Écho. Le tableau 3.10 présente les résultats de l'ANOVA. L'effet de l'écho sur la note moyenne de double parole est significatif. L'ANOVA révèle également que le délai n'a pas d'effet significatif.

3.2.1.4 Synthèse

Le premier objectif du test est atteint : nous avons pu grâce à ce test vérifier que la méthodologie de test était faisable et déterminer que chaque condition durait environ cinq

Facteur	SC	dl	CM	F	P>F
Délai	1.98	3	0.66	1.686	0.184
Écho	2.37	1	2.37	5.778	0.031*
Délai×Écho	0.47	3	0.16	0.733	0.538

Tableau 3.10 : *ANOVA pour le critère de double parole pour le contexte de conversation - Test 1 sur le délai et l'écho*

minutes. Les sujets n'ont pas eu de difficulté particulière pour comprendre le déroulement du test et la notation. La même méthodologie a donc été conservée pour étudier d'autres dégradations dans les tests suivants. La donnée sur la durée moyenne de chaque condition est importante puisqu'elle permet de définir le nombre maximal de conditions tolérables pour les sujets.

Ensuite, les effets du délai et de l'écho ont été étudiés. Le principal résultat est que le délai a un impact faible pour des valeurs inférieures à 400 ms, puis un effet plus important entre 400 et 600 ms. Ce résultat semble confirmer ce qui a été reporté dans la section 1.2.3.2 du chapitre 1 sur l'effet du délai. Cependant il est important de noter qu'ici le délai a été testé en présence d'écho fort ce qui a pu « écraser » l'effet du délai.

3.2.2 Test 2 : pertes de paquets et bruit

3.2.2.1 Objectif

L'objectif de ce test est d'étudier l'effet des pertes de paquets et du bruit (diffusé dans les salles avec des haut-parleurs) sur la qualité de conversation et sur la relation de celle-ci avec les qualités d'écoute, de locution et d'interaction. Ces deux dégradations sont étudiées car :

- les pertes de paquets ont un effet sur la qualité d'écoute et sur la qualité de conversation,
- le bruit a un effet sur la qualité d'écoute, la qualité de locution et la qualité de conversation.

3.2.2.2 Conditions et facteurs expérimentaux

La durée moyenne par condition déterminée lors du premier test (5 minutes/condition) a permis d'augmenter considérablement le nombre de conditions testées par rapport au premier test. Dix-neuf conditions ont été choisies, sous la forme de quinze conditions avec le bruit diffusé dans une seule des deux salles plus quatre conditions avec le bruit diffusé dans les deux salles simultanément. Les conditions sont fournies dans le tableau 3.11. Le bruit diffusé est de type Hoth [Hoth 1941], qui est un bruit gaussien filtré passe-bas avec un spectre fréquentiel similaire à la voix. Il est utilisé de manière habituelle pour simuler le bruit de fond présent dans un bureau. Les niveaux de bruit ont été déterminés pour être gênant (49 dB(A)) et très gênant (59 dB(A)), sans être insupportables. Les pertes de paquets introduites sont de type aléatoire. Les taux de 5 et 10% ont été choisis car ils correspondent à des valeurs réalistes.

Dix paires de sujets non experts (10 femmes et 10 hommes), âgés de 18 à 55 ans, ont participé au test. *A priori* otologiquement sains, ils ont tous été recrutés en dehors de France Télécom et rémunérés. La mise en œuvre du test a été faite à l'aide d'un carré gréco-latin.

Les sujets s'entraînaient au début du test avec deux conditions d'apprentissage : les conditions 3 (taux de pertes de paquets = 5% et pas de bruit) et 4 (bruit du côté A, pas de bruit du côté B et pas de perte de paquets).

Le montage du test, présenté dans la figure 3.6, relie deux ordinateurs portables auxquels sont connectés deux micro-casques monauraux. La communication est effectuée entre les deux ordinateurs grâce au logiciel Microsoft NetMeeting. Les pertes de paquets sont introduites dans la liaison avec le logiciel NetDisturb, qui simule les pertes de paquets au taux souhaité. Le bruit de fond est diffusé dans chacune des salles à l'aide de haut-parleurs. Les communications sont enregistrées grâce à deux autres ordinateurs portables équipés du logiciel Ethereal, qui

Condition	Niveau du bruit diffusé côté A (dB(A))	Niveau du bruit diffusé côté B (dB(A))	Taux de pertes de paquets côtés A et B (%)
1	Pas de bruit	Pas de bruit	0
2	Pas de bruit	Pas de bruit	5
3	Pas de bruit	Pas de bruit	10
4	49	Pas de bruit	0
5	49	Pas de bruit	5
6	49	Pas de bruit	10
7	59	Pas de bruit	0
8	59	Pas de bruit	5
9	59	Pas de bruit	10
10	Pas de bruit	49	0
11	Pas de bruit	49	5
12	Pas de bruit	49	10
13	Pas de bruit	59	0
14	Pas de bruit	59	5
15	Pas de bruit	59	10
16	49	49	0
17	59	59	0
18	49	59	0
19	59	49	0

Tableau 3.11 : *Conditions - Test 2 sur les pertes de paquets et le bruit*

permet de capturer le flux de données IP entre deux adresses IP. Les enregistrements sont synchronisés. Les facteurs expérimentaux du test sont fournis dans le tableau 3.12.

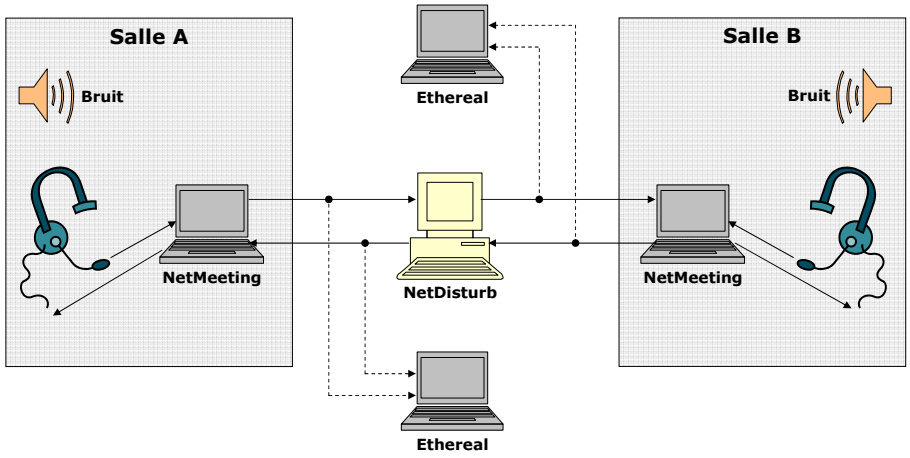


Figure 3.6 : *Schéma du dispositif - Test 2 sur les pertes de paquets et le bruit*

Facteurs	
Terminaux	micro-casques monauraux (Microsoft NetMeeting)
Codec	G.711 loi A (32 ms/paquet)
Nombre de paires de sujets	10 paires de sujets non experts
Mise en œuvre du test	carré gréco-latin
Méthodes de notation	ACR et DCR
Langue	français

Tableau 3.12 : *Facteurs expérimentaux - Test 2 sur les pertes de paquets et le bruit*

Les questions posées aux sujets dans chaque phase concernent la gêne due aux défauts, la gêne due au bruit et la qualité globale (*cf.* annexe E).

3.2.2.3 Analyse des résultats

L'analyse des enregistrements a montré que les microphones utilisés pendant ce test étaient très sélectifs, ne recueillant ainsi que le signal de parole et pas le bruit ambiant[1]. Cette forte sélectivité a eu deux conséquences :

- les enregistrements effectués à l'émission du bruit, après le microphone, ne contiennent pas le bruit ambiant et ne correspondent donc pas à ce qui a été perçu par les participants en présence de bruit ambiant,
- le bruit ambiant, diffusé du côté émission, n'a pas été transmis par les microphones et ainsi n'a pas été ressenti au niveau réception.

Les conditions avec bruit ambiant ne seront donc pas analysées, restent les neuf conditions avec bruit transmis (*i.e.* 1 à 3 et 10 à 15 pour les sujets A, et 1 à 9 pour les sujets B). Le rapport signal-à-bruit segmental ($RSBseg$) du côté réception du bruit transmis est calculé pour chaque condition grâce à l'équation 3.2

$$RSBseg = \frac{10}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \log \frac{\sum_{n \in \mathcal{N}} x\left(n,i\right)^2}{\sigma_b^2} \tag{3.2}$$

où $\mathcal{S}$ représente l'ensemble des trames qui contiennent de la parole (trames actives) dans le signal bruité x, et $|\mathcal{S}|$ son cardinal. $\mathcal{N}$ représente l'ensemble des échantillons dans la $i$ème trame active, $x\left(n,i\right)$ est le $n$ième échantillon dans la $i$ème trame active et σ_b^2 est la puissance moyenne du bruit estimée sur les trames non actives du signal bruité x. Cette analyse est effectuée avec des fenêtres de Hamming de 256 échantillons de longueur (32 ms) et avec un recouvrement de 128 échantillons entre deux trames successives.

Les conditions finalement retenues, renumérotées de 1 à 9, ainsi que les $RSBseg$ correspondants, sont présentés dans le tableau 3.13. Le niveau du bruit distant est indiqué, c'est-à-dire le niveau du bruit ambiant à l'émission avant transmission par le système. Les rapports signal-à-bruit segmentaux sont élevés ($RSBseg > 35$ dB), même quand il y a du bruit transmis, ce qui est expliqué par la sélectivité des microphones. De plus, le sujet du côté émission élève sa voix pour compenser la perte induite par le bruit ambiant (effet Lombard [Lombard 1911]), ce qui explique que les $RSBseg$ du côté réception sont quasiment identiques pour les trois niveaux de bruit. En présence de pertes de paquets, le bruit transmis est affecté par ces pertes, d'où l'augmentation des $RSBseg$ par rapport aux conditions sans perte de paquets.

Condition	Taux de pertes de paquets côtés A et B (%)	Niveau du bruit distant (dB(A))	Rapport signal-à-bruit segmental (dB)
1	0	0	35.2
2	5	0	38.3
3	10	0	39.9
4	0	49	36.9
5	5	49	39.7
6	10	49	41.9
7	0	59	36.5
8	5	59	38.4
9	10	59	41.1

Tableau 3.13 : *Conditions finales - Test 2 sur les pertes de paquets et le bruit*

L'analyse individuelle des données des sujets a montré qu'aucun sujet n'avait besoin d'être écarté. Les résultats présentés dans la suite sont donc la moyenne des données de vingt sujets.

[1] Nous appellerons « bruit ambiant » le bruit diffusé dans la salle, « bruit transmis » le bruit ambiant après transmission dans le système de test, « côté émission » le côté du système où le bruit ambiant est diffusé et « côté réception » le côté du système où le bruit ambiant est reçu après transmission.

Les effets de trois facteurs ont été analysés : le contexte dans lequel se trouve le sujet, le taux de pertes de paquets et le niveau de bruit. Le facteur Contexte peut prendre trois niveaux (locution, écoute, conversation). Le facteur Pertes de paquets (noté PP) présente trois niveaux (0, 5, 10%). Le facteur Bruit a trois niveaux (pas de bruit, 49, 59 dB(A)). Plusieurs questions ont été posées aux sujets (*cf.* tableau E.1 de l'annexe E). Ces questions portent sur trois critères de qualité (qualité globale, gêne due au bruit, gêne due aux défauts). L'analyse menée ici consiste à étudier l'effet des trois facteurs (Contexte, Pertes de paquets, Bruit) sur le jugement de qualité des sujets (A et B), pour chacun des trois critères.

Corrélations Les matrices de corrélation obtenues entre ces différents critères sont fournies dans les tableaux 3.14, 3.15 et 3.16, pour chacun des trois contextes (locution, écoute, conversation), respectivement. En contexte de locution, la qualité globale est corrélée avec le bruit et les défauts, et ce au même niveau. En contexte d'écoute, la corrélation la plus importante est celle entre les critères de gêne due aux défauts et de qualité globale. Le jugement obtenu en fonction du bruit est également corrélé à ceux obtenus avec le critère de qualité globale. Les sujets semblent donc avoir basé en grande partie leur jugement de la qualité globale en situation d'écoute sur les défauts entendus (*i.e.* les pertes de paquets). Les deux critères de gêne due au bruit et de gêne due aux défauts sont également corrélés entre eux, indiquant probablement que les sujets les ont partiellement confondus. En contexte de conversation, le jugement de la qualité globale semble avoir été influencé essentiellement par les défauts, et, à un niveau moindre, par le bruit. De nouveau, les critères de gêne due au bruit et de gêne due aux défauts sont corrélés.

	Bruit	**Défauts**	**Qualité globale**
Bruit	1	0.319	0.596
Défauts	0.319	1	0.607
Qualité globale	0.596	0.607	1

Tableau 3.14 : *Corrélation entre critères en contexte de locution - Test 2 sur les pertes de paquets et le bruit*

	Bruit	**Défauts**	**Qualité globale**
Bruit	1	0.471	0.533
Défauts	0.471	1	0.809
Qualité globale	0.533	0.809	1

Tableau 3.15 : *Corrélation entre critères en contexte d'écoute - Test 2 sur les pertes de paquets et le bruit*

	Bruit	**Défauts**	**Qualité globale**
Bruit	1	0.408	0.511
Défauts	0.408	1	0.768
Qualité globale	0.511	0.768	1

Tableau 3.16 : *Corrélation entre critères en contexte de conversation - Test 2 sur les pertes de paquets et le bruit*

La corrélation entre les trois contextes est aussi étudiée pour le critère de qualité globale et est présentée dans le tableau 3.17. Les contextes de locution et d'écoute sont peu corrélés entre eux. Le contexte de conversation est plus corrélé au contexte d'écoute qu'au contexte de locution. Ceci correspond à ce qui a été observé précédemment, à savoir que le critère de gêne due aux défauts affectant le contexte d'écoute a principalement influencé le jugement de la qualité globale.

Qualité globale La figure 3.7 montre les notes MOS de qualité globale et les intervalles de confiance à 95% correspondants, en fonction des conditions du test (niveau du bruit distant

	Conversation	Locution	Écoute
Conversation	1	0.258	0.465
Locution	0.258	1	0.221
Écoute	0.465	0.221	1

Tableau 3.17 : *Corrélation entre contextes pour le critère de qualité globale - Test 2 sur les pertes de paquets et le bruit*

Facteur	SC	dl	CM	F	P>F
Contexte	118.7	2	59.36	57.30	0.000*
Bruit	0.5	2	0.24	0.41	0.668
PP	95.7	2	47.87	99.25	0.000*
Contexte×Bruit	0.9	4	0.23	0.57	0.682
Contexte×PP	27.5	4	6.87	12.17	0.000*
Bruit×PP	4.2	4	1.04	1.85	0.128
Contexte×Bruit×PP	9.1	8	1.14	2.99	0.004*

Tableau 3.18 : *ANOVA pour le critère de qualité globale et pour les 3 contextes - Test 2 sur les pertes de paquets et le bruit*

en dB(A) et taux de pertes de paquets en %) et du contexte.

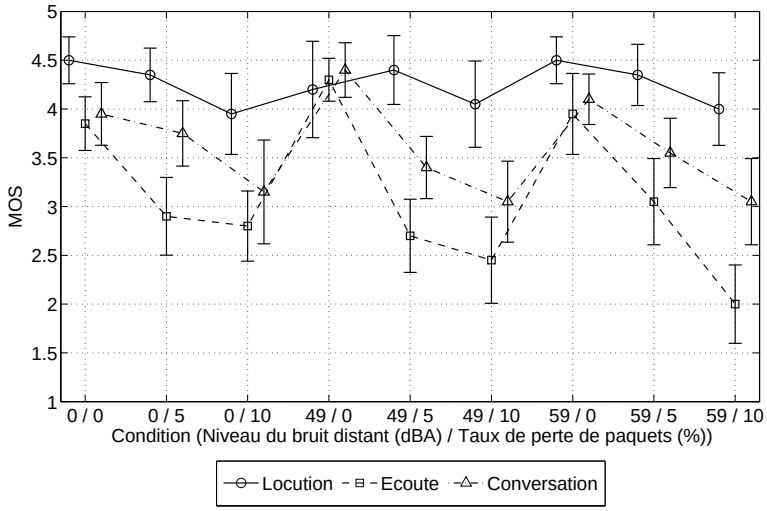


Figure 3.7 : *Notes MOS du critère de qualité globale et intervalles de confiance à 95% correspondants - Test sur les pertes de paquets et le bruit*

Il semble que le jugement global moyen soit influencé à la fois par le contexte et par les conditions testées. En contexte de locution, la qualité varie peu. En contexte de conversation, la qualité varie en fonction du taux de pertes de paquets, mais ne semble pas influencée par le niveau de bruit. En contexte d'écoute, la qualité varie en fonction du taux de pertes de paquets et du niveau de bruit distant.

Une ANOVA conduite sur les notes obtenues pour le critère de qualité globale, dont les résultats sont présentés dans le tableau 3.18, confirme les effets significatifs des deux facteurs Contexte et Pertes de paquets (PP), le facteur Bruit n'ayant pas d'effet significatif.

L'effet des pertes de paquets est très significatif ($p < 0.05$). Cet effet dépend du contexte, comme l'indique l'interaction significative entre les deux facteurs (PP et Contexte). Dans le contexte de locution, la note moyenne globale reste quasiment stable quand le taux de pertes de paquets augmente et quand le niveau du bruit distant augmente, ce qui est cohérent étant donné le niveau du bruit à la réception. Dans le contexte d'écoute, la note moyenne varie logiquement en fonction du taux de pertes de paquets, cette variation dépendant du niveau de bruit distant comme l'indique l'interaction significative entre les trois facteurs (Contexte, Bruit, PP). En présence de bruit, les pertes de paquets affectent le signal « parole + bruit ».

Facteur	SC	dl	CM	F	P>F
Contexte	227.4	2	113.7	101.4	0.000*
Bruit	3.2	2	1.6	1.9	0.160
PP	125.9	2	62.9	78.1	0.000*
Contexte×Bruit	0.8	4	0.2	0.4	0.789
Contexte×PP	44.1	4	11.0	17.8	0.000*
Bruit×PP	4.6	4	1.2	1.4	0.229
Contexte×Bruit×PP	4.9	8	0.6	1.3	0.253

Tableau 3.19 : *ANOVA pour le critère de gêne due aux défauts et pour les 3 contextes - Test 2 sur les pertes de paquets et le bruit*

Elles sont donc perceptibles dans les périodes de parole mais aussi dans les périodes de silence (*i.e.* de bruit seul). Dans le contexte de conversation, la note moyenne varie en fonction du taux de pertes de paquets, mais pas en fonction du niveau de bruit distant. Cette variation a moins d'amplitude que dans le contexte d'écoute, ce qui peut s'expliquer, comme dans le premier test, par l'attention partagée entre la conduite de la discussion et le jugement de la qualité en contexte de conversation. Le bruit n'a pas d'effet significatif, ce qui est probablement dû à son faible niveau à la réception.

Gêne due aux défauts La figure 3.8 montre les notes MOS de gêne due aux défauts et les intervalles de confiance à 95% correspondants, en fonction des conditions du test (niveau du bruit distant en dB(A) et taux de pertes de paquets en %) et du contexte.

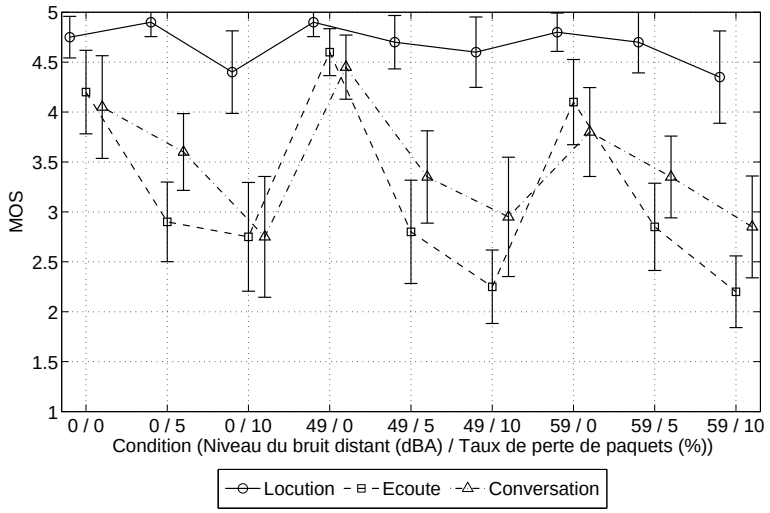


Figure 3.8 : *Notes MOS de gêne due aux défauts et intervalles de confiance à 95% correspondants - Test 2 sur les pertes de paquets et le bruit*

Comme pour le critère de qualité globale, le jugement moyen de la gêne due aux défauts est influencé par le contexte et par le taux de pertes de paquets.

Une ANOVA est conduite sur les notes obtenues pour la gêne due aux défauts pour les trois contextes, en considérant les facteurs Contexte, Bruit et Perte de paquets (PP). Ses résultats sont présentés dans le tableau 3.19. L'ANOVA confirme les effets significatifs des deux facteurs Contexte et Pertes de paquets. L'effet du contexte est très significatif, principalement dû au contexte d'écoute. En toute logique, les pertes de paquets ont un effet significatif, qui dépend du contexte (interaction significative), et le bruit n'a pas d'effet significatif.

Gêne due au bruit La figure 3.9 montre les notes MOS de gêne due au bruit et les intervalles de confiance à 95% correspondants, en fonction des conditions du test (niveau du bruit distant en dB(A) et taux de pertes de paquets en %) et du contexte.

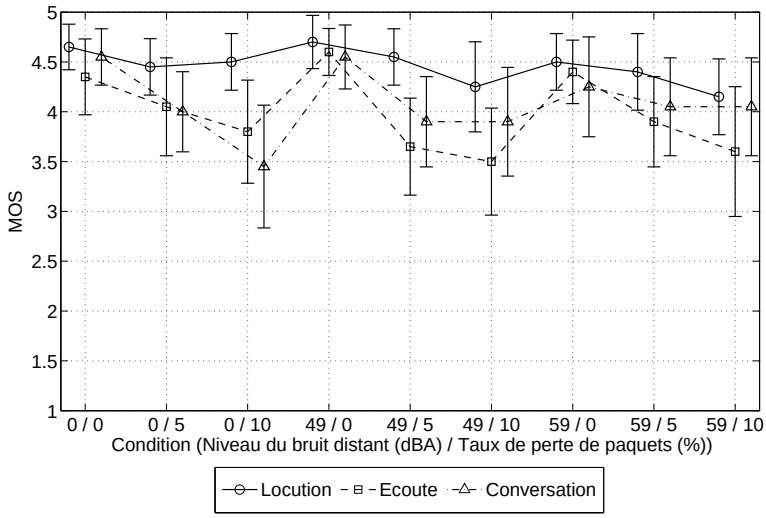


Figure 3.9 : *Notes MOS du critère de gêne due au bruit et intervalles de confiance à 95% correspondants - Test 2 sur les pertes de paquets et le bruit*

Facteur	SC	dl	CM	F	P>F
Contexte	23.05	2	11.52	9.37	0.000*
Bruit	0.28	2	0.14	0.34	0.714
PP	33.07	2	16.54	16.42	0.000*
Contexte×Bruit	2.11	4	0.53	1.13	0.349
Contexte×PP	4.65	4	1.16	2.16	0.081
Bruit×PP	1.99	4	0.50	0.79	0.538
Contexte×Bruit×PP	6.19	8	0.77	2.09	0.040*

Tableau 3.20 : *ANOVA pour le critère de gêne due au bruit et pour les 3 contextes - Test 2 sur les pertes de paquets et le bruit*

Comme pour le critère de qualité globale, le jugement moyen de la gêne due au bruit est influencé par le contexte et par le taux de pertes de paquets, mais peu par le bruit.

Une ANOVA est conduite sur les notes obtenues pour la gêne due au bruit pour les trois contextes, en considérant les facteurs Contexte, Bruit et Perte de paquets (PP). Le tableau 3.20 présente les résultats de l'ANOVA. Elle confirme l'effet significatif du contexte et des pertes de paquets, et la non-significativité du bruit, ce qui est contradictoire. Il semblerait que les sujets n'aient pas fait la différence entre le bruit et les pertes de paquets, ce qui peut probablement être expliqué par la faible gêne due au bruit comparée à la gêne due aux pertes de paquets.

3.2.2.4 Synthèse

L'objectif de ce test était de tester l'impact des pertes de paquets et du bruit. Un seul des deux objectifs est atteint : l'impact des pertes de paquets. L'effet du bruit n'a pas pu être testé car la sélectivité des microphones n'avait pas été anticipée. Un autre test sur le bruit est donc nécessaire pour en étudier l'effet de façon plus approfondie.

3.2.3 Test 3 : bruit

3.2.3.1 Objectif

L'objectif de ce test est d'étudier l'effet du bruit sur la qualité de conversation. Cette dégradation est étudiée car :

- le bruit n'a pas pu être étudié dans le test 2 (sélectivité de la prise de son au niveau acoustique),

– le bruit engendre *a priori* un effet sur la qualité d'écoute, la qualité de locution et la qualité de conversation.

3.2.3.2 Conditions et facteurs expérimentaux

Sept conditions de test ont été choisies et sont fournies dans le tableau 3.21. Le bruit est introduit symétriquement dans le système et est donc au même niveau (électrique et acoustique) des deux côtés. Deux types de bruit à trois niveaux différents sont utilisés : un bruit de Hoth (stationnaire) et des voix enregistrées dans un restaurant (non stationnaire). Le rapport signal-à-bruit segmental ($RSBseg$) est calculé sur le signal reçu par chaque participant pour chaque condition grâce à l'équation 3.2 et est présenté dans le tableau 3.21. Le rapport signal-à-bruit diminue rapidement quand le niveau acoustique du bruit augmente. Le bruit est introduit au niveau électrique dans le circuit et transmis sans atténuation dans le système, ce qui aboutit à des RSB faibles en présence de bruit.

Condition	Type de bruit	Niveau du bruit côtés A et B (dB(A))	Rapport signal-à-bruit segmental (dB)
1	Pas de bruit	Pas de bruit	28.9
2	Hoth	48	14.5
3	Hoth	53	11.2
4	Hoth	59	10.6
5	Restaurant	51	16.9
6	Restaurant	57	13.7
7	Restaurant	63	11.6

Tableau 3.21 : *Conditions - Test 3 sur le bruit*

Le dispositif de test, dont le schéma est donné dans la figure 3.10, relie deux téléphones analogiques et le PABX gère les indicatifs correspondant chacun à une condition de test. Le bruit est introduit symétriquement dans la communication grâce à deux extracteurs/inserteurs de trames TE820 et à un lecteur numérique Edirol R-4 contenant les fichiers de bruit au format wav. Les signaux sont enregistrés en extrayant les trames de parole avec les deux TE820 et en les envoyant vers un enregistreur numérique Edirol R-4. Les quatre signaux enregistrés sont synchrones.

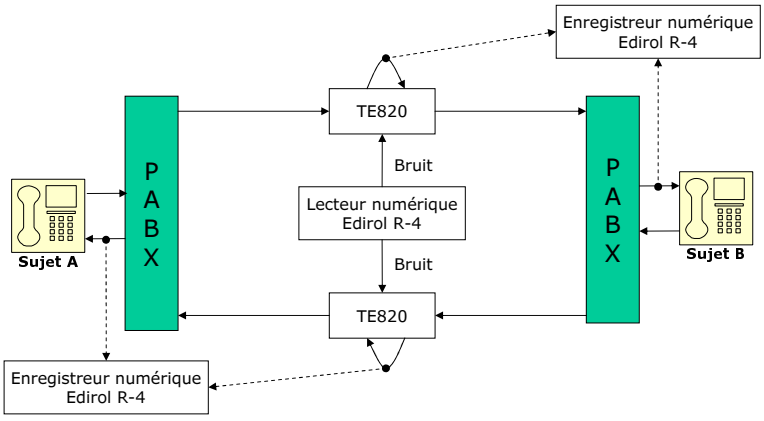


Figure 3.10 : *Schéma du dispositif - Test 3 sur le bruit*

Sept paires de sujets non experts (7 femmes et 7 hommes), âgés de 18 à 55 ans, ont participé au test. *A priori* otologiquement sains, ils ont tous été recrutés en dehors de France Télécom et rémunérés. La mise en œuvre du test a été faite à l'aide d'un carré gréco-latin.

Les sujets s'entraînaient au début du test en passant deux conditions d'apprentissage : les conditions 1 (pas de dégradation) et 3 (bruit de Hoth à 53 dB(A) des deux côtés). Les

facteurs expérimentaux sont fournis dans le tableau 3.22. Les questions posées aux sujets à l'issue de chaque phase concernent la gêne due au bruit et la qualité globale (*cf.* annexe E).

Facteurs	
Terminaux	téléphones analogiques
Codec	G.711
Nombre de paires de sujets	7 paires de sujets non experts
Mise en œuvre du test	carré gréco-latin
Méthodes de notation	ACR et DCR
Langue	français

Tableau 3.22 : *Facteurs expérimentaux - Test 3 sur le bruit*

3.2.3.3 Analyse des résultats

Les données d'un sujet ont été écartées, suite à des difficultés de compréhension et d'utilisation du logiciel rencontrées pendant le test. Les résultats présentés dans la suite correspondent donc à la moyenne des données de treize sujets.

Les effets de deux facteurs ont été analysés : le contexte dans lequel se trouve le sujet et le bruit. Le facteur Contexte peut prendre trois niveaux (locution, écoute, conversation). Le facteur Bruit présente sept niveaux. Plusieurs questions ont été posées aux sujets selon le contexte dans lequel ils se trouvaient (*cf.* tableau E.1 de l'annexe E). Ces questions portent sur deux critères de qualité (qualité globale, gêne due au bruit). L'analyse menée ici consiste à étudier l'effet des deux facteurs (Contexte, Bruit) sur le jugement de qualité des sujets, pour chacun des deux critères.

Corrélations Les matrices de corrélations obtenues entre ces deux critères sont fournies dans les tableaux 3.23, 3.24 et 3.25, pour chacun des trois contextes (locution, écoute, conversation), respectivement. Dans les trois contextes, la qualité globale est très corrélée avec le bruit, ce qui s'explique par les dégradations testées.

	Bruit	**Qualité globale**
Bruit	1	0.860
Qualité globale	0.860	1

Tableau 3.23 : *Corrélation entre critères en contexte de locution - Test 3 sur le bruit*

	Bruit	**Qualité globale**
Bruit	1	0.822
Qualité globale	0.822	1

Tableau 3.24 : *Corrélation entre critères en contexte d'écoute - Test 3 sur le bruit*

	Bruit	**Qualité globale**
Bruit	1	0.808
Qualité globale	0.808	1

Tableau 3.25 : *Corrélation entre critères en contexte de conversation - Test 3 sur le bruit*

La corrélation entre les trois contextes pour le critère de qualité globale est aussi étudiée et présentée dans le tableau 3.26. Les trois contextes sont très corrélés entre eux, le bruit les affectant tous.

Qualité globale La figure 3.11 montre les notes MOS de qualité globale et les intervalles de confiance à 95% correspondants, en fonction des conditions du test (type de bruit et niveau du bruit en dB(A)) et du contexte.

	Conversation	Locution	Écoute
Conversation	1	0.783	0.767
Locution	0.783	1	0.763
Écoute	0.767	0.763	1

Tableau 3.26 : *Corrélation entre contextes pour le critère de qualité globale - Test 3 sur le bruit*

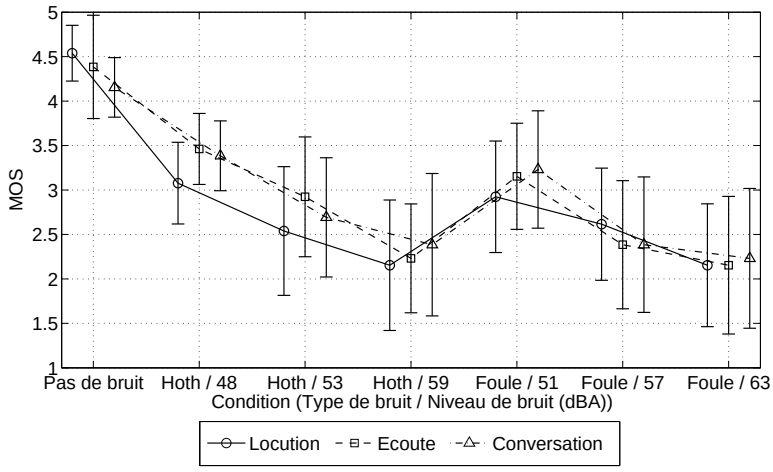


Figure 3.11 : *Notes MOS du critère de qualité globale et intervalles de confiance à 95% correspondants - Test 3 sur le bruit*

Facteur	SC	dl	CM	F	P>F
Contexte	0.5	2	0.23	0.38	0.686
Bruit	136.2	6	22.71	18.03	0.000*
Contexte×Bruit	4.1	12	0.34	1.10	0.368

Tableau 3.27 : *ANOVA pour le critère de qualité globale et pour les 3 contextes - Test 3 sur le bruit*

Facteur	SC	dl	CM	F	P>F
Contexte	0.0	2	0.00	0.01	0.989
Bruit	215.1	6	35.84	39.75	0.000*
Contexte×Bruit	7.9	12	0.66	2.24	0.013*

Tableau 3.28 : *ANOVA pour le critère de gêne due au bruit et pour les 3 contextes - Test 3 sur le bruit*

Il semble que le jugement global moyen soit influencé par les conditions testées, mais pas par le contexte. Quels que soient le contexte et le type de bruit, la qualité diminue quand le niveau du bruit augmente.

Une ANOVA conduite sur les notes obtenues pour le critère de qualité globale, dont les résultats sont présentés dans le tableau 3.27, confirme l'effet significatif du facteur Bruit et la non-significativité de l'effet du facteur Contexte.

Gêne due au bruit La figure 3.12 montre les notes MOS de gêne due au bruit et les intervalles de confiance à 95% correspondants, en fonction des conditions du test (type de bruit et niveau du bruit en dB(A)) et du contexte.

Il semble que, comme pour le critère de qualité globale, le jugement moyen soit influencé par les conditions testées, mais pas par le contexte.

Une ANOVA conduite sur les notes obtenues pour le critère de gêne due au bruit confirme l'effet significatif du facteur Bruit et la non-significativité de l'effet du facteur Contexte. L'interaction entre le contexte et le bruit est significative. Le tableau 3.28 présente les résultats de l'ANOVA.

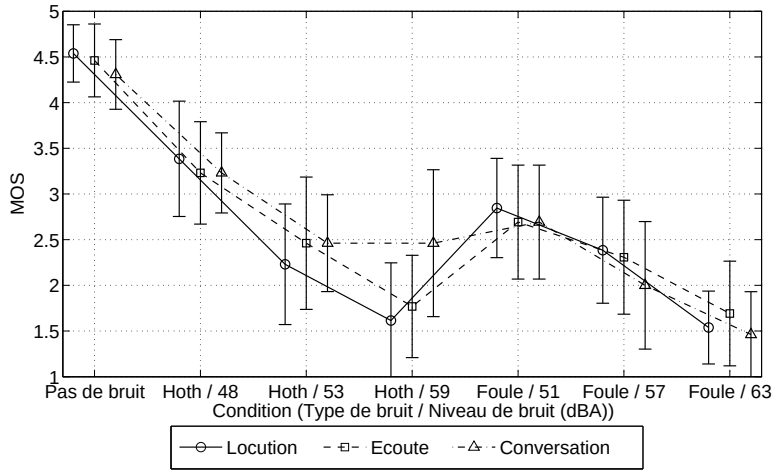


Figure 3.12 : *Notes MOS du critère de gêne due au bruit et intervalles de confiance à 95% correspondants - Test 3 sur le bruit*

3.2.3.4 Synthèse

Le test 3 a permis de tester l'effet du bruit en ligne sur les qualités de locution, d'écoute et de conversation. Celles-ci sont toutes influencées de manière très forte et similaire par le bruit, les notes correspondant à ces trois critères sont donc très corrélées entre elles. Les ANOVA effectuées pour les critères de qualité globale et de gêne due au bruit confirment que le bruit a un effet très significatif et que le contexte n'a pas d'effet significatif sur le jugement des sujets. Ce test, tel qu'il a été conçu, ne permet pas d'appréhender la qualité lorsque le participant est immergé en milieu bruité.

3.2.4 Test 4 : écho, délai et pertes de paquets

3.2.4.1 Objectif

L'objectif de ce test est d'étudier des dégradations déjà testées dans les tests précédents, ainsi que de nouvelles conditions. Les dégradations choisies sont le délai, l'écho et les pertes de paquets aléatoires, car :

- elles ont déjà été testées séparément lors de deux tests précédents,
- elles ont *a priori* un effet sur la qualité d'écoute, la qualité de locution et la qualité de conversation.

3.2.4.2 Conditions et facteurs expérimentaux

Vingt-et-une conditions de test ont été choisies et sont fournies dans le tableau 3.29. Les valeurs de délai (0, 200 et 400 ms) ont déjà été testées (*cf.* Test 1), les valeurs d'atténuation d'écho (20 et 30 dB) n'ont pas encore été étudiées mais sont proches de la valeur de 25 dB testée lors du premier test, les valeurs du taux de pertes de paquets (0, 5 et 10%) sont les mêmes que lors du deuxième test.

Neuf paires de sujets non experts (9 femmes et 9 hommes), âgés de 20 à 60 ans, ont participé au test. *A priori* otologiquement sains, ils ont tous été recrutés en dehors de France Télécom et rémunérés. La mise en œuvre du test a été faite à l'aide d'un carré gréco-latin.

Les sujets s'entraînaient au début du test en expérimentant deux conditions d'apprentissage : les conditions 1 (pas de dégradation) et 20 (délai de 400 ms, atténuation d'écho de 30 dB, taux de pertes de paquets de 5%).

Le montage du test, dont le schéma est fourni dans la figure 3.13, relie deux combinés RNIS, le PABX gère les indicatifs correspondant chacun à une condition de test. Le délai dans chaque sens est généré à l'aide d'une ligne à retard de 200 ms, dans laquelle le signal de

Condition	Délai (ms)	Atténuation de l'écho (dB)	Taux de pertes de paquets (%)
1	0	Pas d'écho	0
2	0	Pas d'écho	5
3	0	Pas d'écho	10
4	200	Pas d'écho	0
5	200	Pas d'écho	5
6	200	Pas d'écho	10
7	200	20	0
8	200	20	5
9	200	20	10
10	200	30	0
11	200	30	5
12	200	30	10
13	400	Pas d'écho	0
14	400	Pas d'écho	5
15	400	Pas d'écho	10
16	400	20	0
17	400	20	5
18	400	20	10
19	400	30	0
20	400	30	5
21	400	30	10

Tableau 3.29 : *Conditions - Test 4 sur l'écho, le délai et les pertes de paquets*

parole passe le nombre de fois correspondant à la condition testée. L'écho est créé en atténuant le signal de parole transitant de A vers B (resp. B vers A) et en le réinjectant dans le sens B vers A (resp. A vers B). Les pertes de paquets aléatoires sont générées symétriquement avec le logiciel NetDisturb. Les signaux sont enregistrés en extrayant les trames de parole et en les envoyant vers un enregistreur numérique Edirol R-4. Les enregistrements sont synchrones.

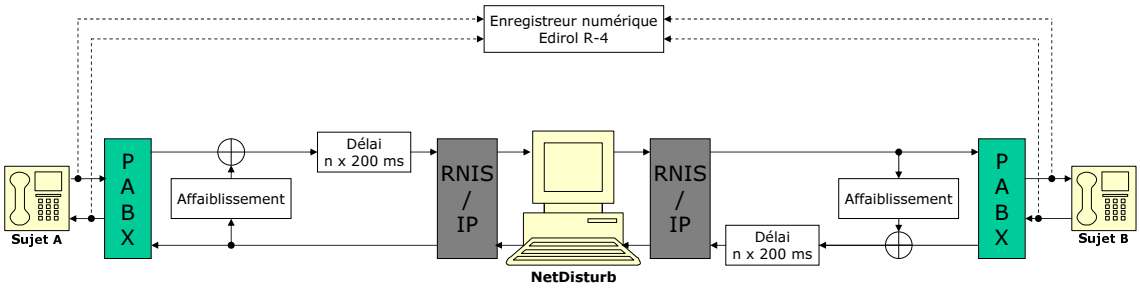


Figure 3.13 : *Schéma du dispositif - Test 4 sur l'écho, le délai et les pertes de paquets*

Les facteurs expérimentaux du test sont fournis dans le tableau 3.30.

Facteurs	
Terminaux	téléphones RNIS
Codec	G.711
Nombre de paires de sujets	9 paires de sujets non experts
Mise en œuvre du test	carré gréco-latin
Méthodes de notation	ACR et DCR
Langue	français

Tableau 3.30 : *Facteurs expérimentaux - Test 4 sur l'écho, le délai et les pertes de paquets*

3.2.4.3 Analyse des résultats

Après analyse des notes individuelles, aucun sujet n'a été rejeté. Les effets de quatre facteurs ont été analysés : le contexte dans lequel se trouve le sujet, le délai, l'atténuation

de l'écho et les pertes de paquets. Le facteur Contexte peut prendre trois niveaux (locution, écoute, conversation). Le facteur Délai présente trois niveaux (0, 200, 400 ms). Le facteur Atténuation a trois niveaux (pas d'écho électrique, 30 dB, 20 dB). Le facteur Pertes de paquets (noté PP) présente trois niveaux (0, 5, 10%). Plusieurs questions ont été posées aux sujets selon le contexte dans lequel ils se trouvaient (*cf.* tableau E.1 de l'annexe E). Ces questions portent sur quatre critères de qualité (qualité globale, gêne due à l'écho, gêne due aux défauts, interruption). L'analyse menée ici consiste à étudier l'effet des quatre facteurs (Contexte, Délai, Atténuation, PP) sur le jugement de qualité des sujets, pour chacun des quatre critères.

Corrélations Nous avons tout d'abord étudié la corrélation qui existait entre ces différents critères pour chacun des trois contextes (locution, écoute, conversation).

En contexte de locution, les critères évalués sont la qualité globale, la gêne due à l'écho et la gêne due aux défauts. La matrice de corrélation obtenue entre ces trois critères est donnée dans le tableau 3.31. La qualité globale est essentiellement corrélée avec l'écho, moins avec les défauts, ce qui est logique en contexte de locution dans lequel les pertes de paquets ne sont perceptibles qu'en présence d'écho et/ou de bruit transmis.

	Écho	**Défauts**	**Qualité globale**
Écho	1	0.224	0.751
Défauts	0.224	1	0.381
Qualité globale	0.751	0.381	1

Tableau 3.31 : *Corrélation entre critères en contexte de locution - Test 4 sur l'écho, le délai et les pertes de paquets*

En contexte d'écoute, les critères évalués sont la qualité globale et la gêne due aux défauts. La matrice de corrélation obtenue entre ces deux critères est donnée dans le tableau 3.32. La qualité globale est très corrélée avec les défauts.

	Défauts	**Qualité globale**
Défauts	1	0.847
Qualité globale	0.847	1

Tableau 3.32 : *Corrélation entre critères en contexte d'écoute - Test 4 sur l'écho, le délai et les pertes de paquets*

En contexte de conversation, les critères évalués sont la qualité globale, la gêne due à l'écho, la gêne due aux défauts et l'interruption. La matrice de corrélation obtenue entre ces quatre critères est donnée dans le tableau 3.33. La qualité globale est le plus corrélée avec les défauts et moins fortement avec l'écho et l'interruption. Les autres critères sont peu corrélés entre eux. Il semblerait donc qu'en contexte de conversation les sujets aient basé leur jugement de la qualité globale en premier lieu sur la gêne due aux défauts et ensuite sur la gêne due à l'écho.

	Écho	**Défauts**	**Interruption**	**Qualité globale**
Écho	1	0.121	0.252	0.426
Défauts	0.121	1	0.418	0.675
Interruption	0.252	0.418	1	0.461
Qualité globale	0.426	0.675	0.461	1

Tableau 3.33 : *Corrélation entre critères en contexte de conversation - Test 4 sur l'écho, le délai et les pertes de paquets*

La corrélation entre les trois contextes pour le critère de qualité globale est aussi étudiée et présentée dans le tableau 3.34. Les contextes de locution et d'écoute sont peu corrélés entre eux. Le contexte de conversation est corrélé de manière quasiment similaire avec les contextes de locution et d'écoute, ce qui est cohérent étant données les dégradations étudiées dans ce test.

	Conversation	Locution	Écoute
Conversation	1	0.497	0.549
Locution	0.497	1	0.228
Écoute	0.549	0.228	1

Tableau 3.34 : *Corrélation entre contextes pour le critère de qualité globale - Test 4 sur l'écho, le délai et les pertes de paquets*

Qualité globale La figure 3.14 montre les notes MOS de qualité globale et les intervalles de confiance à 95% correspondants, en fonction des conditions du test et du contexte.

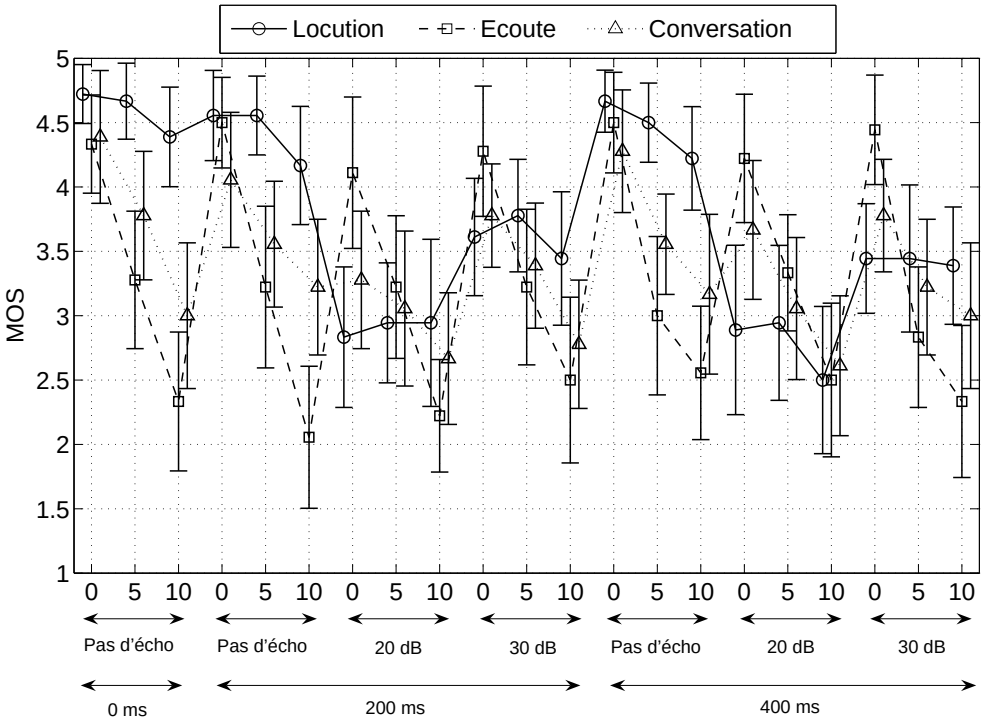


Figure 3.14 : *Notes MOS du critère de qualité globale et intervalles de confiance à 95% correspondants - Test 4 sur l'écho, le délai et les pertes de paquets. La première abscisse indique le taux de pertes de paquets (en %), la deuxième l'atténuation de l'écho et la troisième la valeur du délai unidirectionnel*

Il semble que le jugement global moyen soit influencé par le contexte et par les conditions testées. En contexte de locution, la qualité varie en fonction du délai et de l'atténuation d'écho. En contexte d'écoute, la qualité varie en fonction du taux de pertes de paquets. En contexte de conversation, la qualité semble varier essentiellement en fonction du taux de pertes de paquets et semble moins influencée par le délai et l'écho.

Une ANOVA conduite sur les notes obtenues pour le critère de qualité globale confirme les effets significatifs des facteurs Contexte, Atténuation et Pertes de paquets (PP), le facteur Délai n'ayant pas d'effet significatif. Le tableau 3.35 présente les résultats de l'ANOVA.

L'effet de la perte de paquets est très significatif ($p < 0.05$). Cet effet dépend du contexte, comme l'indique l'interaction significative entre les deux facteurs (PP et Contexte). Dans le contexte de locution, la note moyenne globale reste quasiment stable quand le taux de pertes de paquets augmente. Dans le contexte d'écoute, la note moyenne varie logiquement en fonction du taux de pertes de paquets, tout comme dans le contexte de conversation.

L'atténuation de l'écho a un effet significatif, qui dépend du contexte considéré comme le montre l'interaction significative entre les deux facteurs (Atténuation et Contexte). En contexte de locution, la note moyenne augmente quand l'atténuation d'écho augmente. En contexte d'écoute, il n'y a pas d'effet de l'atténuation d'écho sur la qualité, ce qui est sans surprise. En contexte de conversation, un léger effet de l'atténuation sur la qualité est observé.

Facteur	SC	dl	CM	F	P>F
Contexte	18.9	2	9.47	4.47	0.019*
Délai	0.0	1	0.00	0.02	0.891
Atténuation	89.6	2	44.78	34.84	0.000*
PP	173.6	2	86.78	73.30	0.000*
Contexte×Délai	1.2	2	0.58	1.55	0.226
Contexte×Atténuation	68.6	4	17.16	16.62	0.000*
Délai×Atténuation	1.3	2	0.65	1.31	0.283
Contexte×PP	89.5	4	22.38	26.34	0.000*
Délai×PP	2.0	2	0.99	2.52	0.095
Atténuation×PP	3.2	4	0.80	0.94	0.446
Contexte×Délai×Atténuation	0.8	4	0.21	0.54	0.706
Contexte×Délai×PP	1.3	4	0.31	0.71	0.588
Contexte×Atténuation×PP	1.5	8	0.19	0.35	0.943
Délai×Atténuation×PP	1.5	4	0.38	0.52	0.719
Contexte×Délai×Atténuation×PP	3.2	8	0.41	0.96	0.474

Tableau 3.35 : *ANOVA pour le critère de qualité globale et pour les 3 contextes - Test 4 sur l'écho, le délai et les pertes de paquets*

Gêne due aux défauts La figure 3.15 montre les notes MOS de gêne due aux défauts et les intervalles de confiance à 95% correspondants, en fonction des conditions du test et du contexte.

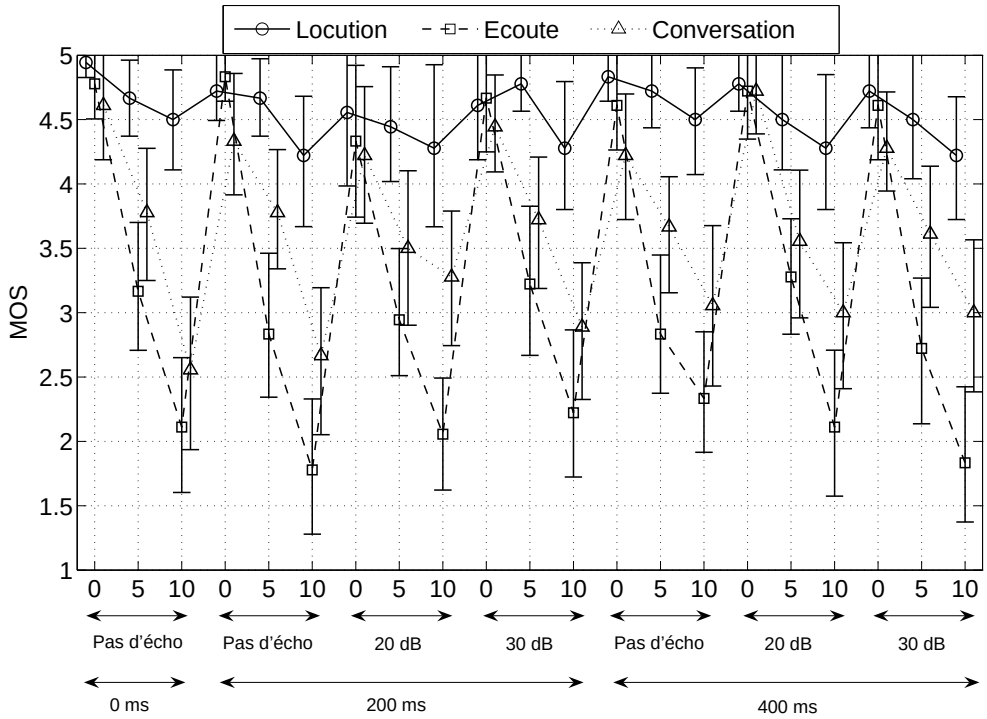


Figure 3.15 : *Notes MOS du critère de gêne due aux défauts et intervalles de confiance à 95% correspondants - Test 4 sur l'écho, le délai et les pertes de paquets. La première abscisse indique le taux de pertes de paquets (en %), la deuxième l'atténuation de l'écho et la troisième la valeur du délai unidirectionnel*

Il semble que le jugement de gêne due aux défauts soit influencé par le contexte et par les conditions testées. En contexte de locution, la qualité varie peu. En contextes d'écoute et de conversation, la qualité varie en fonction du taux de pertes de paquets, de façon plus prononcée en contexte d'écoute.

Une ANOVA conduite sur les notes obtenues pour le critère de gêne due aux défauts confirme les effets significatifs des facteurs Contexte et Pertes de paquets (PP), les facteurs Délai et Atténuation n'ayant pas d'effets significatifs. Le tableau 3.36 présente les résultats

Facteur	SC	dl	CM	F	P>F
Contexte	289.9	2	144.9	80.8	0.000*
Délai	0.3	1	0.3	1.0	0.335
Atténuation	0.0	2	0.0	0.0	0.968
PP	346.0	2	173.0	101.2	0.000*
Contexte×Délai	0.1	2	0.0	0.1	0.939
Contexte×Atténuation	1.6	4	0.4	0.7	0.612
Délai×Atténuation	4.2	2	2.1	3.9	0.030*
Contexte×PP	135.6	4	33.9	45.6	0.000*
Délai×PP	1.0	2	0.5	0.9	0.404
Atténuation×PP	0.8	4	0.2	0.2	0.910
Contexte×Délai×Atténuation	1.7	4	0.4	0.9	0.465
Contexte×Délai×PP	0.1	4	0.0	0.1	0.989
Contexte×Atténuation×PP	4.1	8	0.5	1.2	0.293
Délai×Atténuation×PP	6.4	4	1.6	3.0	0.023*
Contexte×Délai×Atténuation×PP	2.7	8	0.3	0.9	0.549

Tableau 3.36 : *ANOVA pour le critère de gêne due aux défauts et pour les 3 contextes - Test 4 sur l'écho, le délai et les pertes de paquets*

de l'ANOVA.

L'effet de la perte de paquets est très significatif ($p < 0.05$). Cet effet dépend du contexte, comme l'indique l'interaction significative entre les deux facteurs (PP et Contexte). Dans le contexte de locution, la note moyenne globale reste quasiment stable quand le taux de pertes de paquets augmente, alors que dans les contextes d'écoute et de conversation, la note moyenne varie en fonction du taux de pertes de paquets.

Gêne due à l'écho La figure 3.16 montre les notes MOS de gêne due à l'écho et les intervalles de confiance à 95% correspondants, en fonction des conditions du test et du contexte.

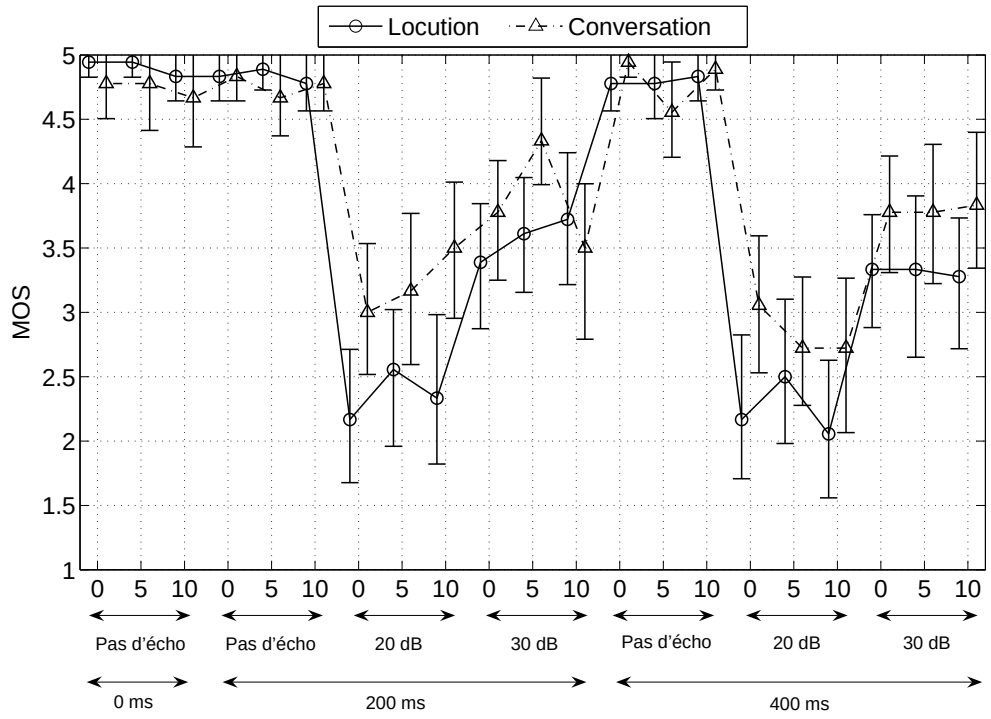


Figure 3.16 : *Notes MOS du critère de gêne due à l'écho et intervalles de confiance à 95% correspondants - Test 4 sur l'écho, le délai et les pertes de paquets. La première abscisse indique le taux de pertes de paquets (en %), la deuxième l'atténuation de l'écho et la troisième la valeur du délai unidirectionnel*

Il semble que le jugement de la gêne due à l'écho soit influencé par le contexte et par les

Facteur	SC	dl	CM	F	P>F
Contexte	21.1	1	21.1	15.06	0.001*
Délai	3.1	1	3.1	8.30	0.010*
Atténuation	493.1	2	246.6	97.87	0.000*
PP	0.6	2	0.3	0.47	0.630
Contexte×Délai	0.0	1	0.0	0.00	0.957
Contexte×Atténuation	16.0	2	8.0	14.76	0.000*
Délai×Atténuation	1.8	2	0.9	1.53	0.232
Contexte×PP	1.0	2	0.5	1.09	0.349
Délai×PP	2.0	2	1.0	2.65	0.085
Atténuation×PP	2.4	4	0.6	1.31	0.275
Contexte×Délai×Atténuation	1.6	2	0.8	2.52	0.095
Contexte×Délai×PP	0.9	2	0.5	1.43	0.254
Contexte×Atténuation×PP	4.2	4	1.1	2.35	0.063
Délai×Atténuation×PP	2.8	4	0.7	1.42	0.236
Contexte×Délai×Atténuation×PP	2.5	4	0.6	2.00	0.105

Tableau 3.37 : *ANOVA pour le critère de gêne due à l'écho et pour les 2 contextes - Test 4 sur l'écho, le délai et les pertes de paquets*

Facteur	SC	dl	CM	F	P>F
Délai	0.077	1	0.077	0.371	0.550
Atténuation	1.210	2	0.605	1.581	0.221
PP	4.321	2	2.160	5.160	0.011*
Délai×Atténuation	0.914	2	0.457	1.373	0.267
Délai×PP	0.691	2	0.346	0.741	0.484
Atténuation×PP	0.346	4	0.086	0.197	0.939
Délai×Atténuation×PP	1.790	4	0.448	1.156	0.338

Tableau 3.38 : *ANOVA pour le critère de possibilité d'interruption et pour le contexte de conversation - Test 4 sur l'écho, le délai et les pertes de paquets*

conditions testées. En contexte de locution, la perception de gêne due à l'écho varie beaucoup, essentiellement en fonction de l'atténuation de l'écho. En contexte de conversation, la perception de gêne due à l'écho varie en fonction de l'atténuation de l'écho et du délai.

Une ANOVA conduite sur les notes obtenues pour le critère de gêne due à l'écho confirme les effets significatifs des facteurs Contexte, Délai et Atténuation, le facteur Pertes de paquets n'ayant pas d'effet significatif. Le tableau 3.37 présente les résultats de l'ANOVA.

L'effet de l'atténuation d'écho est très significatif ($p < 0.05$). Il dépend du contexte, comme l'indique l'interaction significative entre les deux facteurs (Atténuation et Contexte). En effet, dans le contexte de locution, la note moyenne varie plus en fonction de l'atténuation d'écho qu'en contexte de conversation.

Interruption La figure 3.17 montre les notes MOS de possibilité d'interruption et les intervalles de confiance à 95% correspondants, en fonction des conditions du test dans le contexte de conversation.

Le jugement de la possibilité d'interruption est quasiment constant. Une ANOVA conduite sur les notes obtenues pour le critère de possibilité d'interruption, dont les résultats sont présentés dans le tableau 3.38, montre un effet significatif du facteur Pertes de paquets, les facteurs Délai et Atténuation n'ayant pas d'effet significatif.

Contrairement à ce qui était attendu, le délai n'a pas d'effet significatif sur la possibilité d'interruption alors que les pertes de paquets en ont un. Ceci peut être expliqué par la difficulté probable des sujets à évaluer ce critère d'interruption.

3.2.4.4 Synthèse

Ce test a tout d'abord permis de tester des dégradations déjà étudiées dans les tests précédents, ainsi que de nouvelles dégradations et combinaisons de dégradations non encore

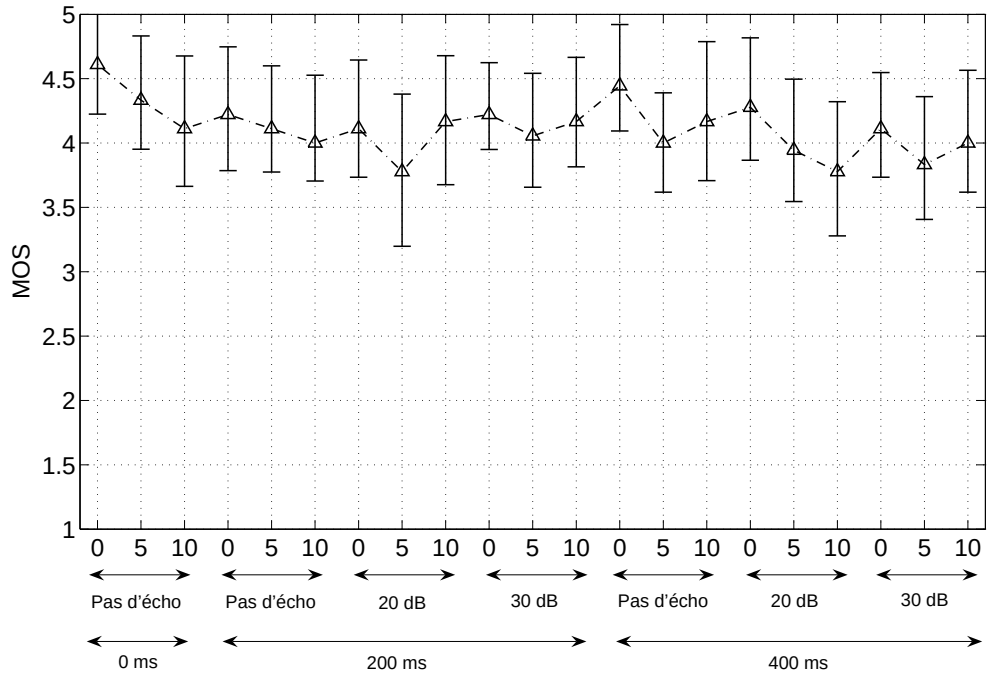


Figure 3.17 : *Notes MOS du critère de possibilité d'interruption et intervalles de confiance à 95% correspondants - Test 4 sur l'écho, le délai et les pertes de paquets. La première abscisse indique le taux de pertes de paquets (en %), la deuxième l'atténuation de l'écho et la troisième la valeur du délai unidirectionnel*

testées. L'analyse des résultats montre une corrélation des contextes d'écoute et de locution avec le contexte de conversation pour le critère de qualité globale, qui s'explique par les dégradations étudiées affectant à la fois la qualité d'écoute et la qualité de locution. L'ANOVA effectuée pour le critère de qualité globale montre que le jugement des sujets a été influencé par le contexte, l'atténuation d'écho et la perte de paquets, mais pas par le délai. Ceci confirme les résultats du premier test sur l'écho et le délai à savoir que le délai a peu d'effet quand il est inférieur à 400 ms.

3.3 Relation entre les différentes composantes de la qualité vocale

Les données recueillies lors de ces différents tests subjectifs ont pour but de déterminer la relation, si elle existe, entre la qualité de conversation et la qualité d'écoute, la qualité de locution et le délai. Le modèle doit idéalement savoir estimer la qualité de conversation d'une communication, sans se focaliser sur une dégradation en particulier. Nous utiliserons donc les notes correspondant au critère de qualité globale récoltées lors des tests subjectifs pour déterminer cette relation.

Nous choisissons d'appliquer une régression linéaire multiple (sans termes d'interaction entre régresseurs) pour estimer la note de conversation à partir des notes de locution et d'écoute et du délai (*cf.* annexe D). Ce choix est tout d'abord guidé par la simplicité de mise en œuvre de la régression linéaire multiple, mais aussi par analogie possible entre l'évaluation objective de la qualité de conversation - telle qu'elle est abordée dans cette étude, c'est-à-dire comme une décomposition en trois composantes (qualités de locution, d'écoute et d'interaction) - et l'évaluation objective de la qualité audiovisuelle, constituée de deux composantes (qualités audio et vidéo). Les études sur la qualité audiovisuelle ont souvent abouti à une régression linéaire multiple (avec ou sans terme d'interaction entre les qualités audio et vidéo) [Jones et Atkinson 1998, Pastrana-Vidal *et al.* 2003, Puglia 2005, Tebaldi 2005, Kitawaki

et al. 2005]. Le terme d'interaction entre composantes ne semble pas nécessaire étant données les matrices de corrélation entre contextes obtenues pour les différents tests présentés précédemment. En effet, la qualité d'écoute et la qualité de locution sont peu corrélées entre elles, à l'exception du test sur le bruit (*cf.* tableau 3.26). Dans ce test, les qualités dans les trois contextes (écoute, locution, conversation) sont très corrélées entre elles, puisque le bruit les affecte toutes.

L'analyse du premier test a montré que le délai a peu d'effet tant qu'il est inférieur à 400 ms. Ce résultat concorde avec des études récentes qui mettent en évidence que le délai n'est plus aussi gênant qu'auparavant. Nous avons donc choisi de prendre en compte l'effet du délai sous la forme d'un seuil dans l'équation de régression. Sur la base de nos constatations, ce seuil est fixé à 400 ms pour les tâches de conversation libre qui ont été utilisées dans nos tests subjectifs. Comme il a été vu dans le chapitre 1, l'effet du délai dépend de l'interactivité de la conversation, le seuil pourrait selon toute vraisemblance être adapté en fonction de cette interactivité. Cette question sera abordée dans le chapitre 5.

La relation entre la qualité de conversation et la qualité d'écoute, la qualité de locution et le délai est étudiée avec l'équation de régression suivante

$$\widehat{MOS}_{conversation} = \alpha \times MOS_{locution} + \beta \times MOS_{écoute} + \delta \times \max(0, délai - délai_{seuil}) + \gamma \quad (3.3)$$

où $\widehat{MOS}_{conversation}$ est la note de qualité de conversation estimée, $MOS_{locution}$ et $MOS_{écoute}$ sont les notes subjectives de qualité de locution et d'écoute, *délai* est le délai mesuré dans le système et $délai_{seuil}$ est le seuil au-delà duquel le délai devient gênant (ici pour une conversation libre, $délai_{seuil} = 400$ ms). Les coefficients α, β, δ et γ sont calculés afin de minimiser l'erreur quadratique moyenne (EQM) entre les notes de conversation subjectives et estimées.

A priori la relation dépend des dégradations présentes dans le système évalué. Si la communication est affectée uniquement par une dégradation de la qualité d'écoute, la qualité de conversation sera, en toute logique, majoritairement corrélée à la qualité d'écoute et peu à la qualité de locution. Pour cette raison, la relation est déterminée pour chacun des quatre tests indépendamment.

Pour chaque test, l'analyse de régression, telle qu'elle est décrite dans l'annexe D, est présentée dans un tableau, fournissant les valeurs des coefficients (Coeff), leurs variances (Var(Coeff)), la significativité de chaque régresseur (t-stat et $Pr > |t|$), la racine de l'erreur quadratique moyenne ($\bar{e}$) et les résultats du test de significativité (statistique F et p-value) du coefficient de détermination ajusté ($\bar{R}^2$) de la régression. Ensuite, une analyse de bootstrap (*cf.* annexe D) est effectuée pour chaque test afin de déterminer la distribution de chaque coefficient de régression α, β, δ et γ, et pour examiner la fiabilité de la régression. Étant donné N le nombre de sujets participant au test, à chaque itération un échantillon de N sujets est tiré aléatoirement avec remise. Pour chaque condition, les notes des sujets choisis aléatoirement sont moyennées pour obtenir une note MOS de conversation, une note MOS de locution et une note MOS d'écoute. Les coefficients α, β, δ et γ sont déterminés à partir de ces 3 notes et de la valeur de délai selon l'équation 3.3. 1000 itérations sont effectuées selon cette procédure pour obtenir la distribution de chaque coefficient.

3.3.1 Test 1 : délai et écho

La régression est appliquée aux huit notes d'écoute, de locution et de conversation récoltées lors du test subjectif pour les huit conditions de délai. Les résultats de la régression multiple sont donnés dans le tableau 3.39. La régression est significative ($F = 65.59, p < 0.05$).

Régression linéaire multiple				Stepwise regression			
Facteur	Coeff	Var(Coeff)	$P > \|t\|$	Facteur	Coeff	Var(Coeff)	$P > \|t\|$
Locution	0.468	0.002	0.00047	Locution	0.490	0.0017	0.00007
Écoute	0.483	0.197	0.3385	Délai	-1.551	0.1654	0.0124
Délai	-1.895	0.2598	0.02052	(Constante)	2.110	0.0239	0.00004
(Constante)	0.284	2.8482	0.8745				
$\bar{e} = 0.095$, $\bar{R}^2 = 0.96$, $F = 65.59$, $p = .0007$				$\bar{e} = 0.097$, $\bar{R}^2 = 0.96$, $F = 94.41$, $p = .0001$			

Tableau 3.39 : *Analyse de régression linéaire - Test 1 sur le délai et l'écho*

Cependant, le test de significativité effectué sur les coefficients de régression montre que le coefficient correspondant au régresseur Écoute (*i.e.* β) est non significativement différent de zéro ($p = 0.3385$). Ce phénomène reflète la colinéarité existant entre la note de qualité d'écoute, qui varie peu (*cf.* figure 3.2), et le terme constant γ. Une sélection des régresseurs, comme présentée dans l'annexe D, du type « stepwise regression » est donc appliquée. Celle-ci rejette le régresseur Écoute (*i.e.* $\beta = 0$). Les résultats correspondants sont fournis dans le tableau 3.39. La stepwise regression est fortement significative ($F = 94.41$, $p < 0.05$) et les tests de significativité pour le régresseur Locution, le régresseur Délai et le terme constant montrent qu'ils sont tous significativement non nuls ($p < 0.05$). La régression obtenue est logique, puisqu'en présence de dégradations affectant la qualité de locution et d'interaction ce sont les régresseurs correspondants qui permettent d'estimer la qualité de conversation.

L'analyse de bootstrap est effectuée sur les notes des quinze sujets participant au test. Les histogrammes des coefficients de régression et des performances correspondantes (coefficient de corrélation de Pearson r et erreur absolue moyenne EAM exprimée en MOS) sont donnés dans la figure 3.18. Les distributions des coefficients α, β et γ sont relativement fines et centrées sur les valeurs des coefficients obtenues avec les notes de l'ensemble des sujets (*cf.* tableau 3.39). La distribution de δ comprend deux maxima, un autour de la valeur du coefficient obtenue avec les notes de l'ensemble des sujets et un autre autour de zéro, ce qui signifie que le délai n'a pas le même impact pour tous les sujets. Les performances de la régression sont centrées autour de 0.9 pour le coefficient de corrélation et autour de 0.15 pour l'erreur absolue moyenne, montrant que quels que soient les sujets considérés, la régression est fiable et proche de celle obtenue avec l'ensemble des sujets.

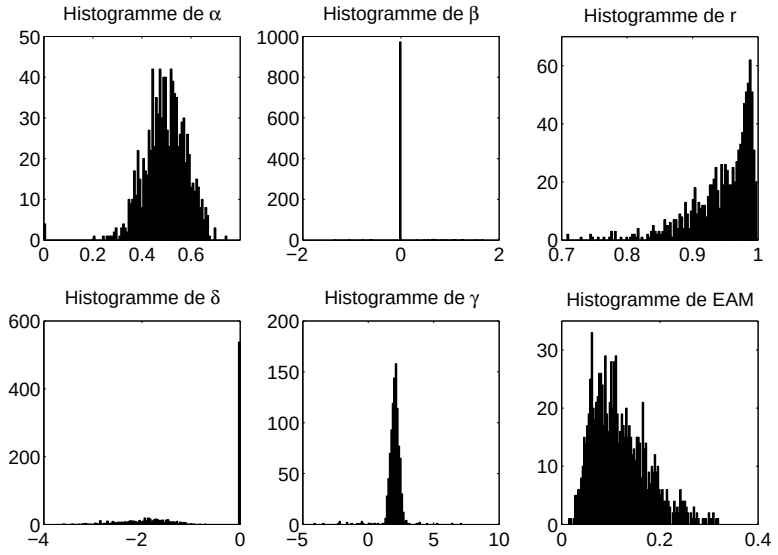
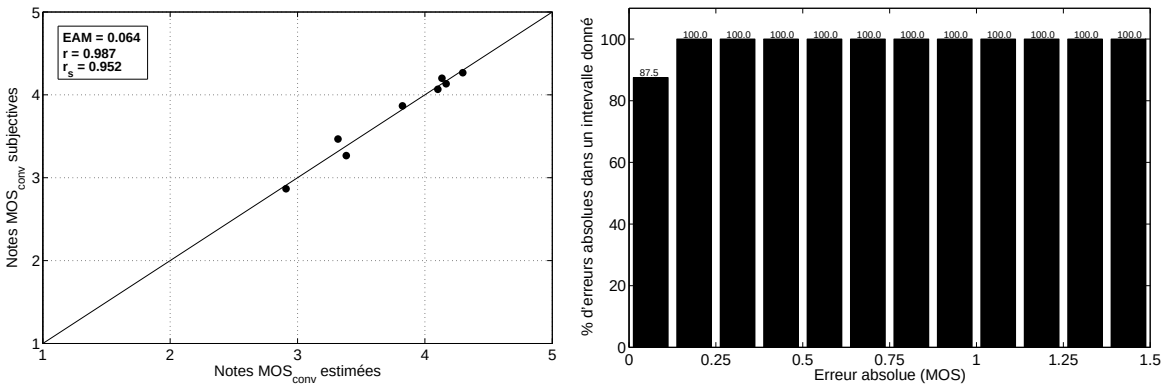


Figure 3.18 : *Résultat du bootstrap - Test 1 sur le délai et l'écho*

Les coefficients obtenus par stepwise regression (*cf.* tableau 3.39) sont appliqués aux notes MOS de l'ensemble des sujets d'après l'équation 3.3. La figure 3.19 présente les performances de l'estimation de la qualité de conversation. Le mapping entre les notes de conversation subjectives et estimées est fourni dans la figure 3.19(a) et la distribution cumulative de l'erreur



(a) Mapping entre notes de conversation subjectives et estimées

(b) Distribution cumulative de l'erreur absolue entre notes de conversation subjectives et estimées

Figure 3.19 : *Performances de l'estimation - Test 1 sur le délai et l'écho*

Régression linéaire multiple				**Stepwise regression**			
Facteur	Coeff	Var(Coeff)	$P > \|t\|$	Facteur	Coeff	Var(Coeff)	$P > \|t\|$
Locution	0.348	0.1190	0.35194	Écoute	0.602	0.0062	0.00012
Écoute	0.546	0.0093	0.0013	(Constante)	1.728	0.0632	0.00024
(Constante)	0.421	1.7400	0.76022				
$\bar{e} = 0.169$, $\bar{R}^2 = 0.86$, $F = 29.81$, $p = .0008$				$\bar{e} = 0.169$, $\bar{R}^2 = 0.86$, $F = 58.46$, $p = .0001$			

Tableau 3.40 : *Analyse de régression linéaire - Test 2 sur les pertes de paquets et le bruit*

absolue entre notes de conversation subjectives et estimées dans la figure 3.19(b). L'estimation par la régression linéaire aboutit à des corrélations de Pearson et de Spearman entre les notes de conversation subjectives et estimées très élevées ($r = 0.987$ et $r_s = 0.952$). La distribution cumulative de l'erreur absolue moyenne entre les notes de conversation subjectives et estimées montre que 100% des notes de conversation estimées diffèrent de moins de 0.25 MOS des notes subjectives.

3.3.2 Test 2 : pertes de paquets et bruit

La régression est appliquée aux neuf notes d'écoute, de locution et de conversation récoltées lors du test subjectif. Il n'y a pas de délai dans ce test, donc le coefficient δ est mis à zéro. Les résultats de la régression multiple sont donnés dans le tableau 3.40. La régression est significative ($F = 29.81$, $p < 0.05$). Le test de significativité effectué sur les coefficients de régression montre que le coefficient correspondant au régresseur Locution (*i.e.* α) est non significativement différent de zéro ($p = 0.3519$). Comme dans le premier test, il existe une colinéarité entre la note de qualité de locution, qui varie peu (*cf.* figure 3.7), et le terme constant γ. Une sélection des régresseurs par stepwise regression est appliquée : elle conserve le régresseur Écoute et rejette le régresseur Locution (*i.e.* $\alpha = 0$). Les résultats correspondants sont fournis dans le tableau 3.40. La stepwise regression est significative ($F = 58.46, p < 0.05$) et les tests de significativité pour le régresseur Écoute et le terme constant montrent qu'ils sont significativement non nuls ($p < 0.05$). Comme pour le premier test, la régression est significative et logique vis-à-vis des dégradations testées.

L'analyse de bootstrap est effectuée sur les notes des vingt participants au test. Les histogrammes des coefficients de régression et des performances correspondantes (coefficient de corrélation de Pearson r et erreur absolue moyenne EAM exprimée en MOS) sont donnés dans la figure 3.20. Les distributions des coefficients α, β et γ sont centrées sur les valeurs des coefficients obtenues avec les notes de l'ensemble des sujets (*cf.* tableau 3.40). Les performances de la régression sont centrées autour de 0.9 pour le coefficient de corrélation et autour de 0.15 pour l'erreur absolue moyenne. Ainsi quels que soient les sujets considérés, il semble que la

régression est fiable et proche de celle obtenue avec l'ensemble des sujets.

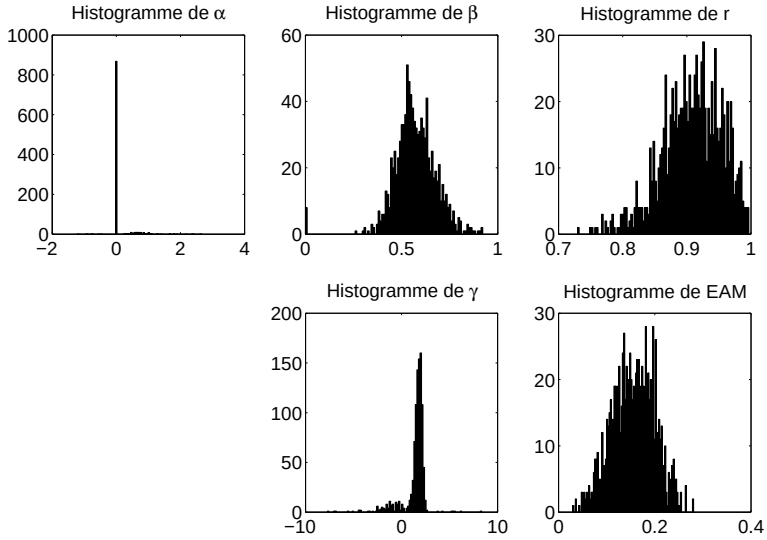
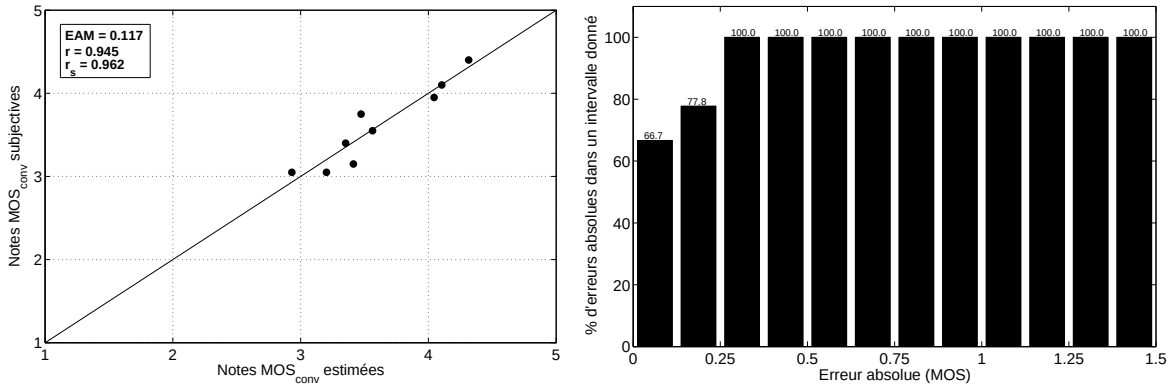


Figure 3.20 : *Résultat du bootstrap - Test 2 sur les pertes de paquets et le bruit*

Les coefficients obtenus par stepwise regression (*cf.* tableau 3.40) sont appliqués aux notes MOS de l'ensemble des sujets d'après l'équation 3.3. La figure 3.21 présente les performances de l'estimation de la qualité de conversation. Le mapping entre les notes de conversation subjectives et estimées est fourni dans la figure 3.21(a) et la distribution cumulative de l'erreur absolue entre notes de conversation subjectives et estimées dans la figure 3.21(b). L'estimation par la régression linéaire aboutit à des corrélations de Pearson et de Spearman entre les notes de conversation subjectives et estimées très élevées ($r = 0.945$ et $r_s = 0.962$). La distribution cumulative de l'erreur absolue moyenne entre les notes de conversation subjectives et estimées montre que 100% des notes de conversation estimées diffèrent de moins de 0.375 MOS des notes subjectives.

3.3.3 Test 3 : bruit

La régression est appliquée aux sept notes d'écoute, de locution et de conversation récoltées lors du test subjectif. Il n'y a pas de délai dans ce test, donc le coefficient δ est mis à zéro. Les résultats de la régression multiple sont donnés dans le tableau 3.41. La régression est significative ($F = 83.03$, $p < 0.05$). Comme pour le test précédent, le test de significativité effectué sur les coefficients de régression montre que le coefficient correspondant au régresseur Locution



(a) Mapping entre notes de conversation subjectives et estimées

(b) Distribution cumulative de l'erreur absolue entre notes de conversation subjectives et estimées

Figure 3.21 : *Performances de l'estimation - Test 2 sur les pertes de paquets et le bruit*

Régression linéaire multiple				Stepwise regression			
Facteur	Coeff	Var(Coeff)	$P > \|t\|$	Facteur	Coeff	Var(Coeff)	$P > \|t\|$
Locution	0.053	0.0477	0.82086	Écoute	0.865	0.0036	0.00003
Écoute	0.813	0.0502	0.02216	(Constante)	0.367	0.0339	0.10271
(Constante)	0.369	0.0419	0.14553				
$\bar{e} = 0.131$, $\bar{R}^2 = 0.959$, $F = 83.03$, $p = .0006$				$\bar{e} = 0.118$, $\bar{R}^2 = 0.967$, $F = 204.5$, $p = 3e^{-5}$			

Tableau 3.41 : *Analyse de régression linéaire - Test 3 sur le bruit*

(*i.e.* α) est non significativement différent de zéro ($p = 0.8208$). Il existe une colinéarité entre les notes de qualité de locution et de qualité d'écoute, qui ont des variations identiques (*cf.* figure 3.11). Les régresseurs sont sélectionnés par stepwise regression, qui conserve le régresseur Écoute et rejette le régresseur Locution (*i.e.* $\alpha = 0$). Les résultats correspondants sont fournis dans le tableau 3.41. La stepwise regression est très significative ($F = 204.5, p < 0.05$). Le test de significativité pour le régresseur Écoute montre qu'il est significativement non nul ($p < 0.05$), ce qui n'est pas le cas pour le terme constant. Par rapport aux tests précédents, les qualités de locution, d'écoute et de conversation sont toutes affectées par le bruit et sont très corrélées. La régression, après sélection des régresseurs, conserve seulement la note d'écoute pour prédire la note de conversation.

L'analyse de bootstrap est effectuée sur les notes des treize participants au test. Les histogrammes des coefficients de régression et des performances correspondantes (coefficient de corrélation de Pearson r et erreur absolue moyenne EAM exprimée en MOS) sont donnés dans la figure 3.22. Les distributions des coefficients α et β dépendent des sujets considérés. En effet, les notes de qualité de locution et d'écoute sont corrélées, la stepwise regression sélectionnera donc l'une ou l'autre en fonction des sujets. La distribution de γ est centrée sur la valeur obtenue avec les notes de l'ensemble des sujets (*cf.* tableau 3.41). Les performances de la régression sont centrées autour de 0.95 pour le coefficient de corrélation et autour de 0.12 pour l'erreur absolue moyenne.

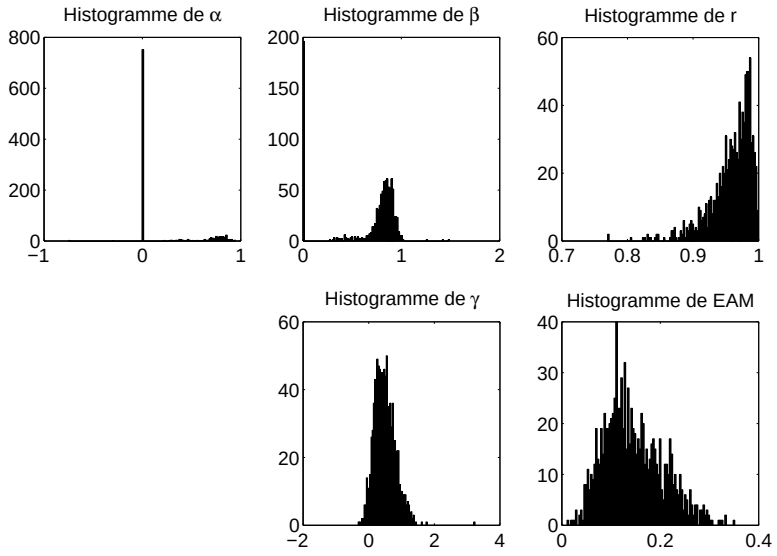
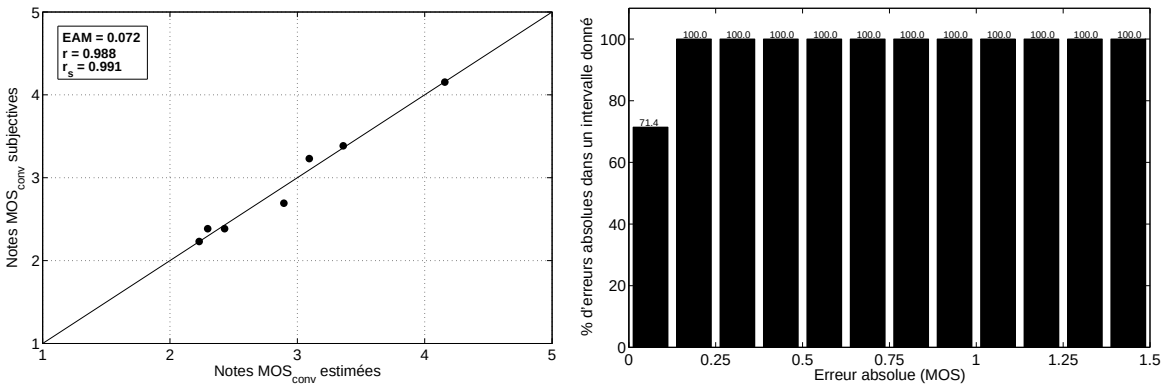


Figure 3.22 : *Résultat du bootstrap - Test 3 sur le bruit*

Les coefficients obtenus par stepwise regression (*cf.* tableau 3.41) sont appliqués aux notes MOS de l'ensemble des sujets d'après l'équation 3.3. La figure 3.23 présente les performances de l'estimation de la qualité de conversation. Le mapping entre les notes de conversation subjectives et estimées est fourni dans la figure 3.23(a) et la distribution cumulative de l'erreur absolue entre notes de conversation subjectives et estimées dans la figure 3.23(b). L'estimation par la régression linéaire aboutit à des corrélations de Pearson et de Spearman entre les notes de conversation subjectives et estimées très élevées ($r = 0.988$ et $r_s = 0.991$). La distribution



(a) Mapping entre notes de conversation subjectives et estimées

(b) Distribution cumulative de l'erreur absolue entre notes de conversation subjectives et estimées

Figure 3.23 : *Performances de l'estimation - Test 3 sur le bruit*

cumulative de l'erreur absolue moyenne entre les notes de conversation subjectives et estimées montre que 100% des notes de conversation estimées diffèrent de moins de 0.25 MOS des notes subjectives.

3.3.4 Test 4 : écho, délai et pertes de paquets

Ce test reprend des conditions déjà testées lors des tests 1 et 2 (conditions avec délai et écho, conditions avec pertes de paquets). Ces conditions vont permettre de vérifier la validité des relations déterminées dans les tests 1 et 2. Les coefficients de régression du test 1 ($\alpha = 0.490$, $\beta = 0$, $\delta = -1.551$ et $\gamma = 2.110$) sont appliqués aux notes subjectives de locution et d'écoute des conditions avec délai et écho (conditions 4, 7, 10, 13, 16 et 19). Le mapping entre les notes de conversation subjectives et estimées ainsi obtenues est donné dans la figure 3.24. Les coefficients déterminés lors du premier test permettent de très bien estimer les notes de conversation du test 4 en présence de délai et d'écho. Les coefficients de régression du test 2 ($\alpha = 0$, $\beta = 0.602$, $\delta = 0$ et $\gamma = 1.728$) sont appliqués aux notes subjectives de locution et d'écoute des conditions avec pertes de paquets seules (conditions 1, 2 et 3). Le mapping entre les notes de conversation subjectives et estimées ainsi obtenues est donné dans la figure 3.25. L'estimation des notes de conversation du test 4 en présence de perte de paquets est excellente en utilisant les coefficients de régression déterminés lors du test 2 dans les mêmes conditions. Ces deux résultats démontrent une certaine reproductibilité dans les résultats de tests subjectifs et une fiabilité dans les coefficients de régression déterminés lors des tests 1 et 2.

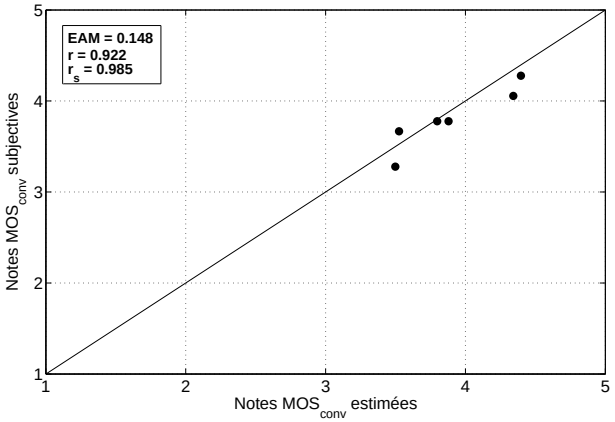


Figure 3.24 : *Mapping entre notes de conversation subjectives et estimées - Conformité avec le test 1 pour les conditions avec délai et écho - Test 4 sur l'écho, le délai et les pertes de paquets*

Régression linéaire multiple

Facteur	Coeff	Var(Coeff)	$P > \|t\|$
Locution	0.367	0.0024	0.00004
Écoute	0.321	0.0055	0.00191
(Constante)	0.915	0.0666	0.00624

$\bar{e} = 0.109$, $\bar{R}^2 = 0.868$, $F = 40.89$, $p = 3e^{-5}$

Tableau 3.42 : *Analyse de régression linéaire - Test 4 sur l'écho, le délai et les pertes de paquets*

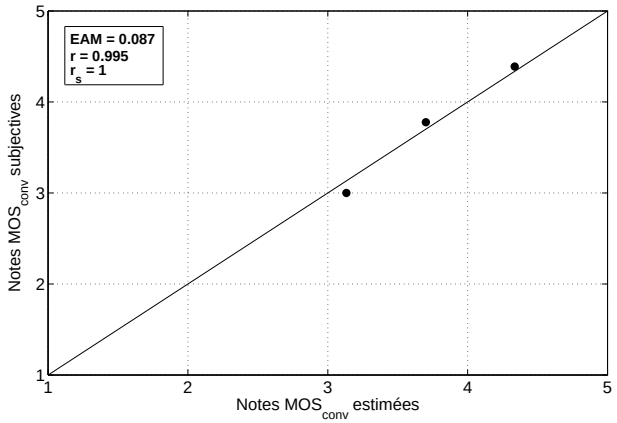


Figure 3.25 : *Notes de conversation subjectives et estimées - Conformité avec le test 2 pour les conditions avec pertes de paquets seules - Test 4 sur l'écho, le délai et les pertes de paquets*

La régression est appliquée aux douze notes d'écoute, de locution et de conversation restantes (conditions 5, 6, 8, 9, 11, 12, 14, 15, 17, 18, 20 et 21) en présence simultanée de délai, d'écho et de pertes de paquets. Il n'y a pas de délai supérieur au seuil de 400 ms dans ce test, donc le coefficient δ est mis à zéro. Les résultats de la régression multiple sont donnés dans le tableau 3.42. La régression est significative ($F = 40.89$, $p < 0.05$). Le test de significativité effectué sur les coefficients de régression montre que tous les coefficients de la régression sont significativement différents de zéro. Dans ce test, à la fois les qualités de locution, d'écoute et de conversation sont affectées par les dégradations présentes, ce qui se retrouve dans les valeurs des coefficients de régression.

L'analyse de bootstrap est effectuée, pour ces douze conditions, sur les notes des vingt participants au test. Les histogrammes des coefficients de régression et des performances correspondantes (coefficient de corrélation de Pearson r et erreur absolue moyenne EAM exprimée en MOS) sont donnés dans la figure 3.26. Les distributions des coefficients α et β semblent dépendre des sujets et de la façon dont ils ont été gênés par les pertes de paquets. En effet, la distribution du coefficient β présente deux maxima, un autour de zéro et l'autre autour d'une valeur proche de celle déterminée sur l'ensemble des sujets. La distribution du coefficient α présente aussi deux maxima, le plus élevé étant celui autour de la valeur proche de celle déterminée sur l'ensemble des sujets. Les performances de la régression sont centrées autour de 0.85 pour le coefficient de corrélation et autour de 0.15 pour l'erreur absolue moyenne.

Les coefficients obtenus par régression linéaire multiple (*cf.* tableau 3.42) sont appliqués aux notes MOS de l'ensemble des sujets d'après l'équation 3.3. La figure 3.27 présente les performances de l'estimation. Le mapping entre les notes de conversation subjectives et estimées est fourni dans la figure 3.27(a) et la distribution cumulative de l'erreur absolue entre notes de conversation subjectives et estimées dans la figure 3.27(b). L'estimation par la régression linéaire aboutit à des corrélations de Pearson et de Spearman entre les notes de conversation subjectives et estimées très élevées ($r = 0.949$ et $r_s = 0.967$). La distribution cumulative de l'erreur absolue moyenne entre les notes de conversation subjectives et estimées montre que 100% des notes de conversation estimées diffèrent de moins de 0.25 MOS des notes subjectives.

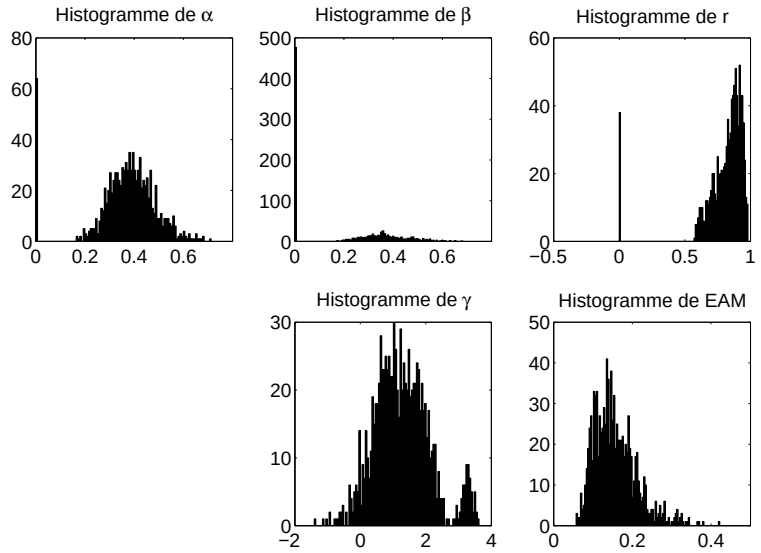
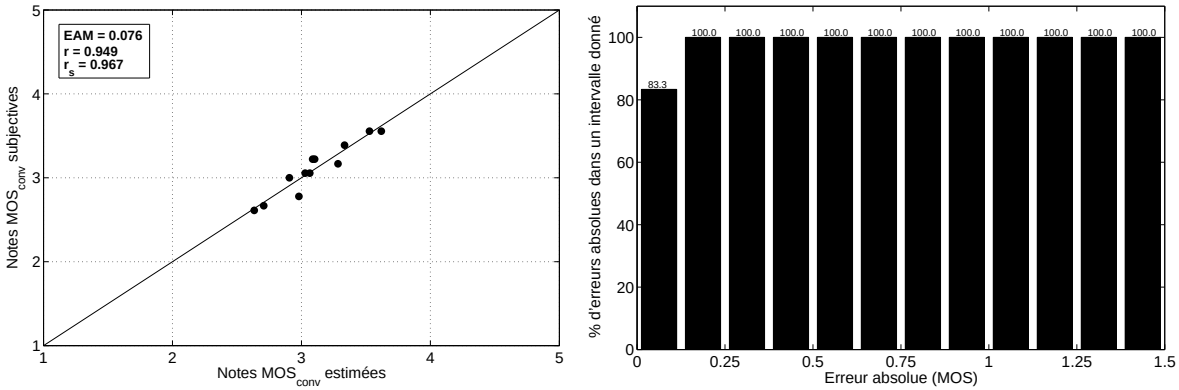


Figure 3.26 : *Résultat du bootstrap - Test 4 sur l'écho, le délai et les pertes de paquets*



(a) Mapping entre notes de conversation subjectives et estimées

(b) Distribution cumulative de l'erreur absolue entre notes de conversation subjectives et estimées

Figure 3.27 : *Performances de l'estimation - Test 4 sur l'écho, le délai et les pertes de paquets*

Coefficient	Test 1	Test 2	Test 3	Test 4
α	0.490	0	0	0.367
β	0	0.602	0.865	0.321
δ	-1.551	0	0	0
γ	2.110	1.728	0.367	0.915

Tableau 3.43 : *Valeurs des coefficients de régression pour les différents tests*

3.3.5 Détection des dégradations

À chaque type de dégradation étudiée lors de ces quatre tests subjectifs correspond donc un jeu de coefficients de régression, récapitulés dans le tableau 3.43. L'application du modèle à une communication implique donc nécessairement de détecter les dégradations présentes pour choisir le jeu de coefficients correspondant dans la base de relations F_i créée pendant la construction du modèle.

Les dégradations à détecter, d'après les quatre tests subjectifs présentés, sont le délai, l'écho, les pertes de paquets et le bruit. Le délai peut être mesuré par différentes méthodes intrusives ou non intrusives, présentées dans le chapitre 2. L'écho peut être détecté de façon intrusive ou non intrusive en comparant les signaux émis et reçu du même côté du système. Des méthodes de corrélation telles que celle brevetée par Barriac et Gilloire dans [Barriac et Gilloire 1996] pourront, par exemple, être utilisées. La détection des pertes de paquets

Régression linéaire multiple

Facteur	Coeff	Var(Coeff)	$P > \|t\|$
Locution	0.4059	0.00137	0
Écoute	0.5519	0.00199	0
Délai	-1.7376	0.32943	0.00665
(Constante)	0.1710	0.02543	0.2964

$\bar{e} = 0.144$, $\bar{R}^2 = 0.948$, $F = 148.2$, $p = .0$

Tableau 3.44 : *Analyse de régression linéaire sur la base d'apprentissage - Tests 1, 2 et 3*

peut être effectuée par différentes méthodes, par exemple grâce à des sondes INMD pour une mesure non intrusive paramétrique [UIT-T Rec. P.561 2002] ou grâce à une analyse du signal dégradé telle que celle proposée dans la Recommandation P.563 par détection des silences dans les trames de parole active [UIT-T Rec. P.563 2004]. La détection de présence de bruit peut être faite par différentes méthodes, par exemple grâce à des sondes INMD pour une mesure non intrusive paramétrique ou grâce au calcul du rapport signal-à-bruit des signaux de la communication.

Cette détection des dégradations pilote ensuite un système de décision, qui, en fonction des dégradations détectées, choisit la relation F_i à appliquer. Une mauvaise détection entraîne un choix erroné des coefficients de régression à appliquer et peut aboutir à une évaluation médiocre de la qualité de conversation. Il semble donc intéressant d'explorer la possibilité de s'affranchir d'une détection des dégradations. Dans cette optique, nous avons déterminé un jeu de coefficients en considérant l'ensemble des conditions explorées lors de ces quatre tests.

3.3.6 Tous tests

La régression est apprise sur l'ensemble des conditions des trois premiers tests confondues (*i.e.* vingt-quatre conditions), puis appliquée à l'ensemble des conditions du quatrième test (*i.e.* vingt-et-une conditions). La base d'apprentissage regroupe les dégradations : délai seul, écho et délai, pertes de paquets seules, pertes de paquets et bruit transmis, et bruit seul. La base de validation reprend certaines dégradations de la base d'apprentissage (à des niveaux identiques ou différents) : délai seul, écho et délai, et pertes de paquets seules, et teste de nouvelles combinaisons de dégradations : délai et pertes de paquets, et écho et délai et pertes de paquets.

3.3.6.1 Apprentissage

Les résultats de la régression multiple sont donnés dans le tableau 3.44. La régression est significative ($F = 148.2$, $p < 0.05$). Le test de significativité effectué sur les coefficients de régression montre que tous les coefficients de la régression sont significativement différents de zéro. Le coefficient de détermination ajusté de la régression est élevé ($\bar{R}^2 = 0.948$), indiquant que la régression est efficace. Les coefficients déterminés ($\alpha = 0.4059$, $\beta = 0.5519$, $\delta = -1.7376$, $\gamma = 0.1710$) ont des signes logiques. α et β sont positifs, indiquant que lorsque les notes de locution et d'écoute augmentent, la note de conversation augmente. δ est négatif, traduisant la diminution de la note de conversation quand le délai augmente.

Pour les conditions considérées, un seul jeu de coefficients de régression semble suffisant pour estimer la qualité de conversation à partir des qualités de locution et d'écoute, et de la valeur de délai, permettant ainsi de s'affranchir de la détection des dégradations. Bien que non attendu *a priori*, ce résultat est de nouveau analogue à ce qui est obtenu avec la qualité audiovisuelle. Dans les différentes études portant sur l'estimation de la qualité audiovisuelle à partir des qualités audio et vidéo [Jones et Atkinson 1998, Pastrana-Vidal *et al.* 2003, Kitawaki *et al.* 2005], un même jeu de coefficients est utilisé pour estimer la qualité dans différentes conditions de dégradation (codecs audio, codecs vidéo, délais audio).

L'analyse de bootstrap est effectuée sur l'ensemble des notes individuelles des trois tests

d'apprentissage. Les histogrammes des coefficients de régression et des performances correspondantes (coefficient de corrélation de Pearson r et erreur absolue moyenne EAM exprimée en MOS) sont donnés dans la figure 3.28. Les distributions des coefficients α, β et γ sont centrées autour des valeurs déterminées sur l'ensemble des sujets. La distribution du coefficient δ présente un maximum à 0 et des valeurs autour de la valeur déterminée sur l'ensemble des sujets, indiquant que le délai n'a pas la même influence selon les sujets considérés. Les performances de la régression sont centrées autour de 0.95 pour le coefficient de corrélation et autour de 0.16 pour l'erreur absolue moyenne.

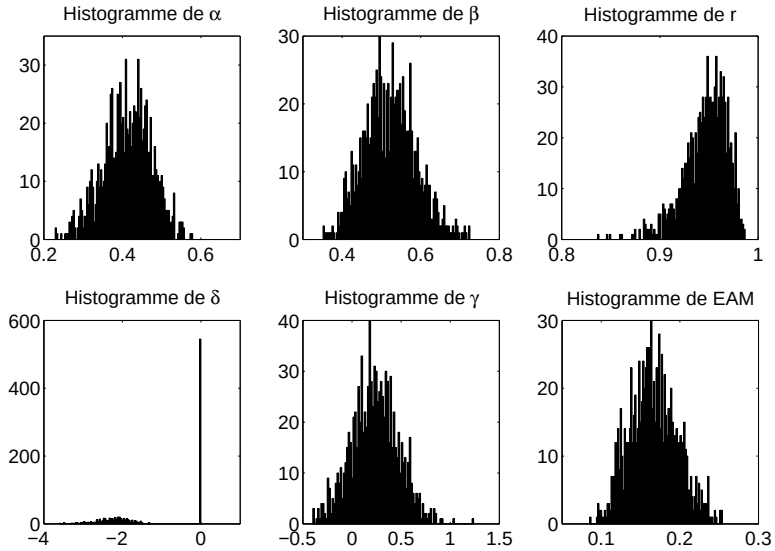


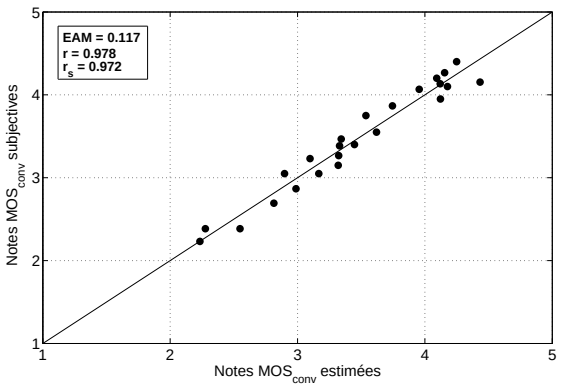
Figure 3.28 : *Résultat du bootstrap sur la base d'apprentissage - Tests 1, 2 et 3*

Les coefficients obtenus par régression linéaire multiple (*cf.* tableau 3.44) sont appliqués aux notes MOS de l'ensemble des sujets de la base d'apprentissage d'après l'équation 3.3. La figure 3.29 présente les performances de l'estimation. Le mapping entre les notes de conversation subjectives et estimées est fourni dans la figure 3.29(a) et la distribution cumulative de l'erreur absolue entre notes de conversation subjectives et estimées dans la figure 3.29(b). L'estimation par la régression linéaire aboutit à des corrélations de Pearson et de Spearman entre les notes de conversation subjectives et estimées très élevées ($r = 0.978$ et $r_s = 0.972$). La distribution cumulative de l'erreur absolue moyenne entre les notes de conversation subjectives et estimées montre que 100% des notes de conversation estimées diffèrent de moins de 0.375 MOS des notes subjectives.

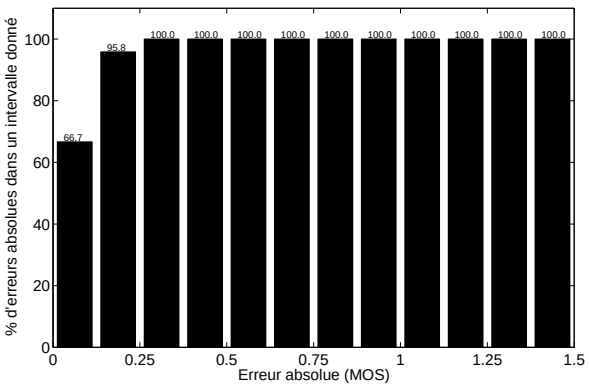
3.3.6.2 Validation

Validation sur le test 4 Les coefficients de régression déterminés à partir des notes subjectives de la base d'apprentissage sont appliqués aux notes subjectives de la base de validation, *i.e.* aux notes MOS du test 4. La figure 3.30 présente les performances de l'estimation. Le mapping entre les notes de conversation subjectives et estimées de la base de validation est fourni dans la figure 3.30(a) et la distribution cumulative de l'erreur absolue entre notes de conversation subjectives et estimées dans la figure 3.30(b). L'estimation aboutit à des corrélations de Pearson et de Spearman entre les notes de conversation subjectives et estimées très élevées ($r = 0.969$ et $r_s = 0.951$). La distribution cumulative de l'erreur absolue moyenne entre les notes de conversation subjectives et estimées montre que 100% des notes de conversation estimées diffèrent de moins de 0.5 MOS des notes subjectives.

Validation sur un test extérieur à la thèse Les coefficients de régression déterminés sur les notes subjectives de la base d'apprentissage sont appliqués aux notes subjectives récoltées

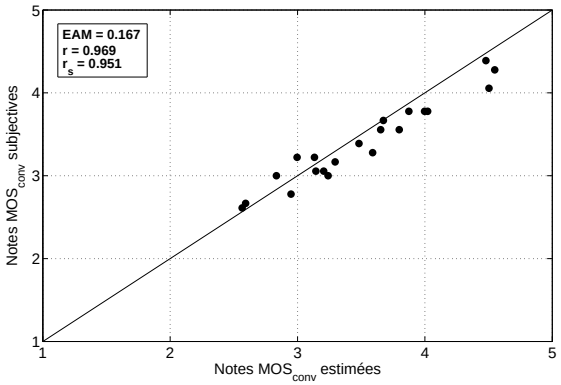


(a) Mapping entre notes de conversation subjectives et estimées

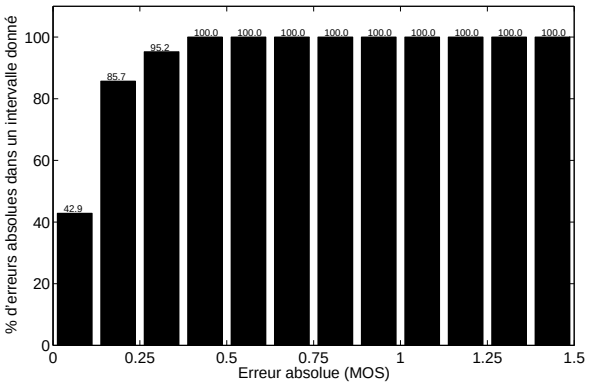


(b) Distribution cumulative de l'erreur absolue entre notes de conversation subjectives et estimées

Figure 3.29 : *Performances de l'estimation sur la base d'apprentissage - Tests 1, 2 et 3*



(a) Mapping entre notes de conversation subjectives et estimées



(b) Distribution cumulative de l'erreur absolue entre notes de conversation subjectives et estimées

Figure 3.30 : *Performances de l'estimation sur la base de validation - Test 4*

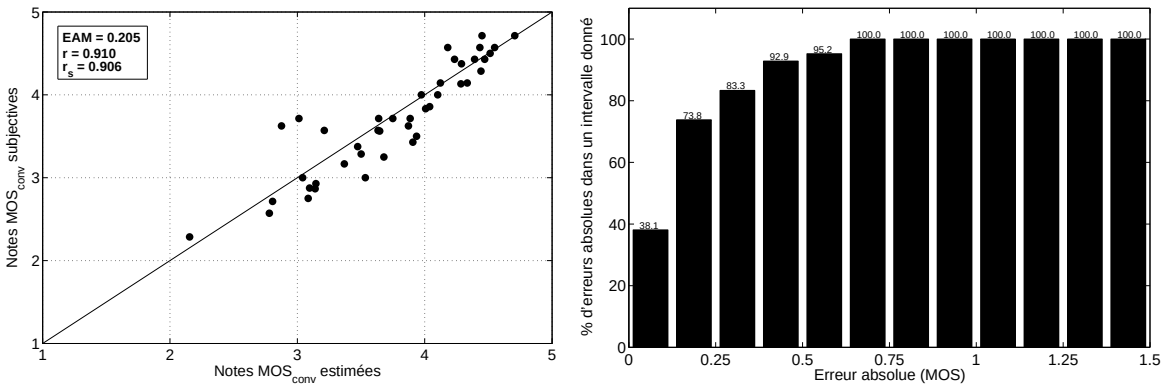
Réseau	RTC (codec G.711)	VoIP (codec GSM 6.10)
Bruit de fond	sans ou 56 dB SPL (Hoth)	sans ou 56 dB SPL (Hoth)
Délai unidirectionnel (ms)	0, 150, 300, 500, 600	150, 300, 500, 600, 900
Atténuation de l'écho (dB)	5, 15, >60	5, 15, >60

Tableau 3.45 : *Conditions du test extérieur à la thèse effectué par MESAQIN*

lors d'un test effectué en dehors du cadre de la thèse par le laboratoire MESAQIN de l'Université de Prague et présenté dans [Holub 2006]. Ce test, dont les conditions sont fournies dans le tableau 3.45, étudie 42 conditions de dégradations (type de réseau/codec, bruit, délai, atténuation de l'écho), selon la méthodologie suivante :

- un participant non expert communique avec un interlocuteur expérimenté,
- 16 participants non experts, chacun testant 11 conditions sur les 42,
- 4 scénarios de conversation [Möller 1997a],
- 2 à 3 minutes de conversation (en langue tchèque),
- le participant non expert juge les qualités d'écoute, de locution, d'interaction et de conversation à la fin de chaque conversation.

Les conditions testées ici permettent d'étudier certaines dégradations déjà testées dans nos propres tests (à des niveaux identiques ou différents), et d'autres qui n'ont pas été testées (réseau/codec différent). La figure 3.31 présente les performances de l'estimation sur ce test, sous la forme d'un mapping entre les notes de conversation subjectives et estimées (*cf.*



(a) Mapping entre notes de conversation subjectives et estimées

(b) Distribution cumulative de l'erreur absolue entre notes de conversation subjectives et estimées

Figure 3.31 : *Performances de l'estimation sur le test extérieur à la thèse effectué par MESAQIN*

Critère de performance	RTC et VoIP	RTC	RTC sans bruit	RTC avec bruit	VoIP	VoIP sans bruit	VoIP avec bruit
r	0.910	0.952	0.954	0.948	0.831	0.875	0.751
r_s	0.906	0.947	0.884	0.970	0.665	0.700	0.588
EAM	0.205	0.154	0.159	0.149	0.272	0.266	0.279

Tableau 3.46 : *Performances de l'estimation sur le test extérieur à la thèse effectué par MESAQIN, en fonction des dégradations*

figure 3.31(a)) et de la distribution cumulative de l'erreur absolue entre notes de conversation subjectives et estimées (*cf.* figure 3.31(b)). L'estimation aboutit à des corrélations de Pearson et de Spearman entre les notes de conversation subjectives et estimées élevées ($r = 0.910$ et $r_s = 0.906$). La distribution cumulative de l'erreur absolue moyenne entre les notes de conversation subjectives et estimées montre que 100% des notes de conversation estimées diffèrent de moins de 0.75 MOS des notes subjectives. Les performances de l'estimation, fournies dans le tableau 3.46, peuvent être analysées en fonction des dégradations testées (réseau, bruit). Elles sont plus élevées dans le cas du réseau RTC que dans le cas du réseau VoIP. Ceci s'explique essentiellement par la valeur de délai unidirectionnel de 900 ms testée en VoIP, puisque notre modèle a été entraîné sur des valeurs de délai unidirectionnel inférieures ou égales à 600 ms. Si les conditions avec un délai égal à 900 ms ne sont pas prises en compte, des performances plus élevées pour le VoIP sont obtenues ($r = 0.940$ et $EAM = 0.216$ MOS). Ces résultats montrent que l'estimation fonctionne correctement pour des dégradations déjà testées lors de nos propres tests (avec des niveaux identiques ou différents), dans une langue étrangère et pour un codec différent de celui utilisé pour l'apprentissage du modèle.

Conclusion

Dans ce chapitre, nous avons présenté quatre tests subjectifs, mis en œuvre d'après la nouvelle méthodologie de test proposée dans le cadre de la thèse, dans différentes conditions de dégradation. À partir des notes subjectives d'écoute et de locution recueillies lors de ces tests et de la valeur du délai, nous avons déterminé une combinaison linéaire de ces trois composantes permettant d'estimer la note subjective de conversation pour chacune des dégradations testées. L'ensemble F des relations ainsi obtenues aboutit à d'excellentes estimations de la qualité de conversation, dans toutes les conditions de dégradation testées. Cependant, une relation F_i par type de dégradation nécessite de détecter les dégradations présentes dans la communication testée pour choisir la relation F_i correspondante : une mauvaise détection peut entraîner une évaluation erronée de la qualité de conversation.

Dans l'optique de s'affranchir de la détection des dégradations, nous avons déterminé un unique jeu de coefficients de régression linéaire multiple ($\alpha = 0.4059$, $\beta = 0.5519$, $\delta = -1.7376$, $\gamma = 0.1710$), à partir d'une base d'apprentissage composée des tests 1, 2 et 3 explorant plusieurs types de dégradation. Le jeu de coefficients ainsi déterminé aboutit à une excellente estimation de la qualité de conversation ($r = 0.978$, $r_s = 0.972$ et $EAM = 0.117$ MOS). Il a ensuite été appliqué à une base de validation (test 4 et test extérieur à la thèse), étudiant des dégradations communes à la base d'apprentissage et de nouvelles dégradations. La régression déterminée sur la base d'apprentissage est confirmée puisqu'elle atteint des performances élevées ($r = 0.969$, $r_s = 0.951$ et $EAM = 0.167$ MOS) sur le test 4 et sur le test extérieur à la thèse ($r = 0.910$, $r_s = 0.906$ et $EAM = 0.205$ MOS). Le modèle fonctionne donc sur l'ensemble des conditions étudiées avec une équation de régression unique, permettant de s'affranchir de la détection des dégradations.

Le chapitre suivant présente plusieurs outils nécessaires à la mise en œuvre du modèle objectif de la qualité de conversation, en se basant sur les valeurs de coefficients de régression déterminées ici.

Chapitre 4

Outils de mesure

Introduction

Dans ce chapitre, nous présenterons les outils nécessaires à la partie mesure du modèle objectif proposé. La partie mesure est composée :
- d'un modèle objectif de la qualité d'écoute,
- d'un modèle objectif de la qualité de locution,
- d'un outil de mesure du délai.

Les modèles objectifs de la qualité d'écoute disponibles à l'heure actuelle et normalisés à l'UIT-T sont ceux décrits dans les Recommandations P.862 (PESQ) [UIT-T Rec. P.862 2001], P.563 [UIT-T Rec. P.563 2004] et P.564 [UIT-T Rec. P.564 2006]. Ces recommandations fournissent une description détaillée de leurs fonctionnements, ainsi que les codes sources des modèles, optimisés sur plusieurs bases de données. Ces trois modèles peuvent être utilisés tels quels dans la partie mesure du modèle que nous proposons. En revanche, il n'existe qu'un seul modèle objectif de la qualité de locution : le modèle PESQM [Appel et Beerends 2002]. Il n'est pas normalisé, donc son code source et les bases de données sur lesquelles il a été optimisé ne sont pas disponibles. La première partie de ce chapitre sera ainsi consacrée à l'optimisation de notre version de PESQM, en s'appuyant sur les résultats d'un test de locution effectué à cet effet.

Le délai peut être mesuré de différentes façons selon l'application visée :
- en mesure intrusive basée sur les signaux : grâce au modèle PESQ, qui permet la mesure du retard entre ses signaux de référence et dégradé, si leurs enregistrements sont synchrones,
- en mesure intrusive ou non intrusive basée sur les signaux : grâce à la mesure du délai de l'écho, s'il est présent,
- en mesure non intrusive pour les communications IP : grâce aux informations contenues dans la trame RTCP ou RTCP-XR sur les délais des paquets transmis,
- en mesure non intrusive : grâce à deux sondes synchronisées (*e.g.* par GPS) placées chacune d'un côté de la communication.

Dans la thèse, le délai sera supposé connu, puisque les enregistrements effectués lors des quatre tests subjectifs n'étaient pas systématiquement synchrones.

Le modèle objectif de conversation développé dans la thèse peut fonctionner avec deux types de signaux :
- de test pour une mesure intrusive : un signal de référence est envoyé dans le système et le signal dégradé est recueilli en sortie du système.
- réels, enregistrés lors de communications réelles ou lors de tests subjectifs.

En mesure intrusive, le premier type de signal peut être utilisé directement avec le modèle objectif proposé pour évaluer la qualité vocale de conversation de la communication entre deux points A et B d'un système, telle qu'elle est perçue au point A, selon le scénario suivant :

1. Envoi d'un signal de référence depuis le point B, recueil du signal dégradé au point A : mesure PESQ de la qualité d'écoute.
2. Envoi d'un signal de référence depuis le point A, recueil du signal dégradé au point A : mesure PESQM de la qualité de locution.

Le signal de référence utilisé pour évaluer la qualité d'écoute respecte alors les contraintes suivantes (définies dans la Recommandation [UIT-T Rec. P.862.3 2005]) pour fonctionner avec le modèle PESQ :

- une longueur entre 8 et 30 secondes,
- un taux d'activité vocale entre 40 et 80% (mesuré d'après la Recommandation P.56 [UIT-T Rec. P.56 1993]),
- un niveau d'activité vocale de -30 dBov.

En mesure non intrusive, le second type de signal peut être utilisé directement avec le modèle objectif proposé pour évaluer la qualité vocale de conversation de la communication entre deux points A et B d'un système, telle qu'elle est perçue au point A, à condition de se conformer aux contraintes posées par le modèle P.563 [UIT-T Rec. P.563 2004] :

- une longueur entre 3 et 20 secondes,
- un taux d'activité vocale entre 25 et 75% [UIT-T Rec. P.56 1993],
- un niveau d'activité vocale entre -36 et -16 dBov.

Dans la seconde partie de ce chapitre, nous présenterons un outil de traitement des signaux qui permet d'utiliser les signaux du second type (réels) avec les modèles intrusifs, tels que PESQ et PESQM. Ainsi les signaux de conversation enregistrés pendant les tests subjectifs de la thèse pourront être exploités. De plus, cet outil permettra d'utiliser les signaux de conversation provenant de tests subjectifs extérieurs.

4.1 Optimisation du modèle de qualité de locution PESQM

Le modèle PESQM, décrit dans l'annexe C, a été reprogrammé d'après les informations fournies dans l'article [Appel et Beerends 2002]. La base de données utilisée par Appel et Beerends pour développer PESQM n'étant pas disponible, nous avons dû récolter d'autres données pour optimiser et valider notre version de PESQM. Nous présentons dans cette partie une étude préliminaire de PESQM sur deux tests de locution disponibles dans la littérature puis son optimisation sur un test de locution mis en œuvre dans le cadre de la thèse.

4.1.1 Étude préliminaire de PESQM sur deux tests de locution de la littérature

Il existe très peu de tests subjectifs de locution dans la littérature. Les deux tests que nous avons recensés sont :

- un test présenté dans la Contribution UIT-T COM 12-16 [Gierlich et Diedrich 2000], qui étudie 20 conditions d'écho seul (*cf.* tableau 4.1) avec des combinés mains-libres et 12 sujets non experts,
- le test présenté dans l'article [Appel et Beerends 2002], qui étudie 40 conditions d'écho seul (*cf.* tableau 4.2) avec 8 sujets non experts.

Les signaux enregistrés pendant ces tests n'étant pas disponibles, la version électrique du modèle PESQM (*cf.* annexe C) a été appliquée à 28 fichiers de référence (format wav, 4 locuteurs, 7 doubles-phrases par locuteur) et aux 28 fichiers dégradés correspondants, générés d'après les conditions des deux tests. La moyenne des scores PESQM obtenus pour chaque condition est calculée, permettant ainsi de minimiser l'effet de la phrase prononcée (locuteur,

Délai (ms)	TELR (dB)	Note MOS
100	99	4.94
25	33	3.65
50	40	4.24
100	47	4.06
200	52	4.65
25	27	3.50
50	34	3.35
100	41	3.82
200	46	4.06
25	23	2.65
50	30	3.18
100	37	2.82
200	42	3.65
25	13	1.50
50	20	1.76
100	27	2.00
200	32	2.68
50	24	2.32
100	24	1.68
200	24	1.76

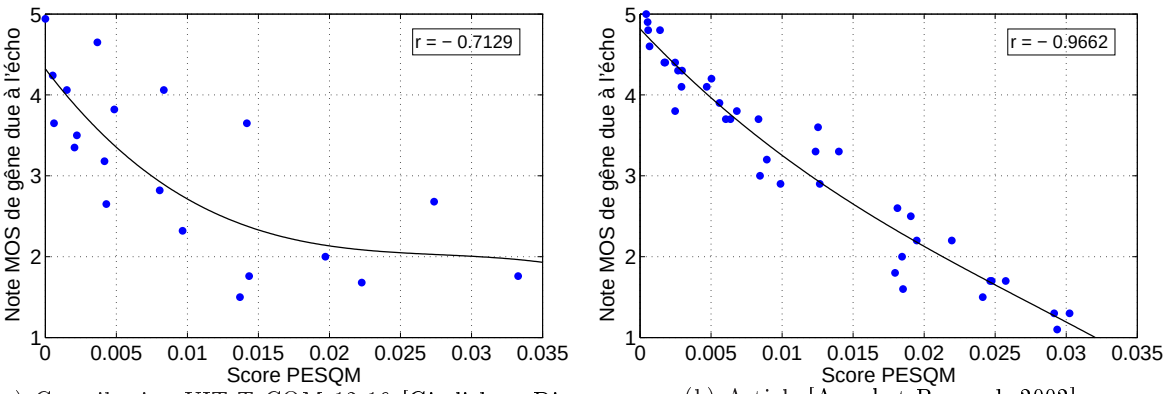
Tableau 4.1 : *Conditions et notes MOS du test de locution avec écho seul - Contribution UIT-T COM 12-16 [Gierlich et Diedrich 2000]*

Délai (ms)	TELR (dB)	Note MOS	Délai (ms)	TELR (dB)	Note MOS
6	19	4.80	86	25	2.00
12	19	4.30	120	25	1.70
20	19	3.70	156	25	1.30
28	19	3.00	30	31	4.80
46	19	3.30	46	31	4.10
60	19	1.60	64	31	3.80
80	19	1.50	90	31	2.90
120	19	1.30	126	31	2.20
8	22	4.90	140	31	2.20
16	22	3.80	170	31	1.70
24	22	4.10	50	39	4.60
40	22	3.20	70	39	4.40
70	22	1.80	100	39	3.70
100	22	1.70	150	39	3.30
140	22	1.10	190	39	2.60
10	25	5.00	200	39	2.50
18	25	4.40	80	43	4.40
24	25	4.30	120	43	4.20
36	25	3.90	160	43	3.70
55	25	2.90	200	43	3.60

Tableau 4.2 : *Conditions et notes MOS du test de locution avec écho seul - Article [Appel et Beerends 2002]*

contenu, langue) sur le score PESQM et de donner un jugement moyen sur 4 locuteurs (2 hommes et 2 femmes), de la même façon que dans les tests subjectifs la note MOS est obtenue en moyennant le jugement de tous les sujets. Le mapping entre les notes MOS subjectives et les scores PESQM est fourni dans la figure 4.1 ainsi que le coefficient de corrélation de Pearson r.

Tout d'abord, le modèle PESQM, implémenté à partir de [Appel et Beerends 2002], donne des résultats cohérents et fortement corrélés avec les notes subjectives des deux tests de locution. En toute logique, la corrélation est très élevée pour le test de l'article [Appel et Beerends 2002], les paramètres p et q du modèle ayant été optimisés sur ces conditions et ces



(a) Contribution UIT-T COM 12-16 [Gierlich et Diedrich 2000]

(b) Article [Appel et Beerends 2002]

Figure 4.1 : *Mapping du score PESQM moyen sur 28 fichiers et des notes MOS subjectives de locution*

notes MOS, et elle est plus faible pour le test de locution de la Contribution UIT-T COM 12-16 [Gierlich et Diedrich 2000].

4.1.2 Test de locution

Afin de pouvoir « vérifier » et optimiser notre implémentation de PESQM, nous avons mis en œuvre un nouveau test de locution.

4.1.2.1 Protocole

Ce protocole de test subjectif détermine la qualité de locution d'une liaison vocale en faisant appel à des sujets non experts. Il consiste à faire parler les sujets dans un combiné téléphonique, relié à un système de communication dans lequel des dégradations de la qualité de locution sont introduites. Ils doivent prononcer deux doubles-phrases. Les dégradations sont l'écho, l'amplification du niveau de l'effet local, un bruit du côté distant pouvant masquer le signal d'écho quand celui-ci est présent et des pertes de paquets. Du fait du grand nombre de conditions à tester, notre test a été divisé en deux sessions.

Les sujets ont pour instruction d'écouter dans le combiné pendant qu'ils parlent, puis de donner un jugement de qualité en répondant aux questions qui leur sont posées sur la qualité globale, la gêne due à l'écho, la gêne due aux bruits et la qualité de l'effet local (*cf.* questions dans l'annexe E). Les échelles de notation ACR de la Recommandation UIT-T P.800 [UIT-T Rec. P.800 1996] et de gêne due à l'écho de la Recommandation UIT-T P.831 [UIT-T Rec. P.831 1998] sont utilisées.

Dans un souci de rapidité du test en lui-même, ainsi que pour faciliter la collecte des notes, le déroulement du test est automatisé avec l'utilisation d'un logiciel de notation par les sujets. Pour chaque condition, le numéro de téléphone à composer et les phrases à prononcer apparaissent à l'écran. Une fois la communication terminée, le sujet répond aux questions présentées à l'écran.

Première session Lors de la première session, les dégradations expérimentées sont l'écho, l'amplification du niveau de l'effet local et le bruit du côté distant. Cette session porte sur l'évaluation de la qualité en situation de locution, avec 57 conditions de test différentes. Vingt-quatre sujets non experts participent à cette session. L'ordre des conditions et des phrases pour chaque sujet est aléatoire (carré gréco-latin). Les facteurs expérimentaux et les conditions du test sont décrits dans les tableaux 4.3 et 4.4 respectivement.

Facteurs	
Terminaux	téléphones analogiques
Niveau d'écoute	79 dB SPL
Nombre de sujets	24 sujets non experts
Mise en œuvre du test	carré gréco-latin
Méthodes de notation	ACR et DCR
Langue	français

Tableau 4.3 : *Facteurs expérimentaux - Session 1 du test de locution*

Condition	Délai de l'écho (ms)	Atténuation de l'écho (dB)	Niveau de l'effet local	Niveau du bruit
1	Pas d'écho	Pas d'écho	Amplifié faiblement	Pas de bruit
2	Pas d'écho	Pas d'écho	Par défaut	Pas de bruit
3	Pas d'écho	Pas d'écho	Amplifié fortement	Pas de bruit
4	100	40	Amplifié faiblement	Pas de bruit
5	100	40	Par défaut	Pas de bruit
6	100	40	Amplifié fortement	Pas de bruit
7	100	20	Amplifié faiblement	Pas de bruit
8	100	20	Par défaut	Pas de bruit
9	100	20	Amplifié fortement	Pas de bruit
10	200	40	Amplifié faiblement	Pas de bruit
11	200	40	Par défaut	Pas de bruit
12	200	40	Amplifié fortement	Pas de bruit
13	200	20	Amplifié faiblement	Pas de bruit
14	200	20	Par défaut	Pas de bruit
15	200	20	Amplifié fortement	Pas de bruit
16	400	40	Amplifié faiblement	Pas de bruit
17	400	40	Par défaut	Pas de bruit
18	400	40	Amplifié fortement	Pas de bruit
19	400	20	Amplifié faiblement	Pas de bruit
20	400	20	Par défaut	Pas de bruit
21	400	20	Amplifié fortement	Pas de bruit
22	100	40	Amplifié faiblement	Niveau 1
23	100	40	Par défaut	Niveau 1
24	100	40	Amplifié fortement	Niveau 1
25	100	20	Amplifié faiblement	Niveau 1
26	100	20	Par défaut	Niveau 1
27	100	20	Amplifié fortement	Niveau 1
28	200	40	Amplifié faiblement	Niveau 1
29	200	40	Par défaut	Niveau 1
30	200	40	Amplifié fortement	Niveau 1
31	200	20	Amplifié faiblement	Niveau 1
32	200	20	Par défaut	Niveau 1
33	200	20	Amplifié fortement	Niveau 1
34	400	40	Amplifié faiblement	Niveau 1
35	400	40	Par défaut	Niveau 1
36	400	40	Amplifié fortement	Niveau 1
37	400	20	Amplifié faiblement	Niveau 1
38	400	20	Par défaut	Niveau 1
39	400	20	Amplifié fortement	Niveau 1
40	100	40	Amplifié faiblement	Niveau 2
41	100	40	Par défaut	Niveau 2
42	100	40	Amplifié fortement	Niveau 2
43	100	20	Amplifié faiblement	Niveau 2
44	100	20	Par défaut	Niveau 2
45	100	20	Amplifié fortement	Niveau 2
46	200	40	Amplifié faiblement	Niveau 2
47	200	40	Par défaut	Niveau 2
48	200	40	Amplifié fortement	Niveau 2
49	200	20	Amplifié faiblement	Niveau 2
50	200	20	Par défaut	Niveau 2
51	200	20	Amplifié fortement	Niveau 2
52	400	40	Amplifié faiblement	Niveau 2
53	400	40	Par défaut	Niveau 2
54	400	40	Amplifié fortement	Niveau 2
55	400	20	Amplifié faiblement	Niveau 2
56	400	20	Par défaut	Niveau 2
57	400	20	Amplifié fortement	Niveau 2

Tableau 4.4 : *Conditions - Session 1 du test de locution. NB : niveau 1 = -55 dBm0p et niveau 2 = -40 dBm0p*

Les sujets s'entraînent au début du test avec trois conditions d'apprentissage : les conditions 1 (pas d'écho, pas de bruit et pas d'amplification du niveau d'effet local), 3 (pas d'écho, pas de bruit et niveau d'effet local amplifié fortement) et 14 (délai = 200 ms, atténuation de l'écho = 20 dB, pas de bruit et pas d'amplification du niveau d'effet local).

Seconde session Lors de la seconde session, les dégradations expérimentées sont l'écho, l'amplification du niveau de l'effet local, le bruit du côté distant et les pertes de paquets. Cette session porte sur l'évaluation de la qualité en situation de locution, avec 74 conditions de test différentes. Un taux de pertes de paquets de 50% a été choisi pour que l'effet des pertes de paquets sur le signal d'écho soit bien perçu par les participants. Vingt-quatre sujets non experts participent à cette session. L'ordre des conditions et des phrases pour chaque sujet est aléatoire (carré gréco-latin). Les conditions de la session sont décrites dans le tableau 4.5. Les facteurs expérimentaux et les conditions d'apprentissage de la seconde session sont les mêmes que ceux de la première.

4.1.2.2 Analyse des résultats

Les outils statistiques nécessaires à cette analyse sont présentés dans l'annexe D. Un seuil de significativité égal à 0.05 sera utilisé systématiquement et un astérisque signalera un niveau significatif ($p < 0.05$).

Première session Cette session porte sur l'écho (délai = 100, 200 et 400 ms, atténuation = 40 et 20 dB), le niveau de l'effet local (par défaut, amplifié faiblement et amplifié fortement) et le bruit (pas de bruit, niveau 1 et niveau 2). Les 24 sujets doivent juger 54 conditions (+ 3 conditions avec de l'effet local seul), selon quatre critères (qualité globale, gêne due à l'écho, gêne due aux bruits, effet local). Une analyse statistique des résultats obtenus est réalisée, en considérant séparément les quatre critères.

Qualité globale : Les résultats de l'ANOVA sont fournis dans le tableau 4.6. Les quatre facteurs (Bruit, Délai, Atténuation et Effet local) ont un effet significatif, les plus importants étant ceux de l'atténuation d'écho et du bruit. D'après le tableau 4.6, l'effet local a un effet significatif, mais faible par rapport aux effets des autres facteurs. D'après les interactions de l'effet local avec les 3 autres facteurs, seule celle entre l'effet local et le délai est significative. L'effet local semble avoir eu très peu d'effet sur le jugement des sujets : la figure 4.2 va donc représenter l'effet des trois autres facteurs en moyennant sur l'effet local. La figure 4.2 met particulièrement bien en évidence l'effet de l'atténuation de l'écho sur le jugement des sujets, qui se traduit par une baisse de la note MOS quand l'atténuation de l'écho diminue. L'effet du bruit est aussi très visible (plus le bruit augmente, plus la note MOS diminue) et est différent selon l'atténuation de l'écho (interaction Bruit×Atténuation significative).

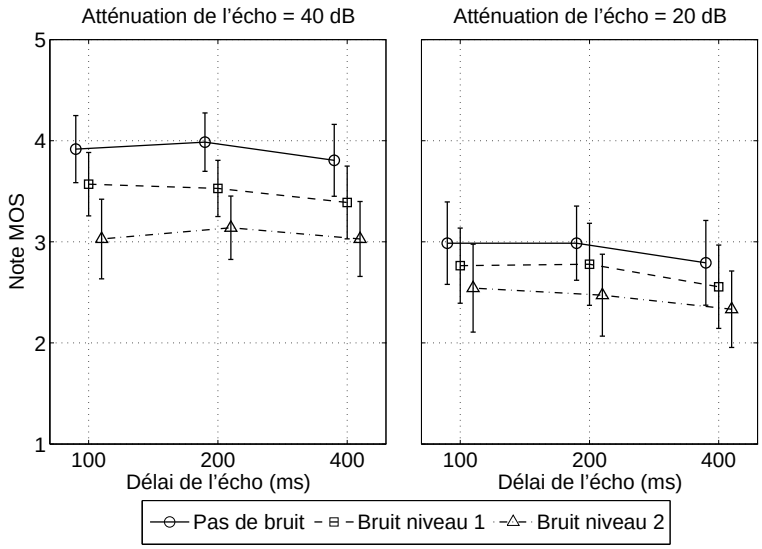


Figure 4.2 : *Interaction Bruit×Délai×Atténuation pour le critère de qualité globale - Session 1*

Condition	Délai de l'écho (ms)	Atténuation de l'écho (dB)	Niveau de l'effet local	Pertes de paquets	Niveau du bruit
1	Pas d'écho	Pas d'écho	Par défaut	Sans	Pas de bruit
2	Pas d'écho	Pas d'écho	Amplifié fortement	Sans	Pas de bruit
3	100	40	Par défaut	Sans	Pas de bruit
4	100	40	Amplifié fortement	Sans	Pas de bruit
5	100	20	Par défaut	Sans	Pas de bruit
6	100	20	Amplifié fortement	Sans	Pas de bruit
7	200	40	Par défaut	Sans	Pas de bruit
8	200	40	Amplifié fortement	Sans	Pas de bruit
9	200	20	Par défaut	Sans	Pas de bruit
10	200	20	Amplifié fortement	Sans	Pas de bruit
11	400	40	Par défaut	Sans	Pas de bruit
12	400	40	Amplifié fortement	Sans	Pas de bruit
13	400	20	Par défaut	Sans	Pas de bruit
14	400	20	Amplifié fortement	Sans	Pas de bruit
15	100	40	Par défaut	Avec	Pas de bruit
16	100	40	Amplifié fortement	Avec	Pas de bruit
17	100	20	Par défaut	Avec	Pas de bruit
18	100	20	Amplifié fortement	Avec	Pas de bruit
19	200	40	Par défaut	Avec	Pas de bruit
20	200	40	Amplifié fortement	Avec	Pas de bruit
21	200	20	Par défaut	Avec	Pas de bruit
22	200	20	Amplifié fortement	Avec	Pas de bruit
23	400	40	Par défaut	Avec	Pas de bruit
24	400	40	Amplifié fortement	Avec	Pas de bruit
25	400	20	Par défaut	Avec	Pas de bruit
26	400	20	Amplifié fortement	Avec	Pas de bruit
27	100	40	Par défaut	Sans	Niveau 1
28	100	40	Amplifié fortement	Sans	Niveau 1
29	100	20	Par défaut	Sans	Niveau 1
30	100	20	Amplifié fortement	Sans	Niveau 1
31	200	40	Par défaut	Sans	Niveau 1
32	200	40	Amplifié fortement	Sans	Niveau 1
33	200	20	Par défaut	Sans	Niveau 1
34	200	20	Amplifié fortement	Sans	Niveau 1
35	400	40	Par défaut	Sans	Niveau 1
36	400	40	Amplifié fortement	Sans	Niveau 1
37	400	20	Par défaut	Sans	Niveau 1
38	400	20	Amplifié fortement	Sans	Niveau 1
39	100	40	Par défaut	Sans	Niveau 2
40	100	40	Amplifié fortement	Sans	Niveau 2
41	100	20	Par défaut	Sans	Niveau 2
42	100	20	Amplifié fortement	Sans	Niveau 2
43	200	40	Par défaut	Sans	Niveau 2
44	200	40	Amplifié fortement	Sans	Niveau 2
45	200	20	Par défaut	Sans	Niveau 2
46	200	20	Amplifié fortement	Sans	Niveau 2
47	400	40	Par défaut	Sans	Niveau 2
48	400	40	Amplifié fortement	Sans	Niveau 2
49	400	20	Par défaut	Sans	Niveau 2
50	400	20	Amplifié fortement	Sans	Niveau 2
51	100	40	Par défaut	Avec	Niveau 1
52	100	40	Amplifié fortement	Avec	Niveau 1
53	100	20	Par défaut	Avec	Niveau 1
54	100	20	Amplifié fortement	Avec	Niveau 1
55	200	40	Par défaut	Avec	Niveau 1
56	200	40	Amplifié fortement	Avec	Niveau 1
57	200	20	Par défaut	Avec	Niveau 1
58	200	20	Amplifié fortement	Avec	Niveau 1
59	400	40	Par défaut	Avec	Niveau 1
60	400	40	Amplifié fortement	Avec	Niveau 1
61	400	20	Par défaut	Avec	Niveau 1
62	400	20	Amplifié fortement	Avec	Niveau 1
63	100	40	Par défaut	Avec	Niveau 2
64	100	40	Amplifié fortement	Avec	Niveau 2
65	100	20	Par défaut	Avec	Niveau 2
66	100	20	Amplifié fortement	Avec	Niveau 2
67	200	40	Par défaut	Avec	Niveau 2
68	200	40	Amplifié fortement	Avec	Niveau 2
69	200	20	Par défaut	Avec	Niveau 2
70	200	20	Amplifié fortement	Avec	Niveau 2
71	400	40	Par défaut	Avec	Niveau 2
72	400	40	Amplifié fortement	Avec	Niveau 2
73	400	20	Par défaut	Avec	Niveau 2
74	400	20	Amplifié fortement	Avec	Niveau 2

Tableau 4.5 : *Conditions - Session 2 du test de locution. NB : Taux de pertes de paquets = 50%, niveau 1 = -55 dBm0p et niveau 2 = -40 dBm0p*

Gêne due à l'écho : Les résultats de l'ANOVA sont donnés dans le tableau 4.7. De nouveau, comme pour le critère de qualité globale, les quatre facteurs (Bruit, Délai, Atté-

Facteur	SC	dl	CM	F	P>F
Bruit	92.7	2	46.4	76.3	0.000*
Délai	7.2	2	3.6	8.6	0.001*
Atténuation	206.2	1	206.2	119.7	0.000*
Effet local	10.6	2	5.3	5.0	0.010*
Bruit×Délai	0.4	4	0.1	0.2	0.958
Bruit×Atténuation	7.2	2	3.6	9.7	0.000*
Délai×Atténuation	0.6	2	0.3	0.4	0.681
Bruit×Effet local	0.9	4	0.2	0.5	0.769
Délai×Effet local	6.0	4	1.5	3.5	0.010*
Atténuation×Effet local	3.2	2	1.6	2.5	0.097
Bruit×Délai×Atténuation	0.6	4	0.1	0.3	0.866
Bruit×Délai×Effet local	4.6	8	0.6	1.3	0.226
Bruit×Atténuation×Effet local	1.2	4	0.3	0.8	0.501
Délai×Atténuation×Effet local	2.1	4	0.5	1.3	0.286
Bruit×Délai×Atténuation×Effet local	2.3	8	0.3	0.7	0.667

Tableau 4.6 : *ANOVA pour le critère de qualité globale - Session 1*

nuation et Effet local) ont un effet significatif, l'effet local ayant aussi un effet beaucoup plus faible que les autres facteurs et l'atténuation ayant un effet prédominant sur les effets des trois autres facteurs. L'effet des facteurs Bruit, Délai et Atténuation est analysé en moyennant sur le facteur Effet local dans la figure 4.3. Le facteur Atténuation a un effet très visible, ce qui est logique étant donné que la question portait sur l'écho. Un effet du délai est également observé : il se traduit par une diminution de la note MOS quand le délai augmente. Le bruit a un effet de masquage de l'écho : plus le bruit augmente, plus la note MOS relative à l'écho augmente.

Facteur	SC	dl	CM	F	P>F
Bruit	55	2	27	43.4	0.000*
Délai	59	2	29	37.5	0.000*
Atténuation	1194	1	1194	443.5	0.000*
Effet local	19	2	10	8.9	0.001*
Bruit×Délai	8	4	2	4.1	0.005*
Bruit×Atténuation	1	2	0.0	0.7	0.491
Délai×Atténuation	1	2	0.0	0.5	0.598
Bruit×Effet local	5	4	1	3	0.023*
Délai×Effet local	5.0	4	1	2.1	0.082
Atténuation×Effet local	8	2	4	4.6	0.015*
Bruit×Délai×Atténuation	5	4	1	2.2	0.071
Bruit×Délai×Effet local	1	8	0.0	0.3	0.959
Bruit×Atténuation×Effet local	3	4	1	1.2	0.324
Délai×Atténuation×Effet local	5	4	1	2.1	0.092
Bruit×Délai×Atténuation×Effet local	3	8	0.0	0.8	0.631

Tableau 4.7 : *ANOVA pour le critère de gêne due à l'écho - Session 1*

Qualité de l'effet local : Les résultats de l'ANOVA sont donnés dans le tableau 4.8. Trois des quatre facteurs (Bruit, Atténuation et Effet local) ont un effet significatif. Contrairement à ce qui pouvait être attendu, le facteur Effet local n'est pas le plus significatif, alors que la question porte sur l'effet local. Il semble que les sujets aient fait la confusion entre effet local et présence de bruits ou d'écho, et aient eu du mal à noter ce critère. D'après les corrélations entre les différents critères, fournies dans le tableau 4.9, le critère d'effet local est effectivement corrélé avec les trois autres critères et en particulier avec la qualité globale. Cela traduit la difficulté que les sujets ont eu à noter ce critère d'effet local.

Gêne due aux bruits : Les résultats de l'ANOVA sont donnés dans le tableau 4.10. Trois des quatre facteurs (Bruit, Atténuation et Effet local) ont un effet significatif. Le facteur Bruit a, de façon logique, un effet prédominant par rapport aux trois autres facteurs.

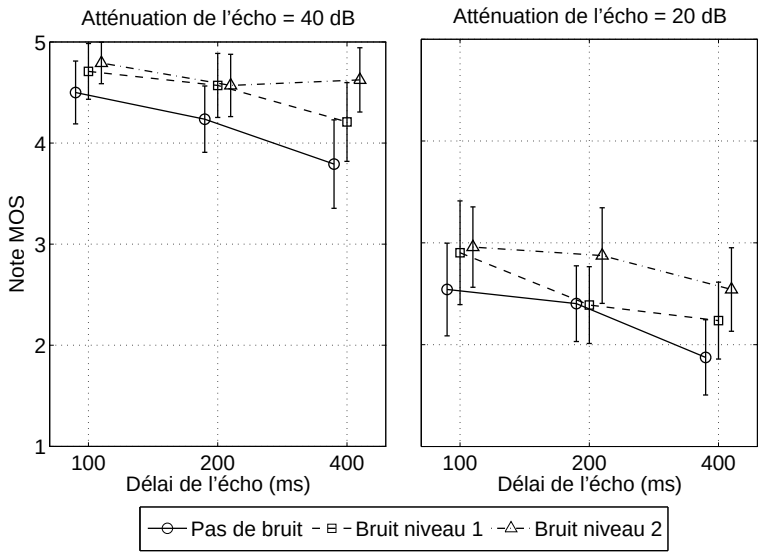


Figure 4.3 : *Interaction Bruit×Délai×Atténuation pour le critère de gêne due à l'écho - Session 1*

Facteur	SC	dl	CM	F	P>F
Bruit	63.8	2	31.9	47.99	0.000*
Délai	3.5	2	1.7	3.15	0.052
Atténuation	129.7	1	129.7	37.12	0.000*
Effet local	55.1	2	27.6	9.71	0.000*
Bruit×Délai	2.0	4	0.5	0.63	0.640
Bruit×Atténuation	6.3	2	3.1	6.27	0.004*
Délai×Atténuation	1.0	2	0.5	0.53	0.594
Bruit×Effet local	0.6	4	0.1	0.25	0.907
Délai×Effet local	6.2	4	1.5	2.09	0.088
Atténuation×Effet local	11.2	2	5.6	7.37	0.002*
Bruit×Délai×Atténuation	1.3	4	0.3	0.58	0.679
Bruit×Délai×Effet local	7.5	8	0.9	1.95	0.055
Bruit×Atténuation×Effet local	3.8	4	1.0	1.92	0.113
Délai×Atténuation×Effet local	2.7	4	0.7	1.10	0.362
Bruit×Délai×Atténuation×Effet local	3.5	8	0.4	0.88	0.535

Tableau 4.8 : *ANOVA pour le critère de qualité de l'effet local - Session 1*

	Qualité globale	Écho	Bruits	Effet local
Qualité globale	1	0.40	0.51	0.72
Écho	0.40	1	0.10	0.29
Bruits	0.51	0.10	1	0.38
Effet local	0.72	0.29	0.38	1

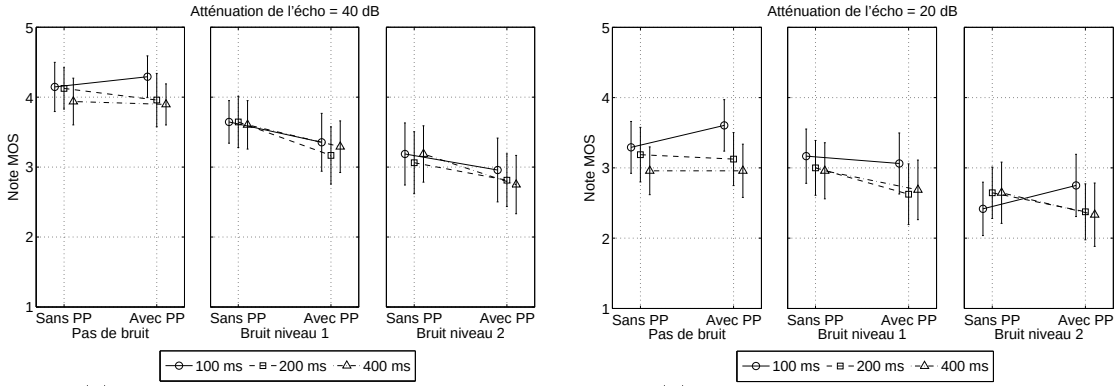
Tableau 4.9 : *Corrélations entre les critères - Session 1*

Seconde session Cette session porte sur l'écho (délai = 100, 200 et 400 ms, atténuation = 40 et 20 dB), le niveau de l'effet local (par défaut et amplifié fortement), le bruit (pas de bruit, niveau 1 et niveau 2) et les pertes de paquets (sans et avec). Les 24 sujets devaient donc juger 72 conditions (+ 2 conditions avec de l'effet local seul), avec 4 critères (qualité globale, gêne due à l'écho, gêne due aux bruits, effet local). Une analyse statistique des résultats obtenus est donc réalisée, en considérant séparément les 4 critères.

Qualité globale : Les résultats de l'ANOVA sont donnés dans le tableau 4.11. Quatre des cinq facteurs (Bruit, Pertes de paquets (notées PP), Délai et Atténuation) ont un effet significatif, alors que le facteur Effet local n'a pas d'effet significatif. L'effet des quatre autres facteurs va donc être analysé en moyennant sur l'effet local et en séparant l'analyse par rapport au facteur ayant l'effet le plus fort, *i.e.* l'atténuation. La figure 4.4 présente les résultats de l'interaction Bruit×PP×Délai×Atténuation, avec Atténuation = 40 dB et Atténuation = 20

Facteur	SC	dl	CM	F	P>F
Bruit	596.5	2	298.3	122.6	0.000*
Délai	0.4	2	0.2	0.4	0.706
Atténuation	18.3	1	18.3	16.5	0.000*
Effet local	8.8	2	4.4	7.9	0.001*
Bruit×Délai	0.3	4	0.1	0.1	0.967
Bruit×Atténuation	2.3	2	1.2	1.8	0.173
Délai×Atténuation	0.2	2	0.1	0.1	0.883
Bruit×Effet local	15.2	4	3.8	6.5	0.000*
Délai×Effet local	4.1	4	1.0	2.0	0.104
Atténuation×Effet local	2.8	2	1.4	2.8	0.072
Bruit×Délai×Atténuation	1.2	4	0.3	0.7	0.591
Bruit×Délai×Effet local	2.5	8	0.3	0.6	0.736
Bruit×Atténuation×Effet local	2.8	4	0.7	1.5	0.220
Délai×Atténuation×Effet local	2.6	4	0.7	1.4	0.231
Bruit×Délai×Atténuation×Effet local	8.6	8	1.1	2.4	0.017*

Tableau 4.10 : *ANOVA pour le critère de gêne due aux bruits - Session 1*



(a) Atténuation de l'écho = 40 dB (b) Atténuation de l'écho = 20 dB

Figure 4.4 : *Interaction Bruit×Pertes de paquets×Délai×Atténuation pour le critère de qualité globale - Session 2*

dB séparément. Elle illustre l'effet très net de l'atténuation, qui se traduit par une diminution des notes MOS quand l'atténuation diminue. La figure confirme également l'interaction significative qui existe entre le bruit et les pertes de paquets. Ainsi, l'effet des pertes de paquets dépend de la présence ou non de bruit. Dans le cas sans bruit, les pertes de paquets n'ont pas ou peu d'effet, ce qui signifie que les sujets ont perçu la présence des pertes de paquets seulement en présence de bruit. En effet, lors du test, le bruit, quand il était présent, était présent en continu pendant la communication (contrairement à l'écho qui n'était présent que quand le sujet parlait), et il était sujet aux pertes de paquets. Les pertes de paquets ont ainsi eu plus d'effet sur la perception du bruit (occurrence des pertes plus fréquentes) que sur la perception de l'écho par les sujets.

Gêne due à l'écho : Les résultats de l'ANOVA sont donnés dans le tableau 4.12. Les cinq facteurs (Bruit, Pertes de paquets, Délai, Atténuation et Effet local) ont un effet significatif, avec une prédominance très forte des facteurs Atténuation et Délai, ce qui est logique puisque le critère évalué est la gêne due à l'écho.

Qualité de l'effet local : Les résultats de l'ANOVA sont donnés dans le tableau 4.13. Les cinq facteurs (Bruit, Pertes de paquets, Délai, Atténuation et Effet local) ont un effet significatif. Une prédominance forte du facteur Atténuation est constatée, ce qui confirme la conclusion de l'analyse de la première session, à savoir que les sujets ont fait la confusion entre effet local et écho ou alors que les sujets ont noté l'effet local comme la qualité globale.

Facteur	SC	dl	CM	F	P>F
Bruit	180.8	2	90.4	56.3	0.000*
PP	32.0	1	32.0	46.8	0.000*
Délai	16.0	2	8.0	20.3	0.000*
Atténuation	168.1	1	168.1	134.0	0.000*
Effet local	0.5	1	0.5	0.2	0.634
Bruit×PP	20.8	2	10.4	26.8	0.000*
Bruit×Délai	9.0	4	2.3	8.1	0.000*
PP×Délai	0.2	2	0.1	0.4	0.649
Bruit×Atténuation	15.3	2	7.6	9.6	0.000*
PP×Atténuation	0.7	1	0.7	1.4	0.244
Délai×Atténuation	1.4	2	0.7	2.3	0.117
Bruit×Effet local	7.3	2	3.7	9.7	0.000*
PP×Effet local	0.0	1	0.0	0.1	0.751
Délai×Effet local	1.8	2	0.9	1.7	0.187
Atténuation×Effet local	5.4	1	5.4	9.5	0.005*
Bruit×PP×Délai	4.7	4	1.2	3.5	0.011*
Bruit×PP×Atténuation	0.3	2	0.1	0.3	0.725
Bruit×Délai×Atténuation	1.7	4	0.4	1.4	0.226
PP×Délai×Atténuation	0.5	2	0.3	0.6	0.544
Bruit×PP×Effet local	0.3	2	0.2	0.7	0.524
Bruit×Délai×Effet local	1.3	4	0.3	1.0	0.424
PP×Délai×Effet local	0.5	2	0.2	0.9	0.415
Bruit×Atténuation×Effet local	0.6	2	0.3	0.5	0.581
PP×Atténuation×Effet local	0.1	1	0.1	0.5	0.496
Délai×Atténuation×Effet local	1.4	2	0.7	1.6	0.214
Bruit×PP×Délai×Atténuation	1.4	4	0.4	1.1	0.358
Bruit×PP×Délai×Effet local	6.6	4	1.6	7.2	0.000*
Bruit×PP×Atténuation×Effet local	2.7	2	1.3	3.8	0.031*
Bruit×Délai×Atténuation×Effet local	1.3	4	0.3	1.2	0.310
PP×Délai×Atténuation×Effet local	1.3	2	0.7	2.2	0.121
Bruit×PP×Délai×Atténuation×Effet local	0.4	4	0.1	0.4	0.832

Tableau 4.11 : *ANOVA pour le critère de qualité globale - Session 2*

L'étude des corrélations entre les différents critères, fournies dans le tableau 4.14, montre, comme pour la première session, que le critère d'effet local est corrélé avec les trois autres critères et en particulier avec la qualité globale. Cela traduit de nouveau la difficulté que les sujets ont eu à noter ce critère d'effet local.

Gêne due aux bruits : Les résultats de l'ANOVA sont donnés dans le tableau 4.15. Trois des cinq facteurs (Bruit, Pertes de paquets et Atténuation) ont un effet significatif. Les facteurs Bruit et Pertes de paquets ont, de façon logique, un effet prédominant par rapport aux trois autres facteurs. L'interaction entre les facteurs Bruit et Pertes de paquets est forte, ce qui s'explique, comme pour le critère de qualité globale, par le fait que les pertes de paquets étaient plus perceptibles en présence de bruit qu'en absence de bruit.

4.1.2.3 Enregistrement des signaux de test

Parallèlement au test, nous avons enregistré des signaux dans les mêmes conditions que celles du test. Ces enregistrements ont été effectués à l'aide d'un simulateur de tête et de torse (HATS) [UIT-T Rec. P.58 1996]. Un signal est envoyé « à la bouche » du HATS. Le HATS « prononce » chaque signal, qui est envoyé dans le système, pour chacune des conditions du test. Le signal reçu à l'oreille du HATS par le combiné est enregistré, ce qui permet ainsi de recueillir les éventuels écho et effet local. Ce signal servira de signal dégradé. Le signal de référence correspondant est le signal envoyé dans le système avec la condition sans écho, sans bruit, sans pertes de paquets et avec un effet local « par défaut » (*i.e.* celui fourni par le combiné utilisé pour la mesure). Ici, nous disposions de 16 signaux, prononcés par 4 locuteurs différents (2 femmes et 2 hommes) en français, sous la forme de 4 doubles-phrases par locuteur.

Facteur	SC	dl	CM	F	P>F
Bruit	62	2	31	31.2	0.000*
PP	8.0	1	8.0	14.1	0.001*
Délai	139	2	69	71.4	0.000*
Atténuation	1169	1	1169	582.1	0.000*
Effet local	57	1	57	33	0.000*
Bruit×PP	3.0	2	1	3.3	0.047*
Bruit×Délai	28.	4	7.0	14.7	0.000*
PP×Délai	4.0	2	2.0	4.7	0.014*
Bruit×Atténuation	13.0	2	6.0	9.7	0.000*
PP×Atténuation	1.0	1	1.0	2.0	0.172
Délai×Atténuation	11.0	2	5.0	6.6	0.003*
Bruit×Effet local	10.0	2	5.0	7.8	0.001*
PP×Effet local	5.0	1	5.0	11.3	0.003*
Délai×Effet local	0.0	2	0.0	0.5	0.608
Atténuation×Effet local	26.0	1	26.0	29.4	0.000*
Bruit×PP×Délai	12.0	4	3.0	7.0	0.000*
Bruit×PP×Atténuation	0.0	2	0.0	0.5	0.604
Bruit×Délai×Atténuation	19.0	4	5.0	10.5	0.000*
PP×Délai×Atténuation	3.0	2	1.0	3.8	0.031*
Bruit×PP×Effet local	2.0	2	1.0	3.1	0.056
Bruit×Délai×Effet local	1.0	4	0.0	0.9	0.475
PP×Délai×Effet local	4.0	2	2.0	4.4	0.018*
Bruit×Atténuation×Effet local	2.0	2	1.0	2.2	0.118
PP×Atténuation×Effet local	0.0	1	0.0	0.0	0.883
Délai×Atténuation×Effet local	13.0	2	6.0	16.6	0.000*
Bruit×PP×Délai×Atténuation	4.0	4	1.0	2.0	0.096
Bruit×PP×Délai×Effet local	1.0	4	0.0	0.6	0.658
Bruit×PP×Atténuation×Effet local	7.0	2	4.0	8.8	0.001*
Bruit×Délai×Atténuation×Effet local	5.0	4	1.0	2.8	0.031*
PP×Délai×Atténuation×Effet local	2.0	2	1.0	2.4	0.099
Bruit×PP×Délai×Atténuation×Effet local	5.0	4	1.0	3.9	0.006*

Tableau 4.12 : *ANOVA pour le critère de gêne due à l'écho - Session 2*

4.1.2.4 Vérification de la reproductibilité des notes subjectives entre les deux sessions

Certaines conditions ont été testées à la fois au cours de la première et de la seconde session. La reproductibilité de ces conditions est vérifiée en comparant les notes attribuées lors des deux sessions, représentées dans la figure 4.5. La corrélation de Pearson entre les deux sessions vaut $r = 0.9053$ et l'erreur absolue moyenne $EAM = 0.25$ MOS, confirmant la bonne reproductibilité entre les deux sessions du test.

4.1.3 Étude de PESQM sur les résultats de notre test de locution

Étant donnée la bonne reproductibilité des notes subjectives entre les deux sessions du test de locution, nous avons considéré la moyenne des notes subjectives obtenues dans les mêmes conditions au cours des deux sessions. Ensuite, nous avons appliqué la version acoustique de PESQM aux résultats de notre test de locution, en conservant les valeurs des paramètres fournies dans [Appel et Beerends 2002], à savoir $p = 1.4$ et $q = 5$. Les conditions sont classées par dégradations : « écho seul », « effet local seul », « écho + effet local », « écho + bruit », « écho + pertes de paquets », « écho + effet local + pertes de paquets », « écho + bruit + effet local », « écho + bruit + pertes de paquets » et « écho + bruit + effet local + pertes de paquets ». Dans tous les cas, la note subjective utilisée pour le mapping note subjective/score PESQM est la note de qualité globale, sauf dans les cas avec « écho + bruit », où la note subjective considérée est la note de gêne due à l'écho. En effet, le modèle PESQM a été construit de telle sorte que le bruit soit pris en compte intrinsèquement comme pouvant masquer l'écho. Il n'est ainsi pas considéré comme une dégradation à part entière, mais plutôt comme un masquage de l'écho.

Facteur	SC	dl	CM	F	P>F
Bruit	126.3	2	63.15	28.37	0.000*
PP	13.5	1	13.55	11.25	0.003*
Délai	9.0	2	4.5	9.53	0.000*
Atténuation	83.1	1	83.13	85.4	0.000*
Effet local	34.7	1	34.74	6.34	0.019*
Bruit×PP	8.5	2	4.23	7.39	0.002*
Bruit×Délai	4.2	4	1.05	2.46	0.051
PP×Délai	0.6	2	0.3	0.8	0.455
Bruit×Atténuation	1.9	2	0.93	1.27	0.290
PP×Atténuation	0.1	1	0.13	0.26	0.612
Délai×Atténuation	0.2	2	0.08	0.2	0.818
Bruit×Effet local	7.6	2	3.8	7.13	0.002*
PP×Effet local	0.0	1	0.00	0.00	0.967
Délai×Effet local	0.8	2	0.42	0.99	0.380
Atténuation×Effet local	3.6	1	3.61	6.76	0.016*
Bruit×PP×Délai	0.7	4	0.17	0.48	0.753
Bruit×PP×Atténuation	0.1	2	0.05	0.09	0.915
Bruit×Délai×Atténuation	1.6	4	0.39	0.93	0.447
PP×Délai×Atténuation	0.8	2	0.39	0.86	0.432
Bruit×PP×Effet local	0.1	2	0.04	0.13	0.876
Bruit×Délai×Effet local	1.5	4	0.39	1.08	0.370
PP×Délai×Effet local	0.2	2	0.12	0.42	0.660
Bruit×Atténuation×Effet local	0.8	2	0.42	0.92	0.405
PP×Atténuation×Effet local	0.2	1	0.17	0.49	0.489
Délai×Atténuation×Effet local	0.2	2	0.1	0.29	0.753
Bruit×PP×Délai×Atténuation	0.7	4	0.18	0.58	0.677
Bruit×PP×Délai×Effet local	6.2	4	1.55	3.98	0.005*
Bruit×PP×Atténuation×Effet local	3.7	2	1.86	5.33	0.008*
Bruit×Délai×Atténuation×Effet local	0.8	4	0.2	0.62	0.649
PP×Délai×Atténuation×Effet local	1.7	2	0.86	1.99	0.149
Bruit×PP×Délai×Atténuation×Effet local	0.7	4	0.17	0.49	0.739

Tableau 4.13 : *ANOVA pour le critère de qualité de l'effet local - Session 2*

	Qualité globale	Écho	Bruits	Effet local
Qualité globale	1	0.37	0.57	0.73
Écho	0.37	1	0.06	0.25
Bruits	0.57	0.06	1	0.42
Effet local	0.73	0.25	0.42	1

Tableau 4.14 : *Corrélations entre les critères - Session 2*

Pour chacune des dégradations, la corrélation r entre les notes subjectives et les scores PESQM correspondants est calculée. Ensuite, un mapping (régression non linéaire) entre les notes subjectives et les scores PESQM correspondants est effectué, avec un polynôme d'ordre 3 ($P(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$) ayant la contrainte d'être monotone décroissant ($P'(x) \leq 0$, pour tout x) sur l'intervalle de valeurs considérées. Cela revient à un problème de minimisation avec contraintes

$$\min_{\underline{\theta} \in \mathbf{R^4}} \sum_{i=1}^{n} \left[MOS_i - \left(\theta_0 + \theta_1 PESQM_i + \theta_2 PESQM_i^2 + \theta_3 PESQM_i^3 \right) \right]^2 \tag{4.1}$$

sous la contrainte

$$\theta_1 + 2\theta_2 PESQM_i + 3\theta_3 PESQM_i^2 \leq 0, \text{pour tout } 1 \leq i \leq n \tag{4.2}$$

où n est le nombre de conditions, $PESQM_i$ est le score PESQM de la condition i et MOS_i est la note subjective de la condition i.

Le vecteur $\underline{\theta}$ des coefficients optimaux et une estimation des notes subjectives à partir des scores PESQM sont obtenus. Le coefficient de détermination R^2 de la régression non

Facteur	SC	dl	CM	F	P>F
Bruit	917.6	2	458.8	149.7	0.000*
PP	105.5	1	105.5	192.9	0.000*
Délai	1.1	2	0.5	1.0	0.383
Atténuation	5.0	1	5.0	6.3	0.020*
Effet local	1.9	1	1.9	1.4	0.246
Bruit×PP	45.7	2	22.8	61.1	0.000*
Bruit×Délai	7.0	4	1.7	4.3	0.003*
PP×Délai	0.2	2	0.1	0.3	0.776
Bruit×Atténuation	0.1	2	0.0	0.0	0.953
PP×Atténuation	0.1	1	0.1	0.2	0.651
Délai×Atténuation	3.3	2	1.6	3.2	0.049*
Bruit×Effet local	47.6	2	23.8	36.2	0.000*
PP×Effet local	0.6	1	0.6	1.7	0.207
Délai×Effet local	0.7	2	0.4	0.8	0.476
Atténuation×Effet local	2.6	1	2.6	6.6	0.017*
Bruit×PP×Délai	1.9	4	0.5	1.3	0.271
Bruit×PP×Atténuation	0.1	2	0.1	0.2	0.858
Bruit×Délai×Atténuation	3.0	4	0.8	1.7	0.153
PP×Délai×Atténuation	0.2	2	0.1	0.3	0.710
Bruit×PP×Effet local	0.2	2	0.1	0.3	0.772
Bruit×Délai×Effet local	0.1	4	0.0	0.1	0.992
PP×Délai×Effet local	2.6	2	1.3	4.5	0.017*
Bruit×Atténuation×Effet local	2.1	2	1.0	2.2	0.118
PP×Atténuation×Effet local	0.1	1	0.1	0.4	0.528
Délai×Atténuation×Effet local	2.4	2	1.2	3.0	0.057
Bruit×PP×Délai×Atténuation	1.7	4	0.4	1.2	0.302
Bruit×PP×Délai×Effet local	1.5	4	0.4	0.8	0.558
Bruit×PP×Atténuation×Effet local	1.7	2	0.8	1.6	0.220
Bruit×Délai×Atténuation×Effet local	2.8	4	0.7	1.4	0.244
PP×Délai×Atténuation×Effet local	0.2	2	0.1	0.2	0.801
Bruit×PP×Délai×Atténuation×Effet local	1.9	4	0.5	0.9	0.455

Tableau 4.15 : *ANOVA pour le critère de gêne due aux bruits - Session 2*

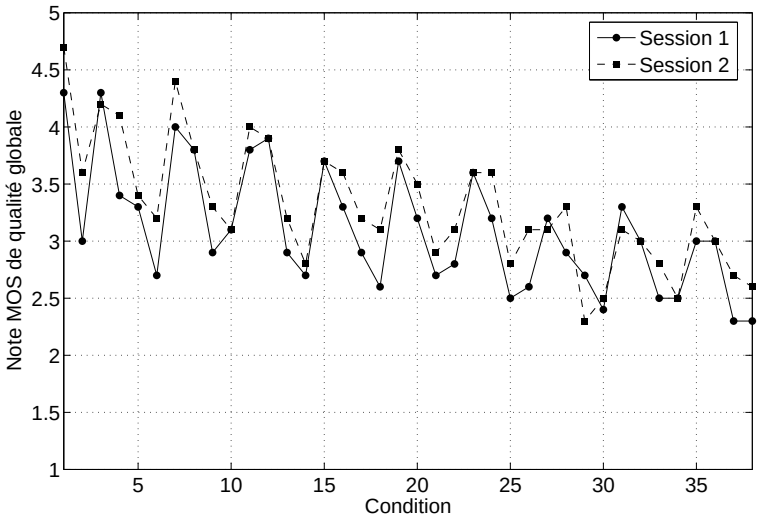


Figure 4.5 : *Comparaison des notes MOS de qualité globale obtenues dans les deux sessions pour les mêmes conditions*

linéaire est calculé. Les mappings entre scores PESQM et notes subjectives, pour chaque type de dégradation, sont donnés dans la figure 4.6. La corrélation obtenue entre les notes MOS et les scores PESQM dépend beaucoup des dégradations présentes. Ainsi, dès qu'il y a amplification de l'effet local, la corrélation est faible. Ceci s'explique par les résultats du test qui ont montré que les sujets n'étaient pas gênés par l'amplification de l'effet local, alors que le modèle PESQM fait la différence entre les trois niveaux de l'effet local. De bonnes corrélations ($|r| > 0.8$) sont obtenues dans les cas avec « écho seul », avec « écho + bruit »

et avec « écho + pertes de paquets ». Plus il y a de dégradations, moins la corrélation est élevée. La moyenne des corrélations (r) est égale à -0.751 et la moyenne des coefficients de détermination (R^2) est égale à 0.64.

4.1.4 Optimisation et validation de PESQM

4.1.4.1 Optimisation

Les valeurs de p et q, telles que $p > 1$ et $q > 1$, sont balayées. Les valeurs de p choisies sont comprises entre 1.1 et 5 avec un pas de 0.1. Les valeurs de q choisies sont comprises entre 2 et 50 avec un pas de 1. Pour chaque couple de paramètres (p, q) et pour chaque type de dégradation, le score PESQM correspondant à chaque condition et la corrélation r entre les scores PESQM et les notes subjectives sont calculés. Les corrélations (en valeur absolue) en fonction des paramètres p et q, pour chaque type de dégradation, sont données dans la figure 4.7.

Tout d'abord, de manière générale, les corrélations (en valeur absolue) sont maximales pour des faibles valeurs de p et q. De plus, les corrélations maximales sont élevées pour les conditions avec :

- « écho seul » : $|r|_{max} = 0.949$,
- « effet local seul » : $|r|_{max} = 0.997$,
- « écho + bruit » : $|r|_{max} = 0.978$,
- « écho + pertes de paquets » : $|r|_{max} = 0.946$.
- « écho + effet local + pertes de paquets » : $|r|_{max} = 0.895$.

Par contre, les corrélations maximales sont plus faibles pour les conditions avec :

- « écho + effet local » : $|r|_{max} = 0.623$,
- « écho + effet local + bruit » : $|r|_{max} = 0.658$,
- « écho + bruit + pertes de paquets » : $|r|_{max} = 0.639$,
- « écho + effet local + bruit + pertes de paquets » : $|r|_{max} = 0.714$.

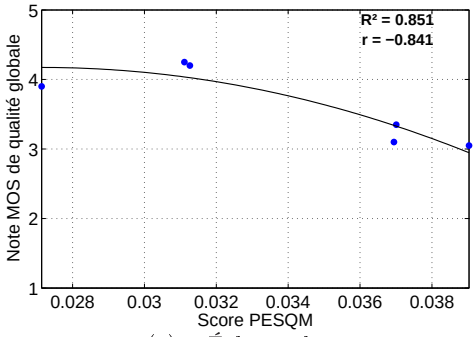
Les conditions affectées par plusieurs dégradations simultanées dont l'effet local semblent poser problème. Ce problème correspond à ce qui avait été déduit de l'analyse du test de locution (*cf.* section 4.1.2.2), à savoir que les sujets n'ont pas fait la différence entre les trois niveaux d'effet local, quand l'amplification de l'effet local était présente en même temps que l'écho. Dans le cas « effet local seul », les sujets ont fait la distinction entre les trois niveaux d'effet local, ce qui explique la bonne corrélation maximale $|r|_{max}$ entre scores PESQM et notes subjectives obtenue dans ce cas. Dans tous les autres cas avec amplification de l'effet local, les notes MOS et les scores PESQM sont peu corrélés puisque le modèle, contrairement aux sujets, fait la différence entre les trois niveaux d'effet local.

Étant donné ce problème particulier, auquel il est difficile de remédier sans modifier profondément le modèle PESQM puisqu'il est dû *à la base* à une prise en compte différente de l'amplification de l'effet local par les sujets et par le modèle, nous avons décidé d'exclure les conditions avec amplification de l'effet local de nos optimisation et validation de PESQM. Restent donc les conditions où l'effet local est à son niveau par défaut.

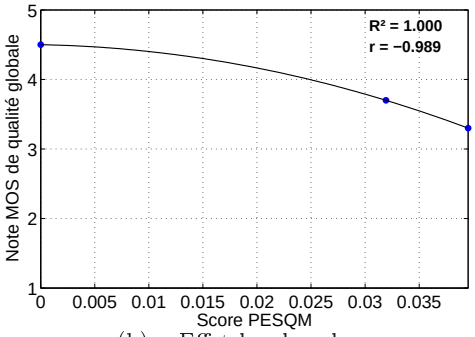
4.1.4.2 Choix des paramètres optimaux

Afin d'avoir une base de validation différente de la base d'optimisation, nous avons optimisé le modèle sur la moitié des signaux enregistrés, c'est-à-dire 8 signaux (2 signaux/locuteur, locuteurs hommes : 4 signaux, locuteurs femmes : 4 signaux) et validé le modèle sur l'autre moitié des signaux. Les conditions considérées sont donc :

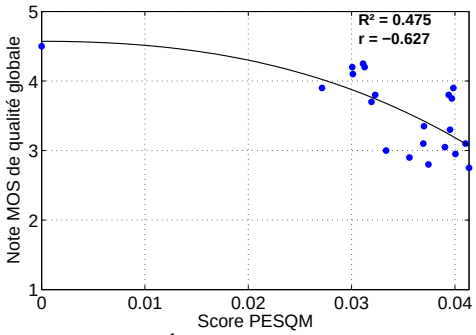
- « écho seul »,
- « écho + bruit »,
- « écho + pertes de paquets »,



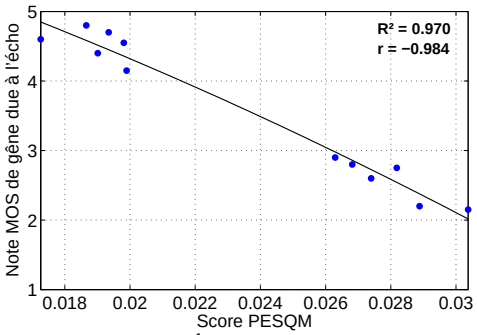
(a) « Écho seul »



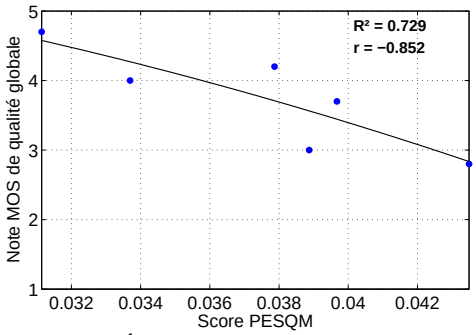
(b) « Effet local seul »



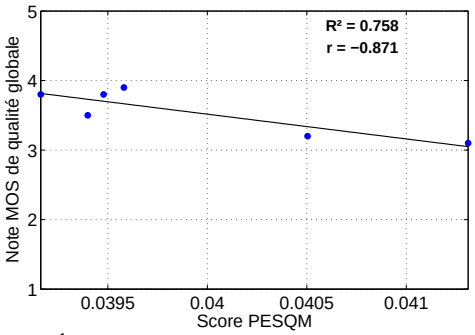
(c) « Écho + effet local »



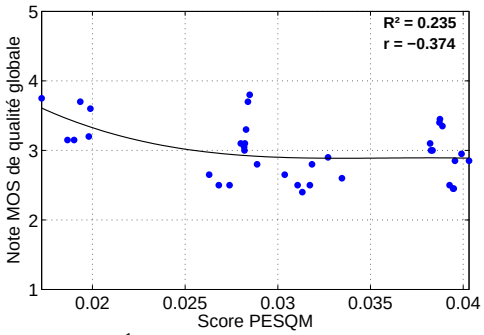
(d) « Écho + bruit »



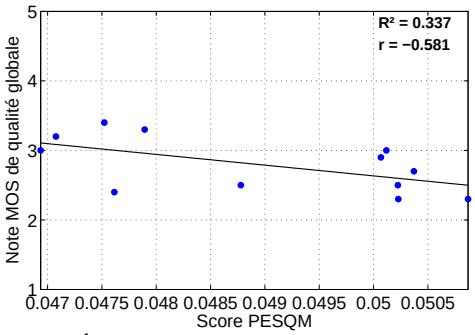
(e) « Écho + pertes de paquets »



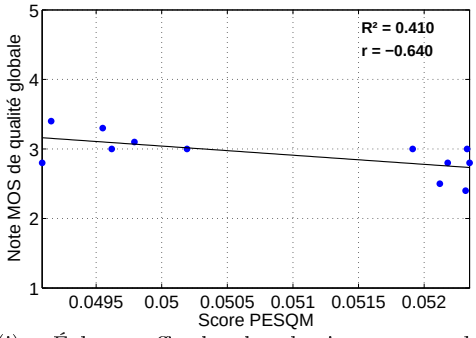
(f) « Écho + effet local + pertes de paquets »



(g) « Écho + effet local + bruit »

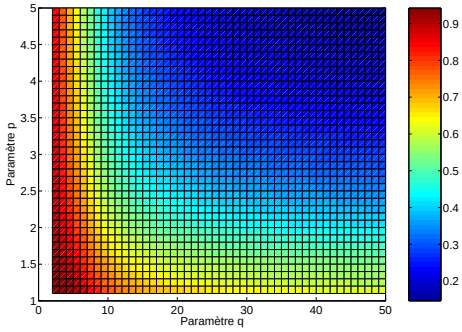


(h) « Écho + bruit + pertes de paquets »

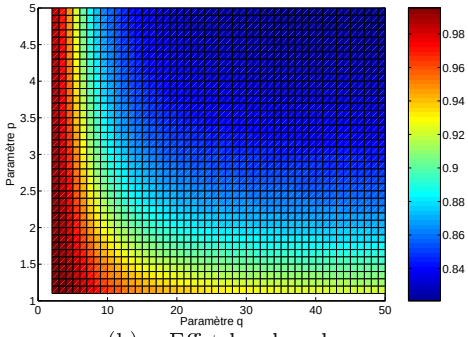


(i) « Écho + effet local + bruit + pertes de paquets »

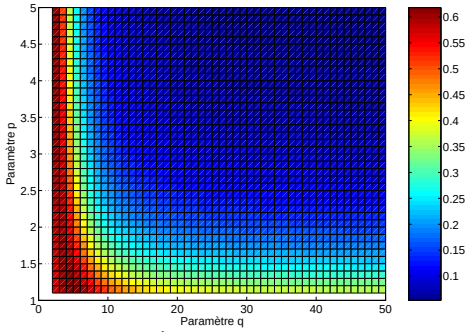
Figure 4.6 : *Mappings et corrélations entre scores PESQM et notes MOS, avec* $p = 1.4$ *et* $q = 5$



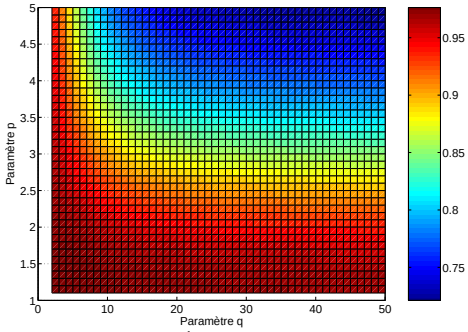
(a) « Écho seul »



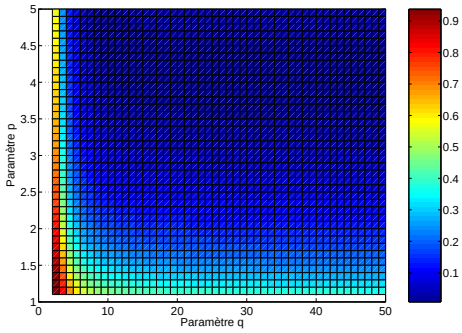
(b) « Effet local seul »



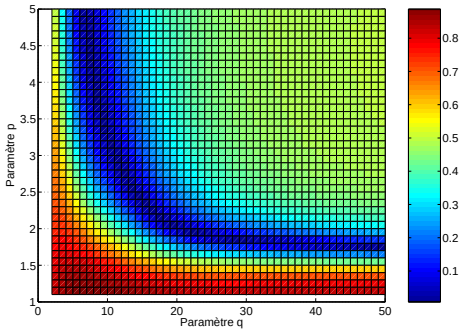
(c) « Écho + effet local »



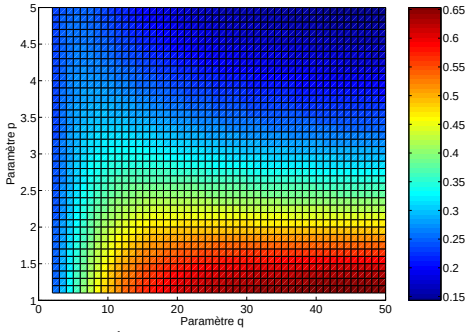
(d) « Écho + bruit »



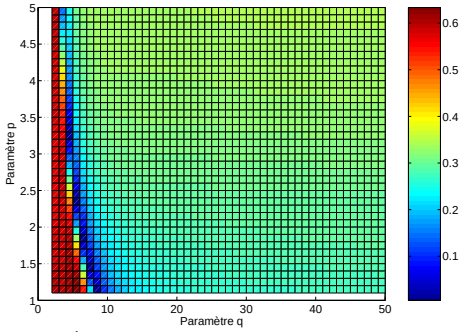
(e) « Écho + pertes de paquets »



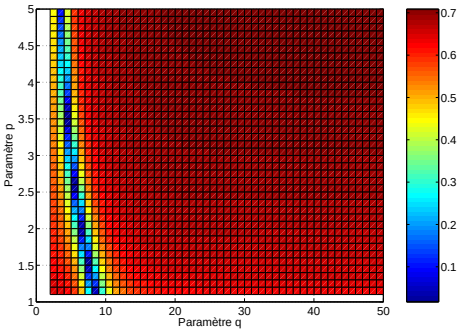
(f) « Écho + effet local + pertes de paquets »



(g) « Écho + effet local + bruit »



(h) « Écho + bruit + pertes de paquets »



(i) « Écho + effet local + bruit + pertes de paquets »

Figure 4.7 : *Corrélations (en valeur absolue) entre scores PESQM et notes subjectives en fonction des paramètres p et q*

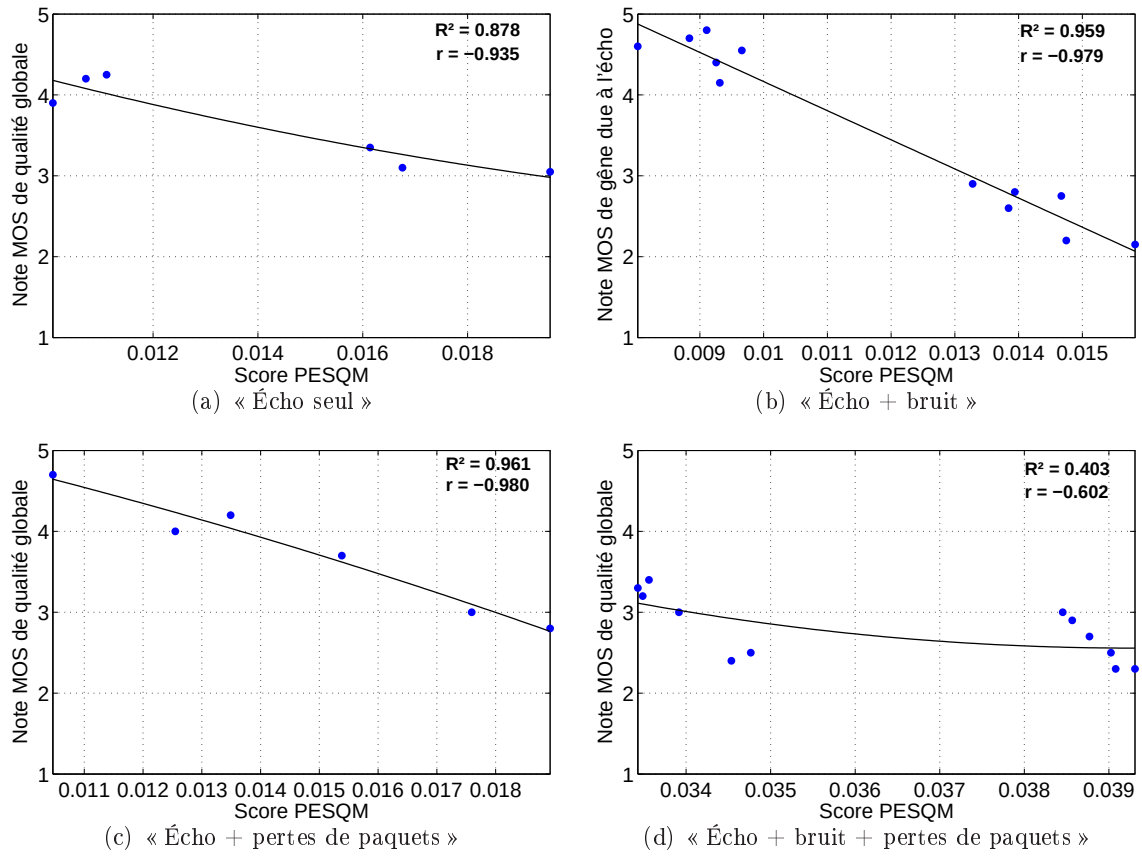


Figure 4.8 : *Mappings et corrélations entre scores PESQM et notes subjectives, avec* $p = 1.1$ *et* $q = 2.0$

– « écho + bruit + pertes de paquets ».

Nous avons choisi comme paramètres optimaux les maximisant la moyenne des corrélations entre scores PESQM et notes subjectives pour chacune des 4 dégradations (sur la base d'optimisation). Les paramètres obtenus valent $p_{opt} = 1.1$ et $q_{opt} = 2.0$, aboutissant à une corrélation moyenne $|r_{optim}| = 0.8697$ sur la base d'optimisation et à une corrélation moyenne $|r_{valid}| = 0.8589$ sur la base de validation. Ces valeurs p_{opt} et q_{opt} seront utilisées dans la suite du travail.

4.1.4.3 Version optimisée de PESQM appliquée au test de locution

Les mappings entre scores PESQM et notes subjectives avec les paramètres optimaux, pour chaque type de dégradation, sont donnés dans la figure 4.8.

Les performances finales du modèle optimisé sur l'ensemble des signaux sont données dans le tableau 4.16. Les corrélations sont très élevées, sauf dans le cas « écho + bruit + pertes de paquets » du fait de la particularité des conditions avec écho et bruit simultanément soulevée dans le paragraphe 4.1.3 pour lesquelles la note de gêne due à l'écho doit être utilisée.

Dégradation	Corrélation après mapping	Erreur absolue moyenne après mapping (en MOS)
« écho seul »	0.9373	0.1475
« écho + bruit »	0.9791	0.1822
« écho + pertes de paquets »	0.9805	0.1108
« écho + bruit + pertes de paquets »	0.6348	0.2508

Tableau 4.16 : *Performances du modèle PESQM optimisé* ($p = 1.1$ *et* $q = 2.0$)

La figure 4.8 donne une courbe de mapping par type de dégradation. Nous regroupons maintenant les conditions avec « écho seul » et avec « écho + pertes de paquets » pour estimer une courbe de mapping unique pour ces deux types de dégradations, comme cela est fait dans [Appel et Beerends 2002]. La note objective de locution $\widehat{MOS}$ est calculée à partir du score PESQM correspondant grâce à la courbe de mapping ainsi obtenue telle que

$$\widehat{MOS} = 4.89 + 2.42PESQM - 5667.2PESQM^2 - 220.3PESQM^3. \quad (4.3)$$

Les trois courbes de mapping (« écho seul », « écho + pertes de paquets » et « écho seul et écho + pertes de paquets ») sont comparées dans la figure 4.9, pour les conditions avec écho seul et écho + pertes de paquets. La courbe de mapping obtenue sur l'ensemble des conditions « écho seul et écho + pertes de paquets » aboutit à la meilleure corrélation de Pearson r entre les notes de locution subjectives et objectives. Cette courbe de mapping sera utilisée dans la suite du document pour transformer le score PESQM en note objective de qualité de locution.

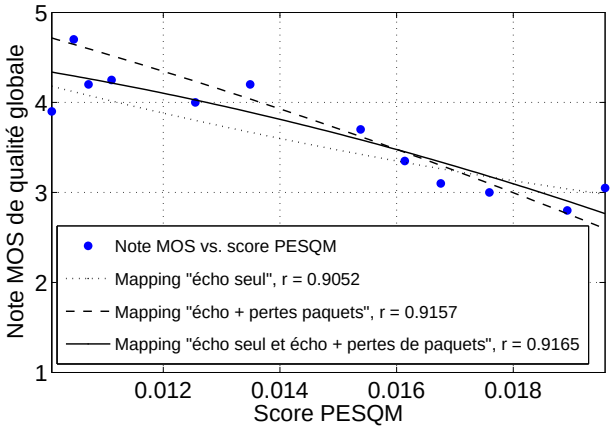


Figure 4.9 : *Comparaison des équations de mapping pour les conditions avec écho seul et avec écho + pertes de paquets*

4.1.4.4 Synthèse

La version de PESQM développée dans le cadre de la thèse à partir de [Appel et Beerends 2002] a été validée et optimisée grâce à un test de locution, conçu à cet effet. Les scores PESQM obtenus dans différentes conditions de dégradations sont très bien corrélés avec les notes subjectives de locution. Une fonction de mapping unique a été déterminée pour les conditions avec « écho seul » et avec « écho + pertes de paquets », aboutissant à une corrélation de Pearson $r = 0.9165$ entre les notes subjectives et objectives de locution.

4.2 Découpage des signaux de conversation

Nous présentons dans ce paragraphe un outil de traitement des signaux qui permet d'utiliser les signaux de conversation, enregistrés pendant des tests subjectifs ou pendant des communications réelles, avec les modèles intrusifs tels que PESQ et PESQM. Ces modèles, pour fonctionner de façon optimale, utilisent en entrée des signaux de parole de longueur comprise entre 8 et 30 secondes [UIT-T Rec. P.862.3 2005]. Les signaux enregistrés lors de la communication testée doivent donc être dimensionnés afin de respecter ces contraintes.

Le découpage des signaux pour PESQ est simple : il suffit de découper les signaux de référence et dégradé à la longueur désirée, le modèle PESQ effectuant l'alignement des deux signaux. Pour PESQM, le découpage est plus critique : il faut obtenir un signal dégradé « propre », c'est-à-dire ne contenant que l'écho éventuel et pas le signal de parole du participant distant qui perturberait le modèle dans son évaluation de la qualité de locution. Un exemple

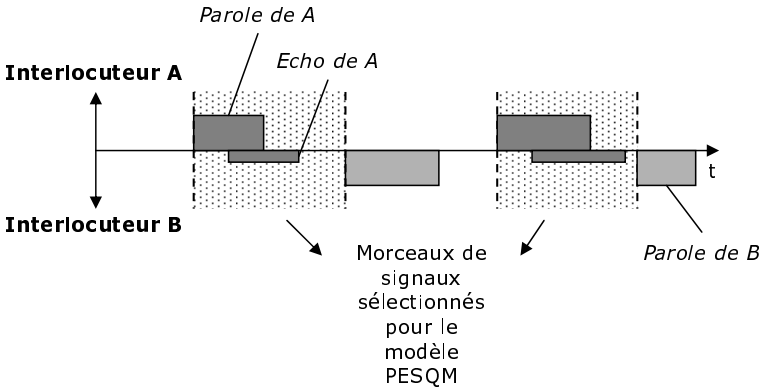


Figure 4.10 : *Illustration du découpage des signaux pour utilisation avec le modèle PESQM (qualité de locution évaluée pour l'interlocuteur A)*

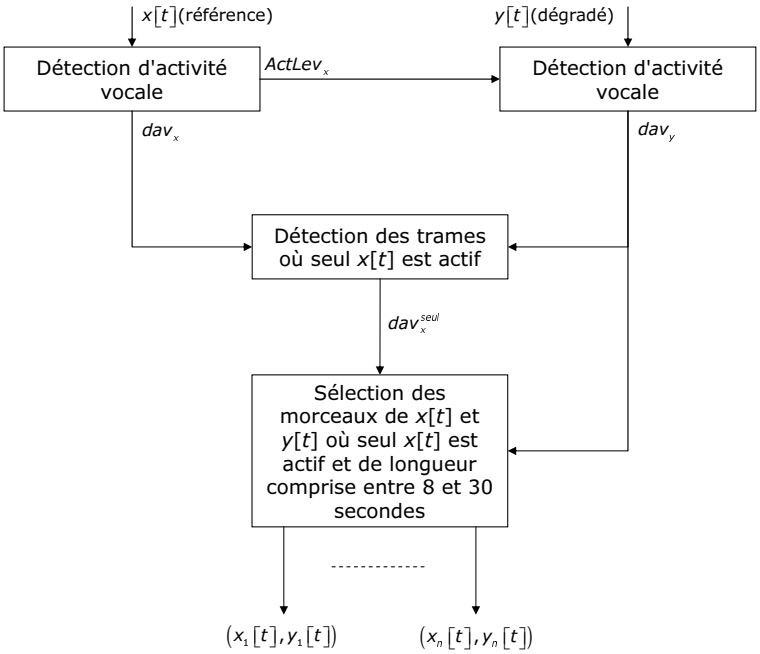


Figure 4.11 : *Algorithme de découpage des signaux pour utilisation avec le modèle PESQM*

est illustré dans la figure 4.10, quand la qualité de locution de l'interlocuteur A est évaluée. Cette sélection des morceaux de signaux appliqués en entrée de PESQM s'effectue grâce à une détection d'activité vocale (DAV), décrite dans la Recommandation P.56 [UIT-T Rec. P.56 1993] :

1. Détection de l'enveloppe du signal.
2. Comparaison avec un seuil fixe en prenant en compte un temps de maintien de la parole.
3. Détermination des trames au-dessus du seuil (DAV = 1 pour les trames actives) et en-dessous (DAV = 0 pour le silence). Le seuil est fixé à 16 dB en-dessous du niveau vocal actif moyen (puissance du signal sur les trames actives).

L'algorithme de découpage proposé est présenté dans la figure 4.11. La DAV est tout d'abord appliquée au signal de référence $x[t]$ de PESQM, ce qui permet d'en déterminer le niveau de parole actif, noté $ActLev_x$. Ce niveau est ensuite utilisé comme niveau actif de référence pour la détection d'activité vocale du signal dégradé $y[t]$ de PESQM. En effet, le signal d'écho éventuellement présent dans $y[t]$ ne doit pas être détecté comme de la parole, ni confondu avec le signal de parole du participant distant : la détection d'activité vocale doit donc être effectuée avec un seuil approprié pour discriminer le signal d'écho du signal de parole

du participant distant. À l'heure actuelle, le seuil de détection de la parole est fixé à 16 dB en-dessous du niveau actif moyen du signal, tel que décrit dans la Recommandation P.56. La DAV est donc capable de distinguer l'écho du signal de parole venant du côté distant si l'écho est atténué de plus de 16 dB par rapport au signal de référence $x[t]$. Une alternative pour déterminer ce seuil consisterait à utiliser un détecteur d'écho, tel que celui breveté par Barriac et Gilloire dans [Barriac et Gilloire 1996], qui adapterait le seuil du détecteur d'activité vocale en fonction du niveau d'atténuation de l'écho. La DAV appliquée à $y[t]$ permet ainsi de détecter les trames de parole du participant distant comme de la parole (DAV = 1) et de considérer les silences et l'écho éventuel comme du silence (DAV = 0). À partir des vecteurs d'activité vocale dav_x et dav_y des signaux $x[t]$ et $y[t]$, l'algorithme détermine le vecteur des trames durant lesquelles seul $x[t]$ est actif (*i.e.* seul le participant proche parle), noté dav_x^{seul}. Les indices de début et de fin des parties durant lesquelles seul $x[t]$ est actif et ceux des parties durant lesquelles $y[t]$ est actif sont calculés. Enfin, à partir des indices de fin des parties de signaux durant lesquelles seul $x[t]$ est actif et des indices de début des parties de signaux durant lesquelles $y[t]$ est actif, l'algorithme sélectionne les morceaux de signaux durant lesquels seul $x[t]$ est actif et dont la durée est comprise entre 8 et 30 secondes.

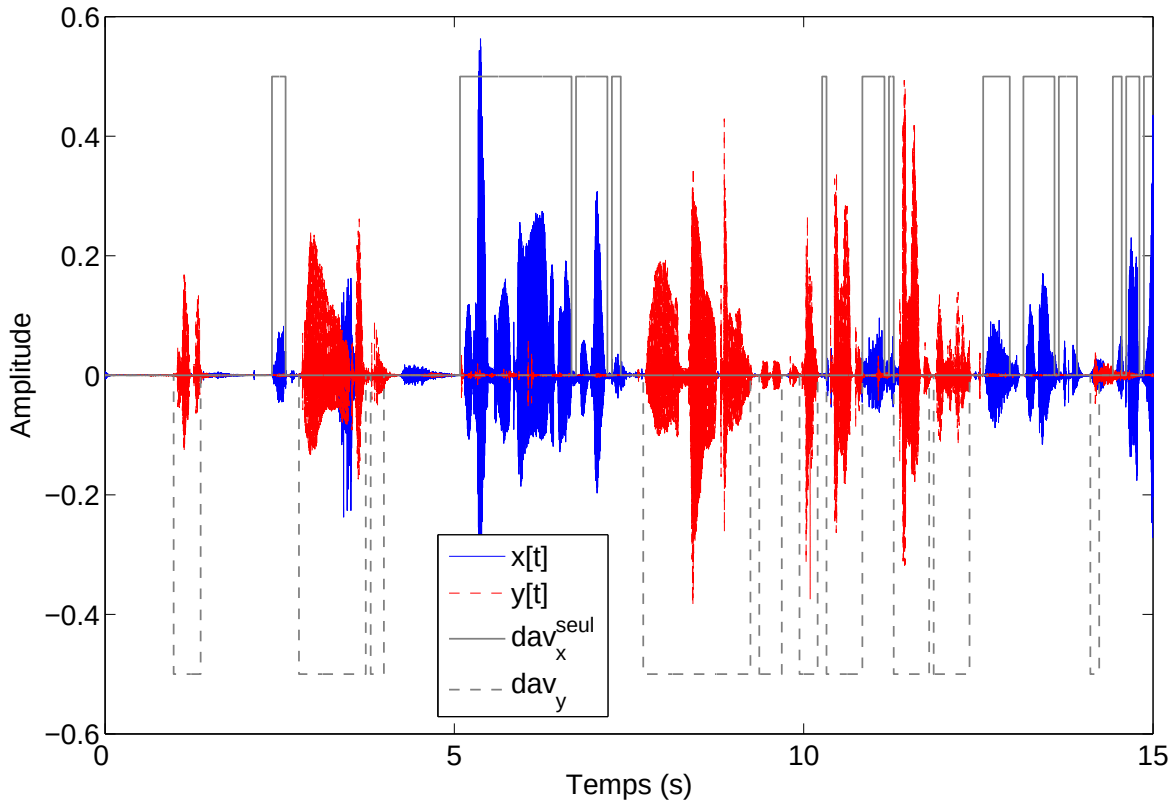
La figure 4.12 présente deux exemples d'application de cet algorithme à des signaux de conversation enregistrés dans le test 1 (*cf.* chapitre 3) dans des conditions respectivement sans écho et avec écho (atténuation = 25 dB, délai = 400 ms). Dans le cas sans écho, les périodes durant lesquelles seul $x[t]$ est actif et les périodes durant lesquelles $y[t]$ est actif sont correctement délimitées par l'algorithme de découpage. Dans le cas avec écho, l'algorithme fait de plus la distinction entre les trames de parole du participant distant et les trames d'écho dans le signal dégradé $y[t]$, ainsi les trames durant lesquelles $x[t]$ est actif et le signal d'écho est présent sont considérées comme des trames où seul $x[t]$ est actif. Une étude expérimentale plus poussée montre que l'algorithme fonctionne correctement et fournit un signal dégradé « propre » (*i.e.* ne contenant que l'écho éventuel et pas le signal de parole du participant distant). Les indices de début et de fin des vecteurs dav_x^{seul} et dav_y sont ensuite utilisés pour sélectionner les morceaux de signaux $x[t]$ et $y[t]$ à appliquer en entrée de PESQM.

Conclusion

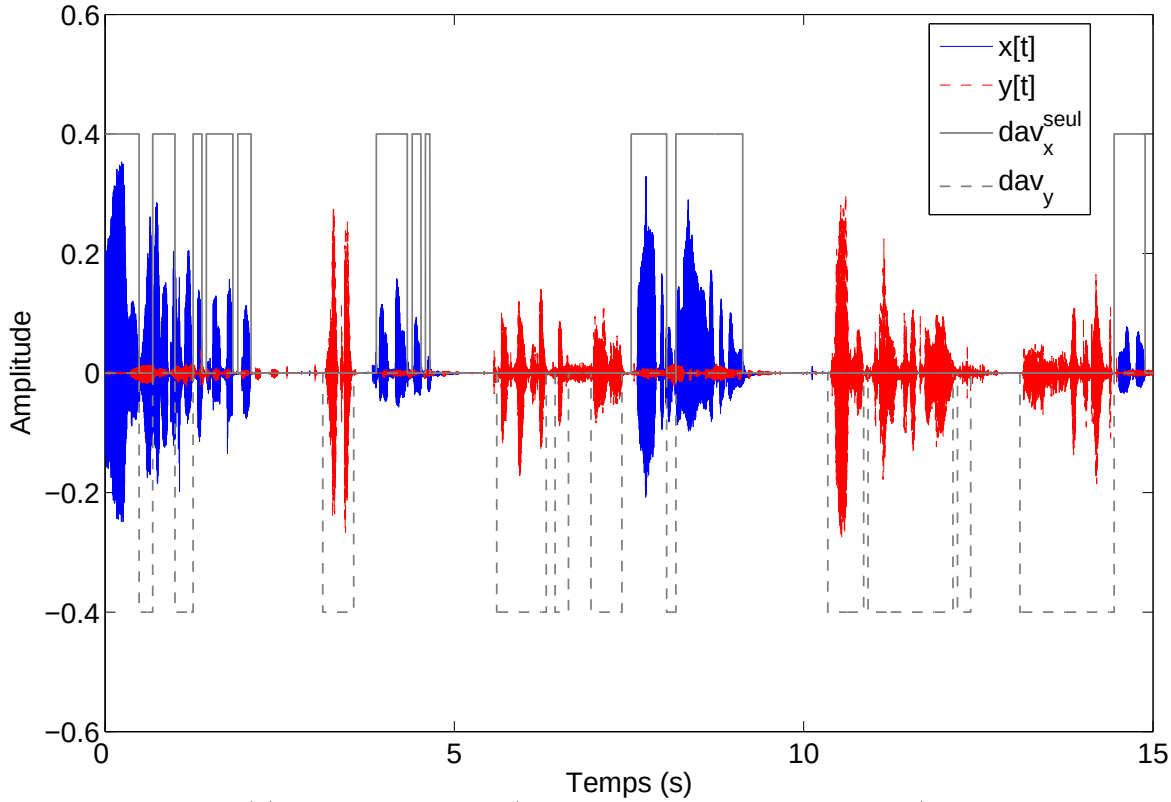
Dans ce chapitre, nous avons présenté des outils nécessaires au modèle objectif pour fonctionner intégralement avec deux types de signaux (signaux de test et signaux de conversation).

Tout d'abord, le modèle objectif de locution PESQM proposé dans [Appel et Beerends 2002] a été implémenté et optimisé sur les notes subjectives récoltées lors d'un test de locution. Celui-ci portait sur différentes dégradations affectant la qualité de locution (écho, amplification de l'effet local, bruit, pertes de paquets). L'analyse du test a montré que, contrairement aux autres dégradations et aux objectifs du test, l'amplification de l'effet local a eu très peu d'effet sur le jugement des participants. Ceci se traduit par une corrélation élevée entre les notes subjectives et les scores PESQM pour les conditions où l'effet local est à son niveau par défaut, et par une corrélation plus faible pour les conditions avec amplification de l'effet local. L'optimisation et la validation de PESQM ont donc été effectuées pour les conditions où l'effet local est à son niveau par défaut en prenant une moitié des signaux enregistrés comme base d'optimisation et l'autre moitié comme base de validation. Les scores PESQM obtenus avec les paramètres optimaux $p_{opt} = 1.1$ et $q_{opt} = 2.0$ sont très bien corrélés avec les notes subjectives de locution ($|r_{optim}| = 0.8697$ et $|r_{valid}| = 0.8589$). Les scores PESQM sont ensuite transformés en notes MOS grâce à une fonction de mapping unique, déterminée pour les conditions avec écho seul et avec écho + pertes de paquets et aboutissant à une corrélation de Pearson $r = 0.9165$ entre les notes subjectives et objectives de locution.

La seconde partie de ce chapitre a décrit un outil de traitement permettant d'utiliser les



(a) Condition sans écho



(b) Condition avec écho (atténuation = 25 dB, délai = 400 ms)

Figure 4.12 : *Exemples d'application de l'algorithme de découpage des signaux à des signaux de conversation enregistrés durant le test 1*

signaux enregistrés dans des réseaux réels ou lors de conversations avec les modèles intrusifs PESQ et PESQM. En effet, ceux-ci ont des contraintes sur la durée, le taux d'activité vocale, etc. de leurs signaux d'entrée que les signaux réels peuvent ne pas respecter. Cet outil est basé sur une détection d'activité vocale qui :

- pour le modèle PESQ, sélectionne simplement des morceaux de signaux ayant des durées et des taux d'activité vocale appropriés,
- pour le modèle PESQM, sélectionne des morceaux de signaux permettant d'obtenir un signal dégradé « propre » (*i.e.* ne contenant que l'écho éventuel et pas la parole du locuteur distant) et ayant des durées et des taux d'activité vocale appropriés.

Dans le chapitre suivant, le modèle objectif complet est appliqué aux différents types de signaux (de test ou de conversation) enregistrés lors des quatre tests subjectifs présentés dans le chapitre 3. Les performances du modèle objectif sont évaluées en comparant les notes fournies par le modèle aux notes subjectives.

Chapitre 5

Résultats et performances du modèle objectif

Introduction

Dans ce chapitre, le modèle objectif proposé est appliqué aux signaux enregistrés lors des quatre tests subjectifs mis en œuvre pendant la thèse. Comme l'illustre la figure 2.1(b) du chapitre 2, l'évaluation objective de la qualité d'écoute est possible avec plusieurs modèles s'appuyant sur différents types de mesure (mesure intrusive basée sur les signaux avec P.862 (PESQ), mesure non intrusive basée sur les signaux avec P.563 et mesure non intrusive paramétrique avec P.564). Cependant, l'évaluation objective de la qualité de locution n'est possible, à notre connaissance, qu'avec le modèle PESQM, intrusif et basé sur les signaux. Nous présenterons donc les résultats obtenus en utilisant les notes objectives d'écoute et de locution fournies par les modèles intrusifs basés sur les signaux correspondants, à savoir PESQ et PESQM, respectivement. De plus, comme cela a été mentionné dans le chapitre 4, le délai du système testé sera supposé connu pour calculer la note objective de conversation (MOS_{CONV}) par une combinaison linéaire appliquée aux notes objectives de qualité de locution (MOS_{PESQM}) et d'écoute (MOS_{PESQ}) et à la valeur connue du délai unidirectionnel ($d\acute{e}lai$, en ms)

$$\begin{aligned} MOS_{CONV} &= 0.4059 \times MOS_{PESQM} + 0.5519 \times MOS_{PESQ} \\ &\quad - 1.7376 \times \max(0, d\acute{e}lai - 400) + 0.1710. \end{aligned} \tag{5.1}$$

Les deux types de signaux (de test et réels) peuvent être utilisés par le modèle de conversation, les signaux de test correspondant aux signaux enregistrés pendant les phases d'écoute et de locution des tests subjectifs et les signaux réels à ceux enregistrés pendant la phase de conversation de ces mêmes tests. L'application aux signaux de test, présentée dans la première partie de ce chapitre, permet d'évaluer les performances du modèle objectif, constitué des modèles PESQ et PESQM et de la combinaison linéaire présentée dans l'équation 5.1. L'application du modèle objectif aux signaux réels, exposée dans la deuxième partie du chapitre, fournit en supplément les performances de l'outil de traitement des signaux présenté dans le chapitre 4.

Le chapitre 1 a mis en évidence un impact du délai sur la qualité de conversation variable en fonction de l'interactivité de la communication testée. La troisième partie de ce chapitre proposera ainsi une étude, préliminaire, de l'interactivité effectuée à partir des signaux de conversation enregistrés durant les tests subjectifs.

5.1 Application à des signaux de test

Les signaux de test ont été enregistrés pendant les phases d'écoute et de locution des quatre tests subjectifs présentés dans le chapitre 3. Il s'agit de signaux au format wav, échantillonnés à 8 kHz et codés sur 16 bits, respectant les contraintes imposées par la Recommandation P.862.3 pour fonctionner avec le modèle PESQ :

- une longueur entre 8 et 30 secondes,
- un taux d'activité vocale entre 40 et 80% (mesuré d'après la Recommandation P.56 [UIT-T Rec. P.56 1993]),
- un niveau d'activité vocale de -30 dBov.

Le système testé est présenté dans la figure 2.1(a) du chapitre 2, avec les signaux émis et reçu de chaque côté. Quatre signaux sont disponibles par paire de participants et par phase (phase 1 : locution de A/écoute de B, phase 2 : locution de B/écoute de A, phase 3 : conversation libre). Afin d'évaluer la qualité de conversation du côté A, pour une condition donnée, les étapes suivantes sont nécessaires :

1. Les signaux émis par A (①) et reçu par A (②), enregistrés pendant la phase de locution de A, sont utilisés tels quels comme signaux de référence et dégradé de PESQM, respectivement. Une note objective de qualité de locution MOS_{PESQM} est obtenue.
2. Les signaux émis par B (③) et reçu par A (②), enregistrés pendant la phase d'écoute de A, sont utilisés tels quels comme signaux de référence et dégradé de PESQ, respectivement. Une note objective de qualité d'écoute MOS_{PESQ} est obtenue.
3. La note objective de qualité de conversation MOS_{CONV} est déterminée à partir de MOS_{PESQM}, MOS_{PESQ} et de la valeur du délai unidirectionnel *délai* (supposée connue) selon l'équation 5.1.

Pour des signaux d'une durée de 11 secondes, l'ensemble de ces trois étapes nécessite environ 3 secondes de calcul (processeur à 2 GHz, 1 Go de mémoire vive). Pour chaque condition de test, une note objective de qualité de locution MOS_{PESQM}, une note objective de qualité d'écoute MOS_{PESQ} et une note objective de qualité de conversation MOS_{CONV} sont disponibles pour chaque participant. La note moyenne de qualité de locution $\overline{MOS}_{PESQM}$, la note moyenne de qualité d'écoute $\overline{MOS}_{PESQ}$ et la note moyenne de qualité de conversation $\overline{MOS}_{CONV}$ sont calculées par condition à partir de l'ensemble des notes objectives individuelles correspondantes.

Les performances des trois modèles objectifs d'écoute (PESQ), de locution (PESQM) et de conversation (CONV) sur les 45 conditions étudiées lors des quatre tests subjectifs (test 1 : conditions 1-8, test 2 : conditions 9-17, test 3 : conditions 18-24, test 4 : conditions 25-45) sont présentées dans les figures 5.1, 5.2 et 5.3, respectivement, sous la forme d'un mapping entre les notes moyennes subjectives et objectives, d'un histogramme de la distribution cumulative de l'erreur absolue entre notes moyennes subjectives et objectives, et d'un graphe des notes moyennes subjectives et objectives avec intervalles de confiance à 95% en fonction des conditions étudiées. Les performances des trois modèles objectifs sont également détaillées test par test dans le tableau 5.1, fournissant le coefficient de corrélation de Pearson r, le coefficient de corrélation de Spearman r_s et l'erreur absolue moyenne (EAM, exprimée en MOS) entre les notes subjectives et objectives.

5.1.1 Performances du modèle objectif de qualité d'écoute (PESQ)

D'après la figure 5.1(a), la corrélation entre les notes moyennes d'écoute subjectives et objectives est élevée ($r = 0.923$) mais l'erreur absolue moyenne est relativement élevée ($EAM = 0.306$ MOS). La distribution de l'erreur absolue, présentée dans la figure 5.1(b),

Test	Critère de performance	Écoute	Locution	Conversation sans détection du bruit	Conversation avec détection du bruit
Tous tests	r	0.923	0.695	0.860	0.904
	r_s	0.826	0.576	0.862	0.901
	EAM	0.306	0.591	0.236	0.206
Test 1	r	-0.093	0.975	0.951	0.951
	r_s	-0.153	0.539	0.810	0.810
	EAM	0.382	0.279	0.197	0.197
Test 2	r	0.941	0.128	0.886	0.886
	r_s	0.833	0.294	0.862	0.862
	EAM	0.271	0.619	0.246	0.246
Test 3	r	0.917	0.304	0.706	0.893
	r_s	0.893	0.180	0.865	0.883
	EAM	0.273	1.418	0.446	0.264
Test 4	r	0.973	0.947	0.916	0.918
	r_s	0.899	0.880	0.923	0.921
	EAM	0.304	0.422	0.176	0.174

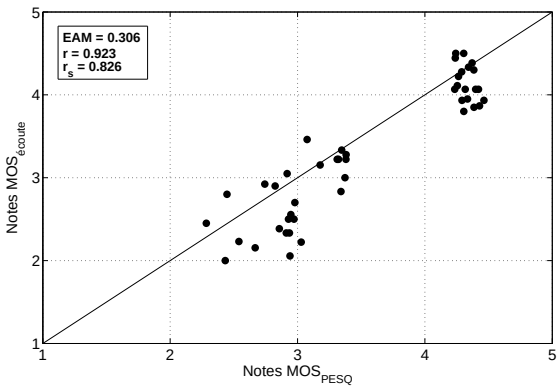
Tableau 5.1 : *Performances des modèles objectifs des qualités d'écoute (PESQ), de locution (PESQM) et de conversation (CONV sans ou avec détection du bruit) pour les différents tests - Signaux de test*

montre que 95.6% des notes moyennes d'écoute objectives diffèrent de moins de 0.625 MOS des notes subjectives et 100% diffèrent de moins de 1 MOS. Les performances de PESQ sont analysées test par test, grâce à la figure 5.1(c). Tout d'abord, les intervalles de confiance à 95% des notes objectives sont faibles par rapport à ceux des notes subjectives pour tous les tests, indiquant la faible variabilité du modèle PESQ pour une même condition. Pour le test 1 sur l'écho et le délai (conditions 1-8), les notes moyennes objectives et subjectives sont quasiment constantes mais le modèle PESQ surestime la qualité d'environ 0.38 MOS. Pour le test 2 sur les pertes de paquets et le bruit transmis (conditions 9-17), il y a une très bonne corrélation et une erreur moyenne faible entre les notes subjectives et objectives. Pour le test 3 sur le bruit (conditions 18-24), la corrélation est élevée, et l'erreur est faible pour les notes de qualité élevées et plus importante pour les notes de qualité basses. Pour le test 4 sur l'écho, le délai et les pertes de paquets (conditions 25-45), PESQ semble moins pénaliser les pertes de paquets à un taux de 10% que les participants aux tests.

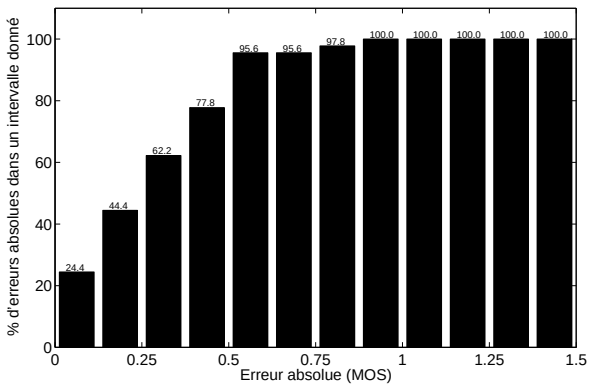
Pour résumer (*cf.* tableau 5.1), le modèle PESQ permet d'estimer la qualité d'écoute étudiée lors des quatre tests subjectifs avec une corrélation et une erreur absolue moyenne conformes aux performances annoncées dans [Beerends *et al.* 2002], et a tendance à surestimer les notes de qualité faibles.

5.1.2 Performances du modèle objectif de qualité de locution (PESQM)

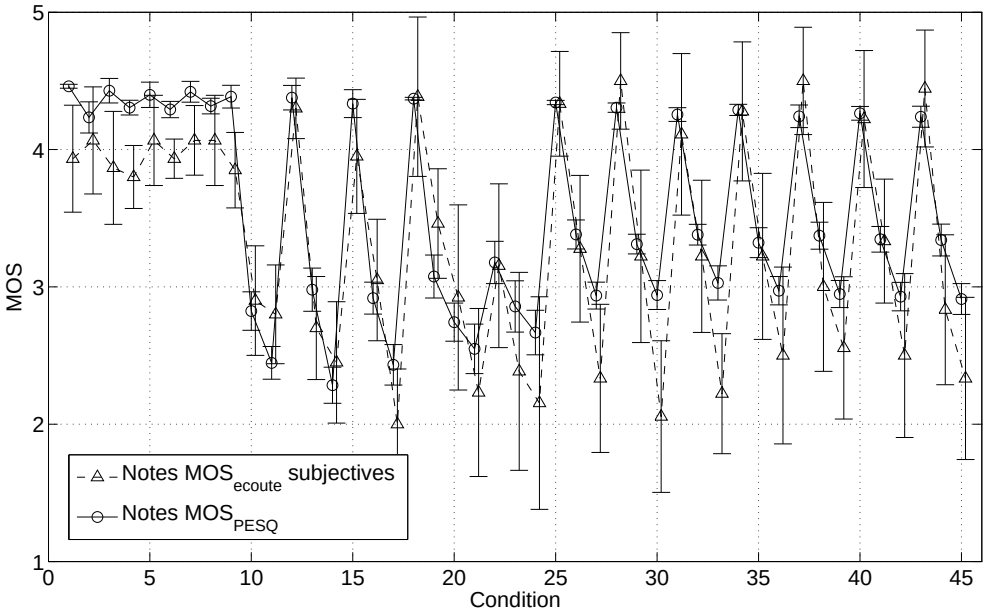
La figure 5.2(a) montre que les performances de PESQM sur l'ensemble des conditions testées sont faibles ($r = 0.695$ et $EAM = 0.591$ MOS). La distribution de l'erreur absolue, présentée dans la figure 5.2(b), confirme ces mauvaises performances, puisque seules 91.1% des erreurs absolues entre notes moyennes de locution subjectives et objectives sont inférieures à 1.5 MOS. La figure 5.2(c) permet d'analyser les performances de PESQM test par test. Pour l'ensemble des quatre tests, les intervalles de confiance à 95% des notes de locution subjectives et objectives sont importants et du même ordre de grandeur. Ceci reflète la difficulté d'évaluer la qualité de locution (subjectivement et objectivement) évoquée dans le chapitre 1, puisque chacun utilise sa propre voix comme référence, contrairement à un test d'écoute où les signaux de référence et dégradé sont les mêmes pour tous les participants. Pour le test 1 sur le délai et l'écho (conditions 1-8), les performances de PESQM sont élevées même si le modèle a tendance à sous-estimer la qualité de locution pour les valeurs de délai unidirectionnel de 400 et 600 ms en présence d'écho. Pour le test 2 sur les pertes de paquets et le bruit transmis (conditions 9-17), les notes objectives diminuent quand le niveau du bruit distant augmente contrairement aux notes subjectives qui restent comprises entre 4 et 4.5 MOS. L'explication la plus probable



(a) Mapping entre notes moyennes d'écoute subjectives et objectives (PESQ)



(b) Distribution cumulative de l'erreur absolue entre notes moyennes d'écoute subjectives et objectives (PESQ)

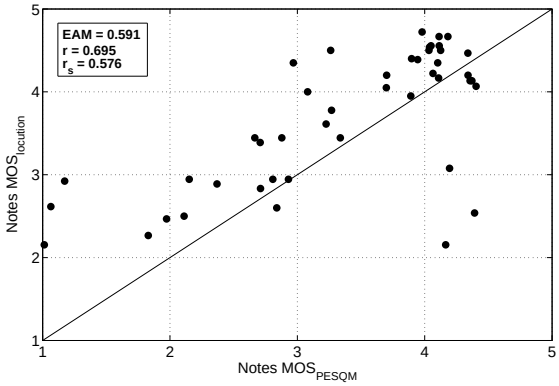


(c) Notes moyennes d'écoute subjectives et objectives (PESQ) avec intervalles de confiance à 95%

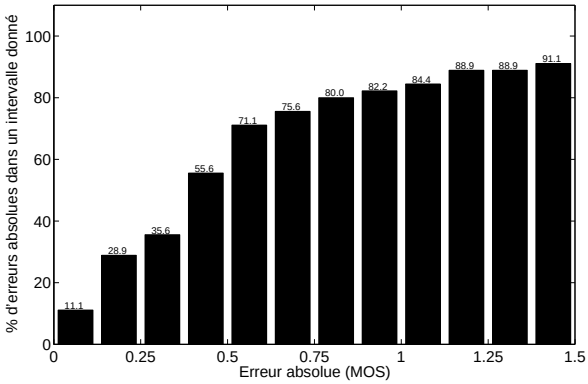
Figure 5.1 : *Performances de PESQ - Signaux de test*

est la présence dans les signaux dégradés de bruit transmis à un niveau faible, pris en compte par PESQM et pas par les participants. Les résultats pour le test 3 sur le bruit (conditions 18-24) sont médiocres, avec une corrélation faible et une erreur très élevée entre les notes de locution subjectives et objectives. Le modèle PESQM ne fonctionne pas en présence de bruit seul, ce qui s'explique par sa structure intrinsèque. En effet, comme cela est expliqué en détail dans l'annexe C, le modèle PESQM intègre un module de suppression du bruit. Celui-ci prend en compte l'influence du masquage de l'écho par un bruit, qui peut ainsi améliorer la qualité de locution, à condition que le bruit reste inférieur à un certain seuil. Cependant, le modèle PESQM, tel qu'il est implémenté, ne peut pas estimer la qualité de locution en présence de bruit seul (en l'absence d'écho), ce qui explique ses mauvaises performances dans ce test. Pour le test 4 sur l'écho, le délai et les pertes de paquets (conditions 25-45), la corrélation entre notes subjectives et objectives est élevée. L'erreur est malgré tout élevée, le modèle ayant tendance à sous-estimer la qualité de locution, en particulier pour un délai de 400 ms (conditions 40-45) comme ce qui avait été constaté pour le test 1.

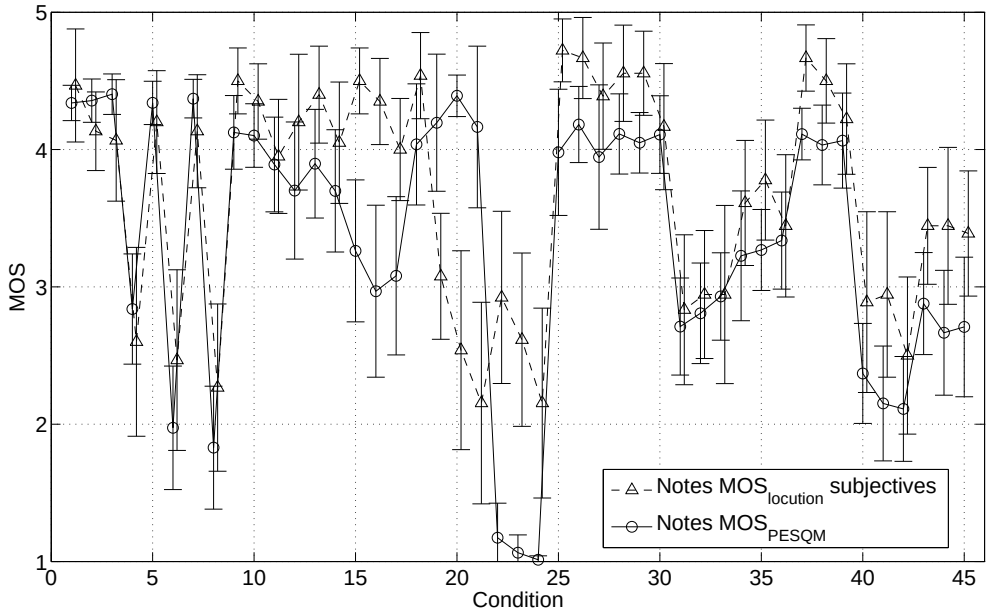
Pour résumer (*cf.* tableau 5.1), le modèle PESQM permet de bien estimer la qualité de locution dans les conditions avec écho seul et avec écho et pertes de paquets (pour lesquelles



(a) Mapping entre notes moyennes de locution subjectives et objectives (PESQM)



(b) Distribution cumulative de l'erreur absolue entre notes moyennes de locution subjectives et objectives (PESQM)



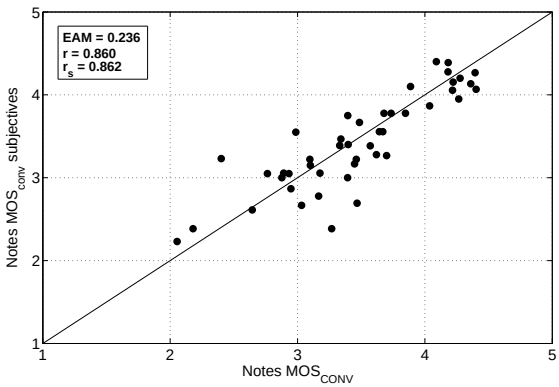
(c) Notes moyennes de locution subjectives et objectives (PESQM) avec intervalles de confiance à 95%

Figure 5.2 : *Performances de PESQM - Signaux de test*

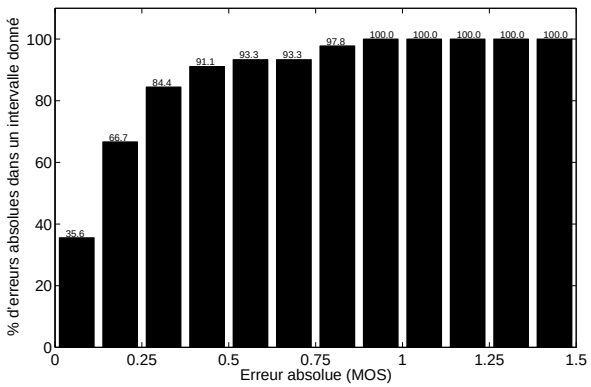
il a été optimisé dans le chapitre 4), mais ne fonctionne pas en présence de bruit seul.

5.1.3 Performances du modèle objectif de qualité de conversation (CONV)

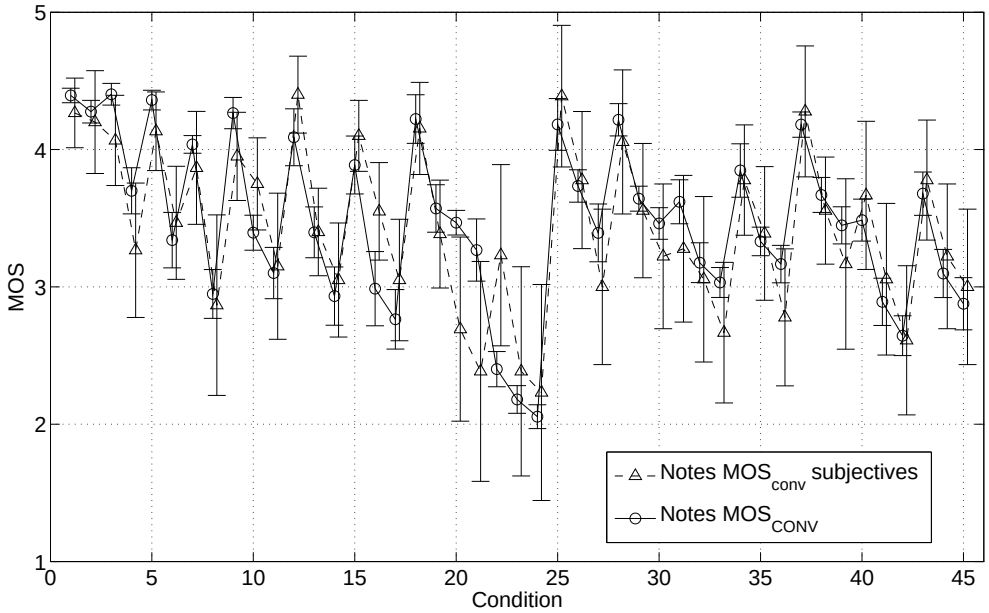
Les performances du modèle objectif de qualité de conversation sur l'ensemble des conditions de test fournies dans la figure 5.3(a) sont bonnes ($r = 0.860$ et $EAM = 0.236$ MOS). La distribution de l'erreur absolue, présentée dans la figure 5.3(b), montre que 91.1% des notes moyennes de conversation objectives diffèrent de moins de 0.5 MOS des notes subjectives et que 100% diffèrent de moins de 1 MOS. Sur l'ensemble des quatre tests, les intervalles de confiance à 95% des notes objectives sont faibles comparés à ceux des notes subjectives, comme l'indique la figure 5.3(c). La corrélation et l'erreur entre notes moyennes de conversation subjectives et objectives sont bonnes, sauf pour le test 3 sur le bruit (*cf.* tableau 5.1). Les erreurs élevées constatées avec les modèles PESQ et PESQM sont en grande partie compensées par les coefficients de régression ($\alpha = 0.4059$ et $\beta = 0.5519$). Cependant, les performances médiocres de PESQM dans les conditions avec bruit seul (test 3, conditions 18-24) se répercutent fortement sur les performances du modèle de conversation pour le test 3 ($r = 0.706$ et $EAM = 0.446$ MOS).



(a) Mapping entre notes moyennes de conversation subjectives et objectives (CONV)



(b) Distribution cumulative de l'erreur absolue entre notes moyennes de conversation subjectives et objectives (CONV)



(c) Notes moyennes de conversation subjectives et objectives (CONV) avec intervalles de confiance à 95%

Figure 5.3 : *Performances du modèle de conversation - Signaux de test*

Afin d'améliorer les performances du modèle de conversation dans ces conditions, la première solution consisterait à modifier le modèle PESQM pour qu'il aboutisse à de meilleures performances dans les conditions avec bruit seul. Cependant, cette solution suppose de disposer de données subjectives étudiant l'impact du bruit seul sur la qualité de locution, pour pouvoir modifier et optimiser PESQM. Ne disposant pas de telles données (indépendantes de celles utilisées pour la construction du modèle objectif), nous proposons une seconde solution pour le court terme. Dans le chapitre 3, une relation de régression linéaire multiple F_i du type

$$\widehat{MOS}_{conversation} = \alpha \times MOS_{locution} + \beta \times MOS_{écoute} + \delta \times \max(0, délai - délai_{seuil}) + \gamma \quad (5.2)$$

a été déterminée pour chaque test i. Pour le test 3 sur le bruit, en particulier, les coefficients de l'équation de régression sont $\alpha = 0$, $\beta = 0.864$, $\delta = 0$ et $\gamma = 0.367$. Comme $\alpha = 0$, la note de qualité de locution n'intervient pas dans l'estimation de la note de qualité de conversation. Appliquée aux notes objectives de qualité, l'équation 5.3 est obtenue, qui ne fait ainsi intervenir que la note objective de qualité d'écoute (fournie par PESQ) et la constante γ pour calculer la note objective de qualité de conversation, sans avoir besoin de la note de qualité de locution

(fournie par PESQM).

$$MOS_{CONV} = 0.864 \times MOS_{PESQ} + 0.367. \tag{5.3}$$

La solution proposée consiste à appliquer l'équation 5.3 dans les conditions du test 3 sur le bruit et l'équation 5.1 dans les autres conditions. Les conditions du test 3 se distinguent des autres conditions par des rapports signal-à-bruit segmentaux moyens $RSBseg$ (calculés du côté réception du bruit et présentés dans le tableau 3.21 du chapitre 3) faibles ($RSBseg < 17$ dB). Un seuil de rapport signal-à-bruit segmental, calculé sur le signal reçu durant la phase d'écoute, permettrait de détecter les conditions avec bruit et de choisir l'équation de régression adéquate. Afin de déterminer ce seuil, l'ensemble des fichiers (enregistrés du côté réception du bruit pendant la phase d'écoute) est divisé en une base d'apprentissage composée de 157 fichiers enregistrés dans les conditions sans bruit (tests 1, 2, 3 et 4) et de 36 fichiers enregistrés dans les conditions avec bruit (test 3), et une base de validation composée du reste des fichiers (517 fichiers sans bruit et 41 fichiers avec bruit). Pour l'apprentissage, le rapport signal-à-bruit segmental de chaque fichier est calculé selon l'équation 3.2 du chapitre 3. Les valeurs obtenues sont présentées dans la figure 5.4, en fonction du type de fichier (sans ou avec bruit). D'après ces valeurs, le seuil de $RSBseg$ retenu est égal à 20 dB et permet de discriminer sans erreur les deux types de fichiers (sans ou avec bruit). Appliqué à la base de validation, ce seuil de 20 dB aboutit à un taux de fausse détection de 0% et un taux de non détection de 2.4% (1 non détection sur 41 fichiers).

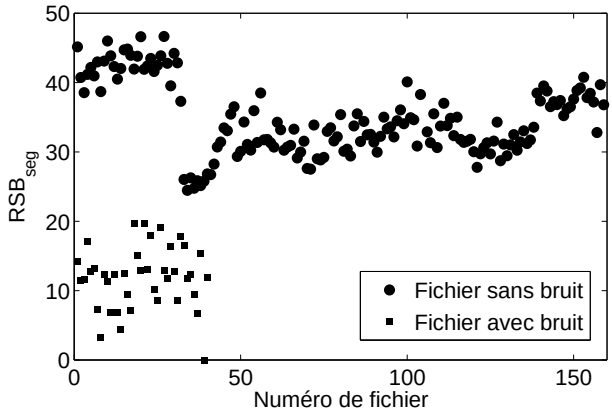


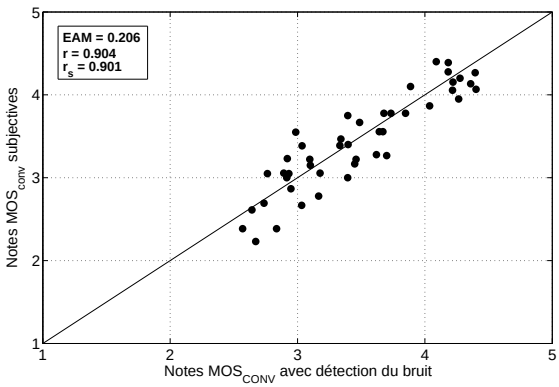
Figure 5.4 : *Calcul du rapport signal-à-bruit segmental sur les fichiers (sans ou avec bruit) de la base d'apprentissage*

5.1.4 Performances du modèle objectif de qualité de conversation (CONV) avec détection du bruit

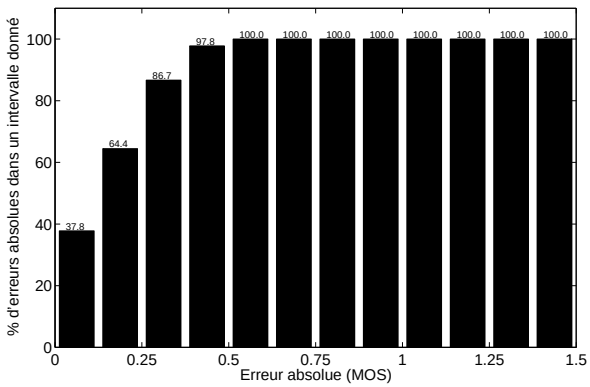
Le rapport signal-à-bruit segmental est calculé sur le signal dégradé de PESQ (*i.e.* le signal reçu pendant la phase d'écoute) pour chaque condition et chaque participant. Pour des signaux d'une durée de 11 secondes, ce calcul supplémentaire augmente d'environ 1 seconde le temps de traitement du modèle objectif (soit 4 secondes). Sur la base des conclusions présentées dans le paragraphe 5.1.3, la règle appliquée pour le choix de l'équation de régression est donc la suivante :

- si $RSBseg \geq 20$ dB, la note objective de qualité de conversation est obtenue à partir de l'équation 5.1,
- sinon, la note objective de qualité de conversation est obtenue à partir de l'équation 5.3.

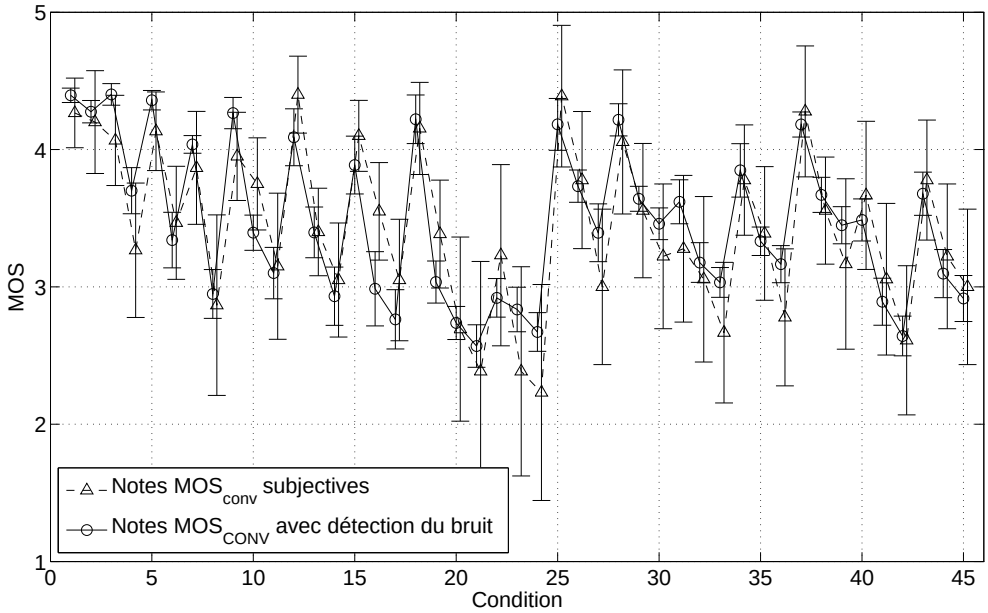
Les performances du modèle objectif de qualité de conversation avec détection du bruit sur l'ensemble des conditions de test sont fournies dans la figure 5.5(a). Elles sont élevées ($r = 0.904$ et $EAM = 0.206$ MOS) et meilleures que celles obtenues avec le modèle objectif



(a) Mapping entre notes moyennes de conversation subjectives et objectives (CONV)



(b) Distribution cumulative de l'erreur absolue entre notes moyennes de conversation subjectives et objectives (CONV)



(c) Notes moyennes de conversation subjectives et objectives (CONV) avec intervalles de confiance à 95%

Figure 5.5 : *Performances du modèle de conversation avec détection du bruit - Signaux de test*

sans détection du bruit. La distribution de l'erreur absolue, présentée dans la figure 5.5(b), montre que 100% des notes moyennes de conversation objectives diffèrent de moins de 0.625 MOS des notes subjectives. Les résultats fournis dans le tableau 5.1 indiquent que, logiquement, seules les performances pour le test 3 sur le bruit changent et s'améliorent par rapport au modèle objectif sans détection du bruit en atteignant $r = 0.893$ et $EAM = 0.264$ MOS au lieu de $r = 0.706$ et $EAM = 0.446$ MOS.

5.1.5 Synthèse

Le modèle objectif de conversation, constitué de PESQ et de PESQM, appliqué aux signaux de test (*i.e.* aux signaux enregistrés pendant les phases d'écoute et de locution des tests subjectifs effectués au cours de la thèse) atteint des performances élevées. Celles-ci sont meilleures lorsque les conditions avec bruit, problématiques pour le modèle de locution PESQM, sont détectées grâce à un seuil du rapport signal-à-bruit segmental. Ce seuil a été déterminé sur une base d'apprentissage de signaux de test et fixé à 20 dB. Cette solution, bien que suffisante à court terme, ne saurait remplacer une amélioration du modèle PESQM dans les conditions bruitées, grâce à d'autres tests subjectifs de locution.

Il est important de noter que les signaux utilisés ici sont des signaux enregistrés dans des conditions réelles et non parfaites. Malgré les instructions clairement formulées, pendant les phases d'écoute, certains auditeurs acquiesçaient ou encourageaient le locuteur à continuer à parler. Ainsi, les signaux reçus par le locuteur pendant sa phase de locution, supposés ne contenir que l'éventuel signal d'écho, peuvent contenir des occurrences de parole de l'auditeur (*e.g.* 'hum', 'oui', ...) qui ont pu perturber le modèle PESQM dans son évaluation de la qualité de locution.

5.2 Application à des signaux de conversation

Comme cela a été évoqué dans le chapitre 4, pour pouvoir appliquer le modèle objectif de conversation à des signaux enregistrés lors de communications réelles ou lors de tests subjectifs de conversation extérieurs à la thèse (ne contenant pas, en général, de phases d'écoute et de locution), il est nécessaire que le modèle puisse fonctionner directement sur les signaux de conversation.

Le modèle objectif est ainsi appliqué aux signaux de conversation enregistrés pendant la phase de conversation des quatre tests subjectifs effectués au cours de la thèse. Il s'agit de signaux au format wav, échantillonnés à 8 kHz et codés sur 16 bits. Le système testé est présenté dans la figure 2.1(a) du chapitre 2, avec les signaux émis et reçu de chaque côté (*i.e.* quatre signaux par paire de participants). Afin d'évaluer la qualité de conversation du côté A, pour une condition donnée, les étapes suivantes sont nécessaires :

1. L'outil de traitement des signaux présenté dans le chapitre 4 est appliqué aux signaux de conversation émis par A (①) et reçu par A (②). m morceaux de signaux ① et ② sont découpés. Les morceaux de signaux $(①_i,②_i)_{1\leq i\leq m}$ sont utilisés comme signaux de référence et dégradé de PESQM, respectivement. Une note objective de qualité de locution MOS_{PESQM_i} par couple de morceaux est obtenue. La note objective de qualité de locution MOS_{PESQM} pour la condition donnée est la moyenne des m notes MOS_{PESQM_i}.
2. L'outil de traitement des signaux est appliqué aux signaux de conversation émis par B (③) et reçu par A (②). n morceaux de signaux ③ et ② sont découpés. Les morceaux de signaux $(③_j,②_j)_{1\leq j\leq n}$ sont utilisés comme signaux de référence et dégradé de PESQ, respectivement. Une note objective de qualité d'écoute MOS_{PESQ_j} par couple de morceaux est obtenue. La note objective de qualité d'écoute MOS_{PESQ} pour la condition donnée est la moyenne des n notes $note_{PESQ_j}$.
3. La note objective de qualité de conversation MOS_{CONV} est déterminée à partir de MOS_{PESQM}, MOS_{PESQ} et de la valeur du délai unidirectionnel *délai* (supposée connue) selon l'équation 5.1.

En moyenne sur l'ensemble des signaux traités, le nombre de morceaux m pour PESQM est 7 et le nombre de morceaux n pour PESQ est 4. Pour des signaux d'une durée de 120 secondes, l'ensemble de ces étapes nécessite environ 35 secondes de calcul (processeur à 2 GHz, 1 Go de mémoire vive). Pour chaque condition de test et chaque participant, une note objective de qualité de locution MOS_{PESQM}, une note objective de qualité d'écoute MOS_{PESQ} et une note objective de qualité de conversation MOS_{CONV} sont disponibles. La note moyenne de qualité de locution $\overline{MOS}_{PESQM}$, la note moyenne de qualité d'écoute $\overline{MOS}_{PESQ}$ et la note moyenne de qualité de conversation $\overline{MOS}_{CONV}$ sont calculées par condition à partir de l'ensemble des notes objectives individuelles correspondantes.

Les performances des trois modèles objectifs d'écoute (PESQ), de locution (PESQM) et de conversation (CONV) (sans et avec détection du bruit) sur les 45 conditions étudiées lors des quatre tests subjectifs (test 1 : conditions 1-8, test 2 : conditions 9-17, test 3 : conditions 18-24, test 4 : conditions 25-45) sont présentées dans les figures 5.6, 5.7, 5.8 et 5.9, respectivement, sous la forme d'un mapping entre les notes moyennes subjectives et objectives,

Test	Critère de performance	Écoute	Locution	Conversation sans détection du bruit	Conversation avec détection du bruit
Tous tests	r	0.883	0.917	0.914	0.916
	r_s	0.870	0.849	0.893	0.895
	EAM	0.322	0.260	0.198	0.203
Test 1	r	-0.061	0.978	0.927	0.927
	r_s	-0.332	0.850	0.881	0.881
	EAM	0.269	0.139	0.146	0.146
Test 2	r	0.943	0.379	0.929	0.929
	r_s	0.867	0.538	0.879	0.879
	EAM	0.226	0.182	0.176	0.176
Test 3	r	0.913	0.829	0.951	0.935
	r_s	0.893	0.847	0.955	0.883
	EAM	0.269	0.462	0.177	0.209
Test 4	r	0.926	0.937	0.919	0.919
	r_s	0.918	0.949	0.891	0.891
	EAM	0.400	0.273	0.235	0.235

Tableau 5.2 : *Performances des modèles objectifs des qualités d'écoute (PESQ), de locution (PESQM) et de conversation (CONV sans ou avec détection du bruit) pour les différents tests - Signaux de conversation*

d'un histogramme de la distribution cumulative de l'erreur absolue entre notes moyennes subjectives et objectives, et d'un graphe des notes moyennes subjectives et objectives avec intervalles de confiance à 95% en fonction des conditions étudiées. Les performances des trois modèles objectifs sont également détaillées test par test dans le tableau 5.2, fournissant l'erreur absolue moyenne (EAM, exprimée en MOS), le coefficient de corrélation de Pearson r et le coefficient de corrélation de Spearman r_s entre les notes subjectives et objectives. Nous présentons les performances du modèle de conversation sans détection du bruit tout d'abord, puis avec afin d'étudier l'apport de la détection de bruit pour les signaux de conversation.

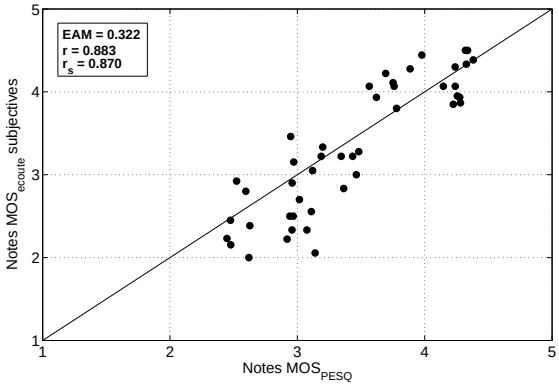
5.2.1 Performances du modèle objectif de qualité d'écoute (PESQ)

D'après la figure 5.6(a) le modèle PESQ atteint de bonnes performances sur l'ensemble des conditions avec un coefficient de corrélation de 0.883 et une erreur absolue de 0.322 MOS. La distribution de l'erreur absolue, présentée dans la figure 5.6(b), montre que 97.8% des notes moyennes d'écoute objectives diffèrent de moins de 0.75 MOS des notes subjectives et que 100% diffèrent de moins de 1.125 MOS. La figure 5.6(c) confirme l'observation faite sur les signaux de test, à savoir que la qualité d'écoute en présence de pertes de paquets à un taux de 10% (test 4, conditions 25-45) est surestimée par PESQ. Dans les conditions des trois autres tests, PESQ atteint de bonnes performances, comme le confirme le tableau 5.2.

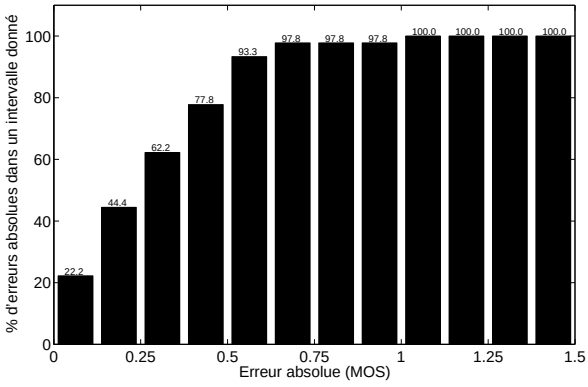
Par rapport aux performances obtenues sur les signaux de test, les performances de PESQ sur signaux de conversation sont légèrement inférieures ($r = 0.883$ et $EAM = 0.322$ MOS au lieu de $r = 0.923$ et $EAM = 0.306$ MOS), à cause de ses mauvaises performances sur les conditions du test 4. Cependant, les intervalles de confiance à 95% des notes d'écoute subjectives observées pour ce test sont particulièrement élevés.

5.2.2 Performances du modèle objectif de qualité de locution (PESQM)

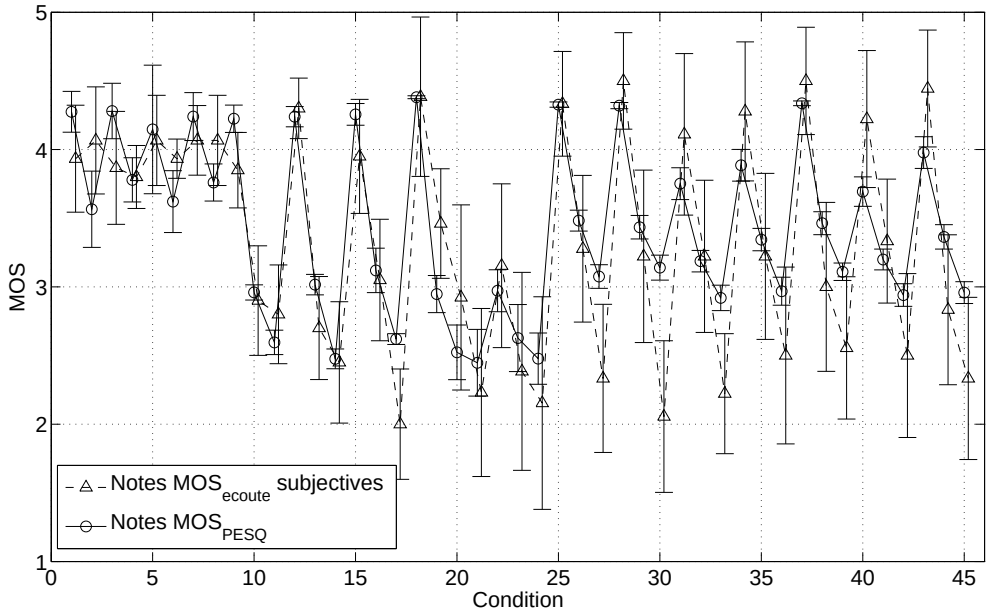
Le modèle PESQM atteint de bonnes performances ($r = 0.917$ et $EAM = 0.260$ MOS) sur l'ensemble des conditions comme le montre la figure 5.7(a). La distribution de l'erreur absolue, présentée dans la figure 5.7(b), montre que 95.6% des notes moyennes d'écoute objectives diffèrent de moins de 0.625 MOS des notes subjectives et que 100% diffèrent de moins de 1.125 MOS. Pour l'ensemble des quatre tests, la figure 5.7(c) et le tableau 5.2 indiquent que les notes moyennes de locution subjectives et objectives sont plus proches qu'avec les signaux de test. En effet, le découpage des signaux de conversation en morceaux permet d'estimer



(a) Mapping entre notes moyennes d'écoute subjectives et objectives (PESQ)



(b) Distribution cumulative de l'erreur absolue entre notes moyennes d'écoute subjectives et objectives (PESQ)



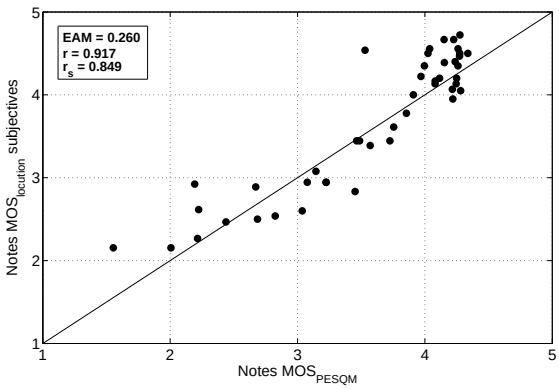
(c) Notes moyennes d'écoute subjectives et objectives (PESQ) avec intervalles de confiance à 95%

Figure 5.6 : *Performances de PESQ - Signaux de conversation*

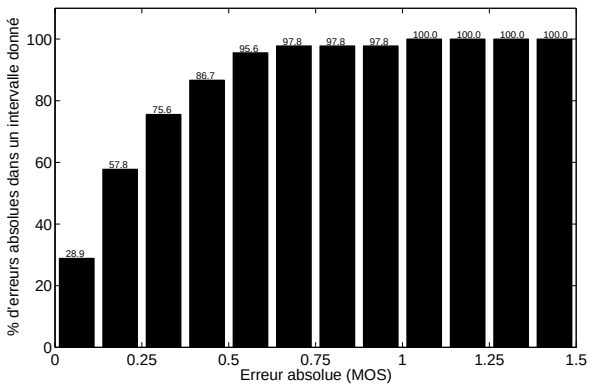
les notes de locution objectives sur un plus grand nombre de données pour chaque condition et ainsi de diminuer la grande variabilité de PESQM. Pour le test 3 sur le bruit (conditions 18-24), les notes moyennes subjectives et objectives sont effectivement mieux corrélées et plus proches qu'avec les signaux de test ($r = 0.829$ et $EAM = 0.462$ MOS au lieu de $r = 0.304$ et $EAM = 1.418$ MOS). Cependant les intervalles de confiance à 95% pour ce test, présentés dans la figure 5.7(c), sont très importants.

5.2.3 Performances du modèle objectif de qualité de conversation (CONV)

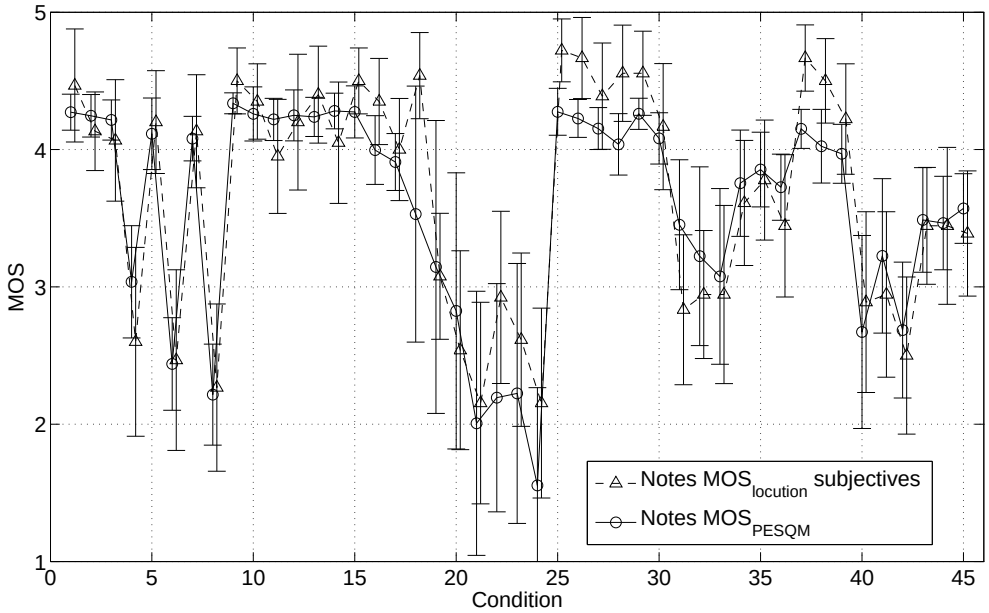
Le modèle objectif de conversation présente des performances très élevées, comme le montre la figure 5.8(a), avec une corrélation $r = 0.914$ et une erreur absolue moyenne $EAM = 0.198$ MOS entre les notes moyennes de conversation subjectives et objectives. D'après la figure 5.8(b), 100% des erreurs absolues entre notes subjectives et objectives sont inférieures à 0.625 MOS. Ces performances sur l'ensemble des tests sont confirmées test par test, par la figure 5.8(c) et le tableau 5.2. Pour chacun des quatre tests, la corrélation de Pearson r est supérieure à 0.919 et l'erreur absolue moyenne est inférieure à 0.235 MOS. Les intervalles de confiance à 95% du test 3 (conditions 18-24) sont les plus importants, ceux de PESQM dans



(a) Mapping entre notes moyennes de locution subjectives et objectives (PESQM)



(b) Distribution cumulative de l'erreur absolue entre notes moyennes de locution subjectives et objectives (PESQM)



(c) Notes moyennes de locution subjectives et objectives (PESQM) avec intervalles de confiance à 95%

Figure 5.7 : *Performances de PESQM - Signaux de conversation*

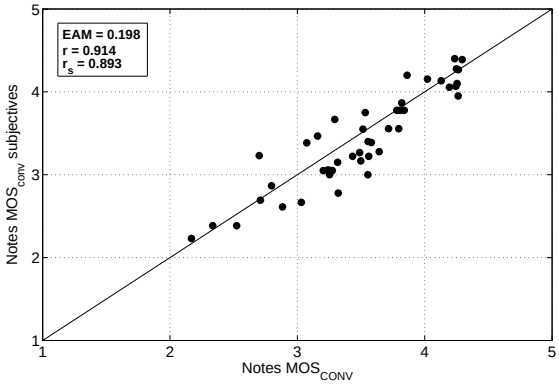
ces conditions étant également très grands.

5.2.4 Performances du modèle objectif de qualité de conversation (CONV) avec détection du bruit

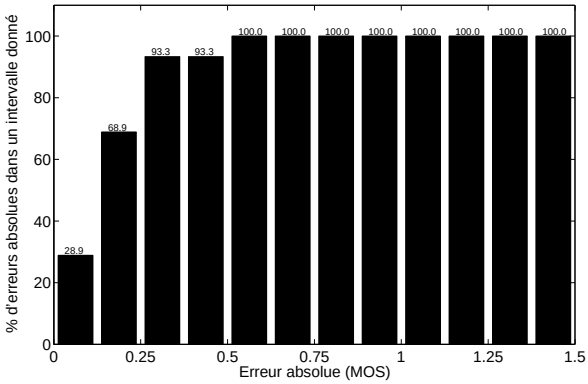
Afin d'éviter l'utilisation de PESQM dans les conditions de bruit pour lesquelles il fonctionne mal, et ainsi de diminuer les intervalles de confiance à 95% des notes de conversation objectives dans ces conditions, la détection de bruit telle qu'elle a été présentée dans le paragraphe 5.1.3 est appliquée aux signaux de conversation. Comme pour les signaux de test, la règle appliquée pour le choix de l'équation de régression est la suivante :

- si $RSBseg \geq 20$ dB, la note objective de qualité de conversation est obtenue à partir de l'équation 5.1,
- sinon, la note objective de qualité de conversation est obtenue à partir de l'équation 5.3.

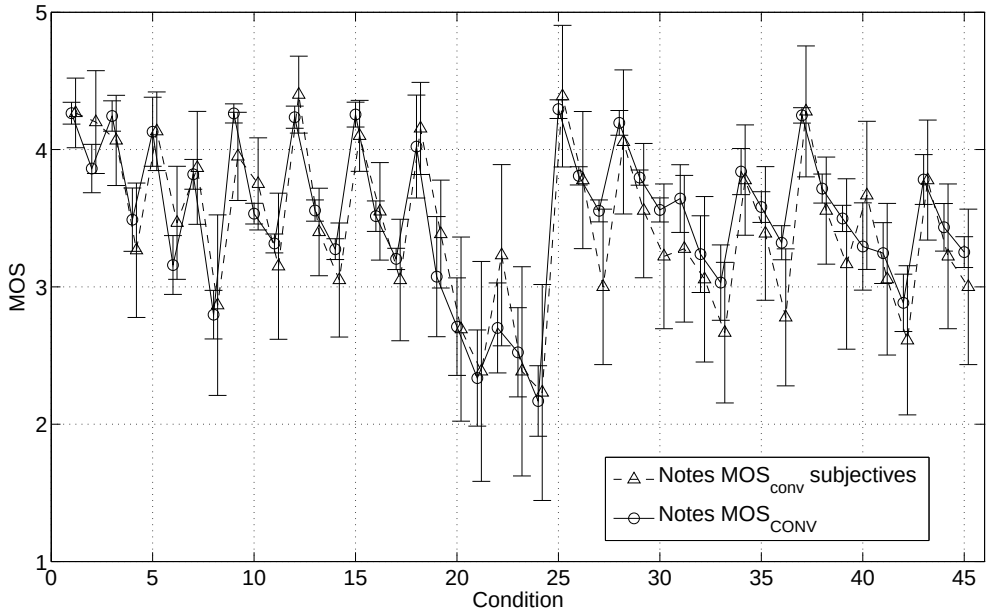
Les performances du modèle objectif de qualité de conversation avec détection du bruit sont fournies dans la figure 5.9(a). Elles sont très élevées ($r = 0.916$ et $EAM = 0.203$ MOS) et quasiment identiques à celles obtenues avec le modèle objectif sans détection du bruit. La distribution de l'erreur absolue, présentée dans la figure 5.9(b), montre que 100% des



(a) Mapping entre notes moyennes de conversation subjectives et objectives (CONV)



(b) Distribution cumulative de l'erreur absolue entre notes moyennes de conversation subjectives et objectives (CONV)



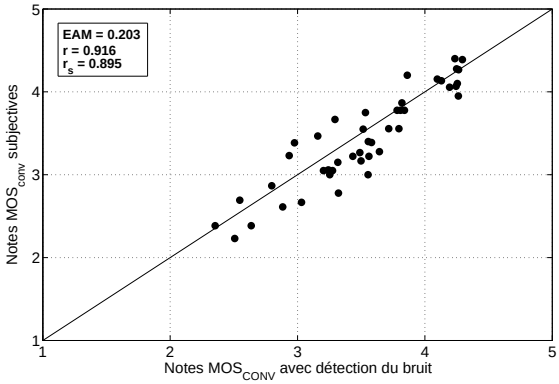
(c) Notes moyennes de conversation subjectives et objectives (CONV) avec intervalles de confiance à 95%

Figure 5.8 : *Performances du modèle de conversation - Signaux de conversation*

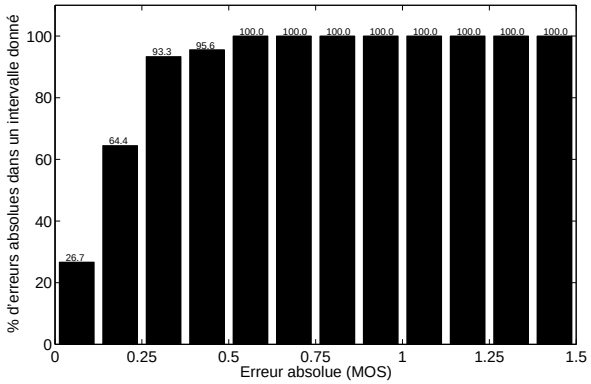
notes moyennes de conversation objectives diffèrent de moins de 0.625 MOS des notes subjectives. D'après le tableau 5.2, les performances sur les notes moyennes de conversation pour le test 3 diminuent légèrement par rapport au modèle sans détection de bruit ($r = 0.935$ et $EAM = 0.209$ MOS au lieu de $r = 0.951$ et $EAM = 0.177$ MOS). Cependant, la figure 5.9(c) montre que la détection de bruit permet de diminuer les intervalles de confiance à 95% des notes moyennes de conversation objectives pour les conditions du test 3, ce qui donne une meilleure confiance en les notes objectives de conversation que sans détection du bruit.

5.2.5 Synthèse

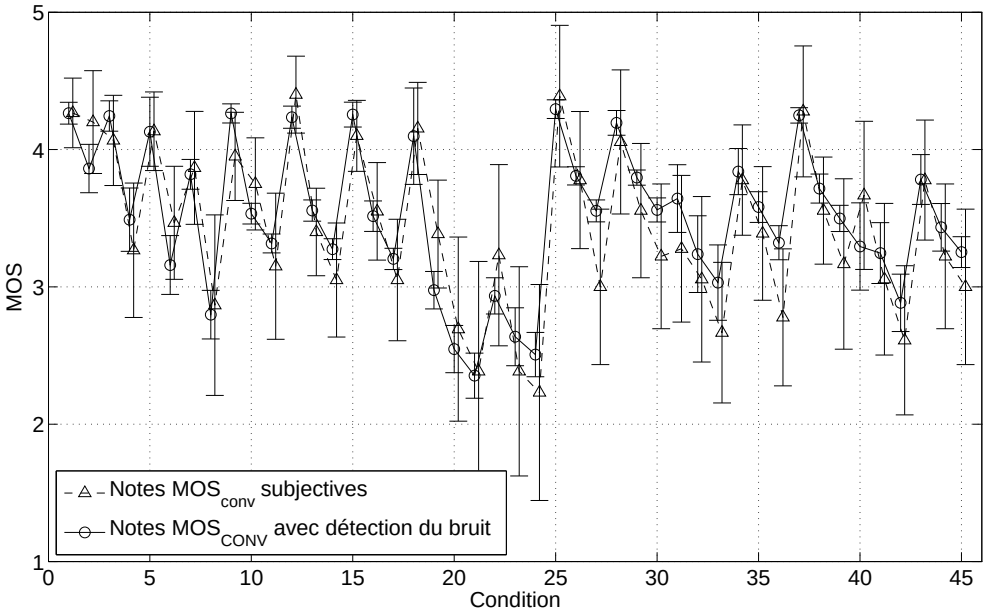
Le modèle objectif sans détection du bruit permet d'atteindre, sur les signaux de conversation, des performances élevées avec une très bonne corrélation, une erreur absolue moyenne faible et une distribution de l'erreur absolue moyenne étroite. L'ajout de la détection de bruit n'améliore pas la corrélation et l'erreur entre les notes de conversation objectives et subjectives, mais diminue sensiblement les intervalles de confiance à 95% des notes objectives moyennes de conversation. Ces résultats indiquent que le modèle objectif, moyennant l'utilisation d'un outil de découpage des signaux pour PESQM, peut fonctionner directement sur les signaux



(a) Mapping entre notes moyennes de conversation subjectives et objectives (CONV)



(b) Distribution cumulative de l'erreur absolue entre notes moyennes de conversation subjectives et objectives (CONV)



(c) Notes moyennes de conversation subjectives et objectives (CONV) avec intervalles de confiance à 95%

Figure 5.9 : *Performances du modèle de conversation avec détection du bruit - Signaux de conversation*

enregistrés durant des conversations, ce qui en étend le domaine d'application. Par rapport aux signaux de test, les signaux de conversation permettent d'améliorer l'estimation de la qualité de conversation, puisque le modèle dispose alors de plus de données, en particulier pour estimer la note de qualité de locution avec PESQM. D'après les résultats obtenus, l'outil de traitement des signaux fonctionne correctement. De plus, il permet d'éviter le problème rencontré avec les signaux de test (occurrences de parole dans les signaux reçus lors des phases de locution) qui perturbait le modèle PESQM.

5.3 Étude de l'interactivité

5.3.1 Motivations

Comme cela a été mis en évidence dans le chapitre 1, le délai a un impact sur la qualité de conversation, par l'intermédiaire de la qualité d'interaction, l'ampleur de cet impact dépendant de l'interactivité de la conversation. Ce phénomène doit donc être pris en compte pour évaluer objectivement la qualité de conversation. Dans la thèse, nous avons proposé d'estimer la qualité

d'interaction et son impact sur la qualité de conversation sous la forme d'un seuil de la valeur du délai, au-delà duquel celui-ci deviendrait gênant (*cf.* équation 5.2). Les tests subjectifs effectués dans le cadre de cette thèse se sont limités à un type de tâche de conversation, à savoir la conversation libre basée sur des scénarios de conversation proposés dans [Möller 1997a]. L'analyse des résultats du premier test subjectif a montré que, pour ce type de tâche de conversation, le délai devenait gênant au-delà de 400 ms. Pour étendre le modèle objectif proposé à d'autres types de conversation, l'idée serait de déterminer un seuil par type de conversation, *i.e.* par niveau d'interactivité. Ceci nécessite donc de disposer (i) de résultats de tests subjectifs examinant l'impact du délai pour différents niveaux d'interactivité et (ii) d'un outil mesurant le niveau d'interactivité d'une conversation à partir des signaux de parole enregistrés. Ne disposant pas de résultats de tests subjectifs explorant différents niveaux d'interactivité, nous proposons dans cette partie une étude préliminaire de l'interactivité.

Il existe peu d'études et de tests subjectifs sur l'interactivité, les deux principales sources étant les travaux de Kitawaki et Itoh [Kitawaki et Itoh 1991], qui ont permis de montrer que l'impact du délai sur le jugement dépend de la tâche de conversation, et de Hammer [Hammer 2006]. Ses travaux se basent sur l'étude des signaux de conversation, *i.e.* des temps de parole et de pause de chacun des interlocuteurs, des temps de double parole et des temps de silence mutuel, initiée par Brady. Dans [Brady 1968], il présente une analyse statistique de seize conversations (de type conversation libre, 7 minutes environ). Les salves de parole et les temps de pause sont détectées grâce à un détecteur d'activité vocale. Les salves de parole durant moins de 15 ms sont exclues et les temps de pause durant moins de 200 ms sont considérés comme des temps de parole. La conversation est vue comme un modèle à quatre états I, présenté dans la figure 5.10 (état A = parole de A et silence de B, état B = parole de B et silence de A, état M = silence mutuel, état D = double parole). De la détection d'activité vocale sont déduits la probabilité de chaque état (notée π_I, avec $I \in \{A, B, M, D\}$), le temps de séjour moyen de chaque état (noté t_I, avec $I \in \{A, B, M, D\}$) et les probabilités de transition d'un état à l'autre (notées π_{IK}, avec $I, K \in \{A, B, M, D\}$).

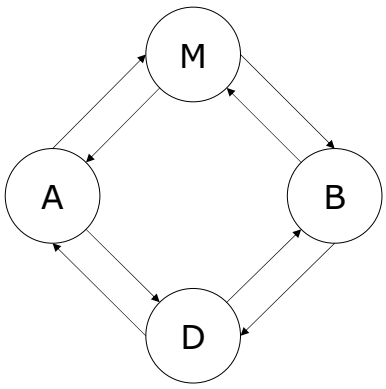


Figure 5.10 : *Modèle de la conversation à quatre états : état A = parole de A et silence de B, état B = parole de B et silence de A, état M = silence mutuel, état D = double parole*

À partir de plusieurs analyses statistiques du même type, l'UIT-T a produit une recommandation sur la voix conversationnelle artificielle, modélisant une conversation à partir des caractéristiques de ces quatre états [UIT-T Rec. P.59 1993]. Le tableau 5.3 énumère les valeurs des principaux paramètres temporels de la conversation humaine, présentés dans la Recommandation UIT-T P.59.

L'approche proposée par Hammer consiste à utiliser ces paramètres temporels pour déduire une métrique du niveau d'interactivité, nommée « température de conversation ». Elle est définie dans [Hammer 2006] en interprétant le modèle de conversation à 4 états comme une chaîne de Markov et en introduisant des notions de thermodynamique. Le temps de séjour moyen dans l'état I peut être déterminé en fonction de la température de conversation τ

État	Temps de séjour moyen $\bar{t}_I$ (s)	Probabilité d'état $\bar{\pi}_I$ (%)
A	0.78	35.2
B	0.78	35.2
M	0.51	22.5
D	0.23	6.6

Tableau 5.3 : *Paramètres temporels de la conversation (moyenne pour l'anglais, l'italien et le japonais) présentés dans [UIT-T Rec. P.59 1993]*

$$t_I(\tau) = \exp\left(\frac{1}{\kappa\tau} - \frac{1}{\kappa\tau_0}\right)\bar{t}_I \tag{5.4}$$

où τ_0 est la température « standard » (par exemple, $\tau_0 = 20\,^\circ$ en termes humains), $\bar{t}_I$ est le temps de séjour moyen dans l'état I pour la conversation standard correspondant à τ_0, et κ est l'équivalent de la constante de Boltzmann définie telle que

$$\frac{1}{\kappa} = -\ln\left(\frac{\tau_0 - 1}{\tau_0}\right)\tau_0(\tau_0 + 1). \tag{5.5}$$

L'équation 5.4 permet ainsi de déduire une estimation de la température de conversation $\hat{\tau}$, en fonction des temps de séjour moyens mesurés t_I et des temps de séjours moyens standard $\bar{t}_I$

$$\hat{\tau} = \operatorname{argmin}\left(\sum_I \alpha_I \left(\bar{t}_I \exp\left(\frac{1}{\kappa\tau} - \frac{1}{\kappa\tau_0}\right) - t_I\right)^2\right) \tag{5.6}$$

où α_I sont des facteurs de pondération, par exemple $\alpha_I = \pi_I$ pour $I \in \{A, B, M, D\}$.

5.3.2 Application

Dans [Hammer 2006], l'auteur présente l'analyse de 28 conversations effectuées par 7 paires de sujets et enregistrées lors de tests subjectifs. Ces tests consistaient en des connexions VoIP utilisant le codec G.729 avec différents taux de pertes de paquets de type rafale (0, 3, 5 et 15 %) combinés avec un délai de transmission unidirectionnel de 60, 360, 660 et 960 ms. Il était demandé aux sujets d'effectuer une conversation basée sur les scénarios de conversation interactifs (*interactive Short Conversation Test*, iSCT) [Raake 2004a]. Ceux-ci ont été conçus pour créer une situation de conversation plus interactive que celle obtenue avec les scénarios SCT, tout en restant naturelle et réaliste. Le scénario consiste en un échange rapide de données comme les adresses email et numéros de téléphones d'employés d'une entreprise. Chaque interlocuteur possède les informations complémentaires de l'autre. De plus, les deux participants d'une même paire se connaissent.

L'auteur a analysé les conversations sans perte de paquets, pour étudier l'effet du délai de transmission sur les paramètres conversationnels. Les résultats obtenus sont reproduits dans la figure 5.11, la température de conversation a été recalculée à partir des paramètres de conversation fournis dans [Hammer 2006], de l'équation 5.6 et des paramètres standard de la Recommandation P.59 fournis dans le tableau 5.3. Les probabilités d'état moyennes et les temps de séjour moyens des états A et B diminuent quand le délai augmente, alors que les valeurs de l'état M (silence mutuel) augmentent significativement entre 360 et 660 ms. D'après l'auteur, ceci montre que les sujets adaptent leur comportement conversationnel aux propriétés de la connexion. Par contre, l'état D (double parole) augmente peu avec le délai.

Les signaux de conversation enregistrés pendant le premier test de la thèse, dans les conditions sans écho et pour les délais unidirectionnels de 0, 200, 400 et 600 ms, ont également été analysés. La détection d'activité vocale est effectuée d'après la Recommandation P.56. Les salves de parole durant moins de 15 ms sont exclues et les temps de pause durant moins de

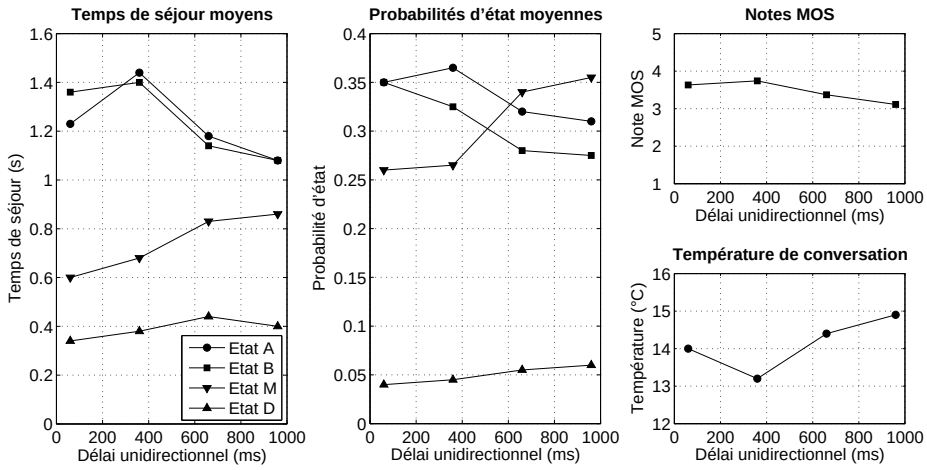


Figure 5.11 : *Paramètres conversationnels, température de conversation moyenne et notes MOS en fonction du délai (d'après [Hammer 2006])*

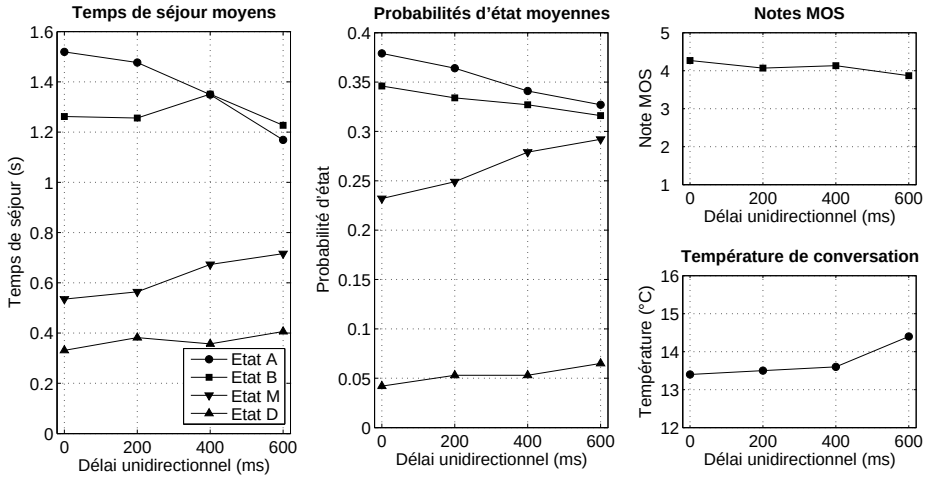


Figure 5.12 : *Paramètres conversationnels, température de conversation moyenne et notes MOS en fonction du délai (conversations enregistrées au cours du test 1)*

200 ms sont considérés comme des temps de parole. Pour chaque conversation, les paramètres temporels sont mesurés et la température de conversation est calculée selon l'équation 5.6 et avec les paramètres standard de la Recommandation P.59 fournis dans le tableau 5.3. Les moyennes des temps de séjour, des probabilités d'état et des températures de conversation sont représentées en fonction du délai unidirectionnel dans la figure 5.12.

La comparaison des deux figures 5.11 et 5.12 indique que les paramètres de conversation mesurés sont très proches pour les deux études. En particulier, les temps de séjour moyens t_A et t_B sont supérieurs à ceux reportés dans la recommandation P.59. Ceci peut s'expliquer par le fait que les deux tests présentés ici ont été effectués dans des langues non prises en compte dans la recommandation P.59 (allemand et français). Entre les deux études, seuls le temps de séjour et la probabilité de l'état M (silence mutuel) diffèrent, étant plus élevés pour les conversations iSCT que pour les conversations SCT. Les températures de conversation calculées pour les deux études diffèrent très peu, tendant ainsi à montrer que les sujets adaptent leur comportement en fonction du délai. Ce constat est confirmé par les notes MOS correspondantes, présentées dans les figures 5.11 et 5.12, qui varient peu en fonction du délai, même dans le cas de conversations plus interactives iSCT.

5.3.3 Synthèse

Cette étude sur l'interactivité, du fait du manque de données subjectives pour modéliser la qualité d'interaction, est très préliminaire. Elle souligne la difficulté d'appréhender la question de l'interactivité dans les télécommunications et sa prise en compte dans l'évaluation de la qualité de conversation. Les résultats obtenus lors de l'étude comparative de ces deux tests subjectifs montrent (i) l'évolution de la perception des utilisateurs vis-à-vis du délai, qui, bien que modifiant objectivement les paramètres temporels des conversations, semble avoir peu d'impact sur la qualité subjective de conversation dans ces deux situations d'interactivité différentes, et (ii) l'adaptation du comportement des utilisateurs au cours de la communication quand un délai est présent.

Conclusion

Dans les deux premières parties de ce chapitre, nous avons présenté les performances du modèle objectif de conversation, constitué des modèles objectifs existants d'écoute PESQ et de locution PESQM, de la valeur du délai et de l'équation de régression déterminée dans le chapitre 3. L'application aux signaux de test a tout d'abord montré que le modèle PESQM ne fonctionne pas correctement en présence de bruit seul, ce qui pénalise le modèle de conversation dans ces conditions. La solution proposée pour améliorer les performances du modèle de conversation consiste à appliquer une équation de régression ne faisant pas intervenir la note objective de locution fournie par PESQM dans les conditions dites bruitées (*i.e.* $RSBseg < 20$ dB). Le modèle objectif de conversation avec détection du bruit atteint des performances élevées, avec une corrélation $r = 0.904$ et une erreur absolue moyenne $EAM = 0.206$ MOS entre les notes moyennes de conversation subjectives et objectives. Le modèle objectif a ensuite été appliqué aux signaux de conversation, en ayant recours à l'algorithme de traitement des signaux présenté dans le chapitre 4. Sans détection du bruit, le modèle estime très efficacement la qualité de conversation. Le fait de découper les signaux de conversation permet de disposer d'un plus grand volume de données pour évaluer la qualité de locution avec PESQM et ainsi d'en améliorer l'efficacité, même dans les conditions bruitées. L'ajout de la détection du bruit aboutit à des performances identiques ($r = 0.916$ et $EAM = 0.203$ MOS) et, point très important, il s'ensuit une diminution des intervalles de confiance des notes moyennes de conversation objectives.

Avec les signaux de test, les performances du modèle objectif de conversation sont légèrement inférieures car PESQM dispose alors de moins de données qu'avec les signaux de conversation pour estimer la qualité de locution. Le modèle objectif de conversation proposé nécessite donc plusieurs mesures objectives (*i.e.* plusieurs mesures PESQ et plusieurs mesures PESQM) pour une condition donnée afin d'atteindre les performances présentées dans ce chapitre. Sur la base d'une mesure objective unique (*i.e.* une mesure PESQ et une mesure PESQM) pour une condition donnée, les performances du modèle objectif de conversation seront moins élevées. Cette contrainte reste malgré tout compatible avec les objectifs fixés dans le chapitre 2, à savoir que la qualité de conversation doit être évaluée appel par appel. En effet, l'outil de traitement des signaux, découpant les signaux enregistrés lors d'un appel, permet au modèle objectif d'atteindre des performances élevées. Il reste à déterminer, au niveau pratique, combien de mesures sont nécessaires pour une condition donnée pour fournir une évaluation correcte de la qualité de conversation.

Les résultats présentés dans les deux premières parties de ce chapitre confirment la validité de l'approche choisie sur des signaux enregistrés dans des conditions réelles et la faisabilité du modèle objectif de conversation avec les outils objectifs existants. L'application du modèle objectif à d'autres bases de signaux enregistrés lors de tests subjectifs extérieurs à la thèse est à mener. Les difficultés rencontrées avec le modèle objectif de qualité de locution PESQM, bien que résolues pour les conditions avec bruit seul, seront à prendre en compte par la suite

et de nouveaux tests subjectifs de locution seront nécessaires pour évaluer ses performances dans d'autres conditions, en particulier dans les conditions avec bruit et écho.

Le modèle objectif proposé a été construit, dans le chapitre 3, à partir de tests subjectifs qui se limitaient à une tâche de conversation : la conversation libre. L'analyse des résultats du premier test subjectif a montré que, pour ce type de tâche de conversation, le délai devenait gênant au-delà de 400 ms. Cependant, l'impact du délai sur la qualité de conversation varie en fonction de l'interactivité de la communication. Afin d'étendre le modèle objectif proposé à d'autres types de conversation, il faut donc adapter ce seuil au niveau d'interactivité de la communication.

La dernière partie de ce chapitre a présenté l'étude d'un outil, proposé dans la littérature, mesurant le niveau d'interactivité d'une conversation à partir des signaux de parole enregistrés. Cet outil, appelé température de conversation, a été appliqué aux signaux enregistrés pendant le premier test subjectif (sur l'écho et le délai). Les résultats obtenus ont été comparés à ceux rapportés dans un article de la littérature, qui se basait sur les signaux enregistrés lors d'un test subjectif de conversation sur le délai. Cette comparaison a permis de montrer tout d'abord le manque de données subjectives sur la qualité d'interaction, qui rend difficile la question de l'interactivité. Ensuite, les deux tests étudiés, bien qu'effectués avec des tâches de conversation à des niveaux d'interactivité différents, aboutissent à la même conclusion, à savoir que l'impact du délai sur le jugement des utilisateurs a évolué et qu'au cours d'une communication affectée par un délai, les interlocuteurs adaptent leur comportement, ce qui complexifie l'évaluation de l'interactivité. Des tests subjectifs étudiant spécifiquement la qualité d'interaction et son influence sur la qualité de conversation, à différents niveaux d'interactivité et de délai, sont donc nécessaires. Mais ils ne seront possibles que quand une méthodologie de test adéquate sera définie.

Conclusions et perspectives

L'évaluation de la qualité vocale des systèmes de télécommunications est nécessaire pour des raisons techniques et commerciales depuis l'expansion des réseaux numériques, mobiles ou VoIP. La qualité vocale peut être évaluée avec deux types de méthodes : subjectives et objectives. Les méthodes subjectives reflètent directement le jugement humain de la qualité vocale, mais elles sont contraignantes à mettre en œuvre et coûteuses. Des méthodes objectives ont donc été développées par les opérateurs de télécommunications comme alternative aux méthodes subjectives.

L'état de l'art des méthodes objectives d'évaluation de la qualité vocale a mis en évidence (i) la multitude de modèles dans le contexte d'écoute, de différents types (paramétriques, basés sur les signaux, avec ou sans référence) et dont plusieurs sont normalisés à l'UIT-T, et (ii) l'existence du modèle PESQM (basé sur les signaux avec référence, non normalisé) en contexte de locution. Dans la vie quotidienne, un usager des systèmes de télécommunications est rarement placé uniquement en contexte d'écoute ou uniquement en contexte de locution. Dans la situation la plus courante pour les utilisateurs, à savoir le contexte de conversation, l'état de l'art a montré que les seuls modèles existants étaient paramétriques (modèles E et CCI) et n'étaient pas suffisamment efficaces pour estimer la qualité telle qu'elle est perçue par l'utilisateur. Il manquait donc un modèle d'évaluation de la qualité de conversation, basé sur l'analyse des signaux (avec ou sans référence).

La Question 20 de la Commission d'Études 12 de l'UIT-T a pour but de normaliser un modèle objectif de la qualité de conversation, qui prédit l'impact des dégradations du réseau sur la qualité conversationnelle ressentie par l'utilisateur final. Les objectifs du travail présenté dans ce mémoire étaient conformes à ceux formulés par la Question 20, dans l'idée finale d'y contribuer. Le modèle objectif de qualité de conversation proposé devait :

1. Être non paramétrique, *i.e.* basé sur l'analyse des signaux échangés pendant la communication testée.
2. Fonctionner pour une connexion électrique/électrique et bande étroite.
3. Évaluer la qualité de conversation appel par appel.
4. Utiliser des mesures intrusives ou des mesures non intrusives, selon l'application visée.

Pour construire un modèle répondant à ces différents objectifs, il était indispensable de disposer de données subjectives. Un autre objectif de cette thèse a donc consisté en la conception et la mise en œuvre de plusieurs tests subjectifs pour étudier l'impact des dégradations rencontrées dans le contexte de conversation sur la qualité vocale perçue.

L'état de l'art sur l'étude de la structure d'une conversation a montré que la conversation pouvait être décrite, du point de vue d'un interlocuteur, comme une alternance des rôles d'auditeur et de locuteur introduisant de l'interaction entre les interlocuteurs. Du point de vue de la qualité vocale, cela signifie que le contexte de conversation est affecté par les dégradations rencontrées dans le contexte d'écoute et celles rencontrées dans le contexte de locution, auxquelles s'ajoutent les dégradations affectant l'interaction de la conversation (délai et dégradation due aux périodes de double parole). La méthode proposée dans ce travail

s'est appuyée sur ce constat et sur l'hypothèse que la qualité de conversation pouvait être décomposée selon trois dimensions : la qualité d'écoute, la qualité de locution et la qualité d'interaction. L'approche proposée est divisée en deux parties :

- la partie intégration combine les notes de qualité d'écoute, de locution et d'interaction pour estimer une note de qualité de conversation,
- la partie mesure fournit les notes objectives de qualité à la partie intégration en se basant sur les modèles existants de qualité vocale dans les différents contextes.

Le modèle fonctionne pour plusieurs applications selon les modèles utilisés dans la partie mesure, la partie intégration restant commune à toutes les applications.

La qualité d'interaction étant principalement dégradée par le délai, il a été choisi de considérer la valeur du délai comme un indicateur de la qualité vocale d'interaction. La note de qualité de conversation est ainsi estimée par une combinaison des notes de qualité d'écoute et de locution et de la valeur du délai présent dans la communication testée.

Pour vérifier cette hypothèse, une nouvelle méthodologie de test subjective a été conçue pour évaluer au cours d'un même test les qualités de conversation, d'écoute et de locution, dans les mêmes conditions de dégradation. Quatre tests subjectifs ont ainsi été mis en œuvre et ont permis : (i) de valider cette hypothèse et (ii) de déterminer la partie intégration du modèle objectif de conversation, pour plusieurs dégradations (écho et délai, pertes de paquets, bruit).

Tout d'abord, pour chacune des dégradations testées, une combinaison linéaire des notes subjectives d'écoute et de locution recueillies lors de ces tests et de la valeur du délai est calculée afin d'estimer la note subjective de conversation. L'ensemble des relations ainsi obtenues aboutit à d'excellentes estimations de la qualité de conversation. Une détection des dégradations est cependant nécessaire afin de choisir la relation adéquate pour chaque communication. Nous avons donc cherché une unique relation linéaire pour l'ensemble des dégradations : elle aboutit à une excellente estimation de la qualité de conversation (i) sur la base d'apprentissage, composée de trois des quatre tests propres à la thèse, et (ii) sur une base de validation, composée du quatrième test de la thèse et d'un test extérieur à cette étude. La partie intégration du modèle fonctionne donc sur l'ensemble des conditions testées avec une équation de régression unique, permettant de s'affranchir de la détection des dégradations.

La partie mesure du modèle nécessitait un modèle objectif de la qualité d'écoute, un modèle objectif de la qualité de locution, un outil de mesure du délai, un outil de traitement des signaux pour adapter les signaux de conversation aux modèles objectifs de qualités d'écoute et de locution. Nous avons décidé de nous intéresser en particulier à la mesure intrusive avec référence, en utilisant les modèles existants de qualité d'écoute PESQ et de qualité de locution PESQM. Non normalisé, ce dernier a été implémenté et optimisé sur les notes subjectives recueillies lors d'un test de locution mis en œuvre à cet effet et portant sur différentes dégradations affectant la qualité de locution (écho, bruit, pertes de paquets). Un outil de traitement a été conçu : il permet de calibrer les signaux enregistrés dans des réseaux réels ou lors de conversations aux contraintes des modèles intrusifs PESQ et PESQM sur la durée, le taux d'activité vocale, etc. Cet outil, basé sur une détection d'activité vocale, sélectionne : (i) pour le modèle PESQ, des portions de signaux ayant les durées et les taux d'activité vocale appropriés et (ii) pour le modèle PESQM, des portions de signaux permettant d'obtenir un signal dégradé ne contenant que l'écho éventuel sans la parole du locuteur distant et ayant des durées et des taux d'activité vocale appropriés.

Les performances du modèle objectif, constitué des modèles objectifs existants PESQ et PESQM et de la valeur (supposée connue) du délai, ont été évaluées sur deux types de signaux : les signaux de test (directement calibrés pour les modèles intrusifs) et les signaux de conversation (préalablement adaptés aux contraintes des modèles intrusifs grâce à l'outil de traitement développé dans le cadre de la thèse). L'application aux signaux de test a montré

qu'en présence de bruit seul le modèle PESQM ne fonctionne pas correctement. Afin de remédier à ce problème (à court terme), nous avons proposé d'appliquer, dans les conditions dites bruitées (*i.e.* $RSBseg < 20$ dB), une équation de régression dédiée et n'utilisant pas la note objective de locution fournie par PESQM. Le modèle objectif de conversation avec détection du bruit atteint des performances élevées, avec une corrélation $r = 0.904$ et une erreur absolue moyenne $EAM = 0.206$ MOS entre les notes moyennes de conversation subjectives et objectives. Appliqué aux signaux de conversation et sans détection du bruit, le modèle estime très efficacement la qualité de conversation. En effet, le découpage des signaux de conversation permet de disposer d'un plus grand nombre de réalisations pour évaluer la qualité de locution avec PESQM et ainsi d'en améliorer l'efficacité, même dans les conditions bruitées. Avec détection du bruit, le modèle atteint des performances identiques ($r = 0.916$ et $EAM = 0.203$ MOS) tout en diminuant les intervalles de confiance des notes moyennes de conversation objectives.

Le travail présenté dans ce mémoire a fait l'objet de plusieurs publications ainsi que d'une demande de brevet, référencées dans l'annexe F, et a apporté plusieurs contributions au domaine de l'évaluation, tant objective que subjective, de la qualité vocale :

- la construction de la partie intégration a permis une étude « fondamentale » de l'impact de diverses dégradations sur la qualité de conversation et de la relation entre ses différentes dimensions. Cette partie intégration, invariable selon les applications, est entièrement indépendante des modèles objectifs de la partie mesure, permettant ainsi au modèle objectif de conversation d'évoluer sans difficulté avec les modèles de qualité d'écoute et de locution. Ainsi définie, la partie intégration permet d'atteindre les objectifs d'un modèle non paramétrique, fonctionnant pour une connexion électrique/électrique et bande étroite, et utilisant des mesures intrusives ou non intrusives, en fonction des modèles choisis dans la partie mesure,
- la mise en œuvre de la partie mesure a conduit au développement d'un outil de traitement des signaux élargissant le fonctionnement du modèle objectif aux signaux enregistrés lors de conversations réelles. Cet outil permet d'atteindre l'objectif d'une évaluation de la qualité de conversation appel par appel : en découpant les signaux enregistrés lors d'un appel, le modèle dispose de plusieurs portions de signaux (*i.e.* plusieurs réalisations pour une condition donnée) pour évaluer les qualités d'écoute et de locution et ainsi estimer plus efficacement la qualité de conversation,
- afin de mener à bien cette étude, une nouvelle méthodologie subjective de test et de nouveaux tests subjectifs ont été conçus, élargissant ainsi les données et les connaissances sur la qualité de conversation, mais aussi sur la qualité de locution.

Comme l'a montré l'état de l'art, les données subjectives sont véritablement le « nerf de la guerre » dans le domaine de l'évaluation de la qualité vocale. Le manque de données subjectives a limité ce travail sur plusieurs aspects :

- le modèle objectif estime la qualité de conversation uniquement pour certaines dégradations, seules ou combinées, explorées lors des quatre tests mis en œuvre pendant la thèse et d'un test conduit indépendamment,
- les tests subjectifs utilisés pour la construction du modèle objectif de conversation se cantonnent à des conversations libres et donc à un seul niveau d'interactivité.

Ce travail est également limité, dans sa partie mesure, par les modèles objectifs existants, en particulier en contexte de locution. Le modèle PESQM, bien qu'excellent dans les conditions avec écho et pertes de paquets, fournit de mauvaises estimations dans les conditions bruitées. D'autres tests subjectifs de locution seraient nécessaires pour l'améliorer dans les conditions avec bruit et vérifier ses performances dans les conditions avec écho et bruit. De plus, PESQM est le seul modèle de locution existant et il fonctionne de manière intrusive : notre modèle de qualité de conversation est intrusif.

Les contributions et les limitations de ce travail montrent l'intérêt prometteur d'un tel modèle objectif et tracent des perspectives :

- d'autres dégradations et combinaisons de dégradations, telles que le bruit en présence d'écho, le type de codec, la gigue, etc., pourront être étudiées avec la méthodologie de test proposée dans la thèse afin d'étendre le modèle,
- de plus, il serait intéressant d'étudier la relation entre la qualité de conversation et les qualités d'écoute, de locution et d'interaction dans le cas de systèmes bande élargie ou de connexions acoustiques,
- une méthodologie subjective et des tests étudiant spécifiquement la qualité d'interaction et son influence sur la qualité de conversation, à différents niveaux d'interactivité et de délai, seraient nécessaires,
- le même type d'approche que celle proposée dans cette thèse pourrait être transposé aux applications multimédia afin d'évaluer objectivement la qualité de conversation audiovisuelle, à condition de disposer des tests subjectifs correspondants.

Enfin, le modèle objectif de conversation présenté dans ce mémoire au niveau logiciel doit pouvoir être mis en œuvre au niveau pratique et matériel pour être utilisé par les opérateurs de télécommunications dans les réseaux réels. Cette mise en œuvre pratique a été initiée au cours d'un stage de Master mené dans l'équipe TECH/SSTP/MOV de France Télécom R&D. Le modèle a été implémenté dans un instrument de mesure, appelé DSLA (*Digital Speech Level Analyser*) et intégrant déjà le modèle PESQ. Branché au système à tester via une interface électrique ou acoustique, il prend en charge l'envoi des signaux de test dans le système, l'enregistrement des signaux dégradés, la mesure du délai présent dans le système, l'application du modèle PESQ aux signaux de référence et dégradé en contexte d'écoute, l'application du modèle PESQM aux signaux de référence et dégradé en contexte de locution, et enfin le calcul de la note objective de qualité de conversation à partir des coefficients de régression déterminés dans le chapitre 3. Cette mise en œuvre pratique ouvre la voie d'une utilisation et d'une commercialisation du modèle objectif de conversation développé au cours de cette thèse.

Annexe A

Notions de psychoacoustique

La psychoacoustique est décrite par McAdams [McAdams 1994] comme l'étude des relations quantifiables entre paramètres sonores (fréquence, amplitude, durée, etc.) et paramètres perceptifs (hauteur tonale, sonie, timbre, durée subjective).

A.1 Physique du phénomène sonore

La puissance acoustique P d'une vibration correspond à la quantité d'énergie déployée par unité de temps. L'intensité acoustique I correspond à la puissance transmise par unité de surface. Elle est proportionnelle au carré de la pression. L'unité de pression acoustique utilisée en psychoacoustique est le Pascal, noté Pa. L'oreille est sensible à la gamme de pression située entre 20 μPa et 100 Pa. La pression notée p_0 valant 20 μPa, correspondant à une intensité acoustique de 10^{-12} W/m^2, est prise comme pression standard. Le niveau de pression acoustique (exprimé en dB SPL) est tel que

$$\text{Niveau} = 20\log(p/p_0) = 10\log(I/I_0). \tag{A.1}$$

A.2 Capacités sensorielles et dimensions de la perception auditive

L'aire d'audition s'étend environ de 20 Hz à 20 kHz et de 0 dB à 130 dB, comme l'illustre la figure A.1.

A.2.1 Bandes critiques

Différentes expériences et essais acoustiques ont montré que l'ouïe humaine regroupe des excitations sonores ayant des fréquences voisines dans certaines bandes fréquentielles appelées « bandes critiques ». À l'intérieur d'une bande critique un son peut être masqué, alors qu'il ne peut pas l'être à l'extérieur de cette même bande. Ainsi, ces expériences prouvent que l'oreille humaine est dotée de récepteurs sélectifs en fréquence, traitant des zones fréquentielles dont la largeur est précisément la largeur de la bande critique. Donc deux sons séparés de plus d'une bande critique excitent des récepteurs disjoints, ils sont ainsi complètement discriminés. La largeur des bandes critiques :

- est constante ($\simeq$ 100 Hz) jusqu'à 500 Hz,
- augmente au-delà en restant de l'ordre de 10% à 15% de la fréquence centrale.

A.2.2 Masquage

Le phénomène de masquage peut être :

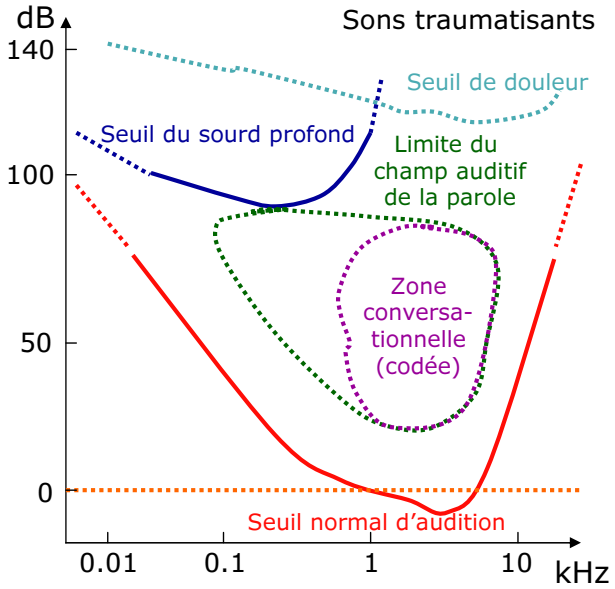


Figure A.1 : *Champ audible pour un jeune adulte entre les seuils d'audibilité et de douleur*

- total : lorsqu'un son à un niveau sonore donné (le masqué) n'est plus détectable en présence d'un autre son (le masquant).
- partiel : réduction de l'intensité subjective d'un son à un niveau donné lorsqu'un autre son est présent simultanément.

Si un son pur est présenté dans un fond de bruit blanc masquant, seules les fréquences du bruit proches de celle du son pur contribuent à l'effet de masque. Le seuil masqué correspond à l'égalité de l'énergie du signal et de celle du bruit dans la bande critique centrée sur le signal. L'oreille détermine ainsi au sein d'une bande de fréquences les seuils d'audition des sons. Elle compare dans cette bande les intensités du bruit masquant et du son test. Le son test est perçu dès que son niveau, dans la bande critique autour de 1 kHz, se situe ~ 4 dB en-dessous du niveau du son masquant dans cette même bande de fréquences. La décomposition du spectre de fréquence en bandes critiques correspond à une propriété fondamentale de l'ouïe. Dans la zone de fréquences de 20 Hz à 20 kHz, il y a 24 bandes critiques.

A.2.3 Audiogramme masqué

L'audiogramme masqué est le niveau que doit atteindre le son masqué, en fonction de sa fréquence, pour être juste audible en présence d'un son masquant de fréquence et d'intensité fixes. Il possède une pente raide du côté des fréquences basses et une pente moins raide du côté des fréquences élevées. L'effet de masque s'étend de plus en plus vers les fréquences élevées lorsque l'intensité du masquant s'accroît. Un son fort de fréquence basse affecte le comportement temporel des fibres accordées aux fréquences plus élevées, l'inverse ne se produit pas.

A.3 Perception de l'intensité acoustique

L'intensité acoustique d'une source sonore est appelée *sonie*. Pour mesurer la sonie d'un son pur, sa fréquence (1 kHz) et sa durée (1 s) sont maintenues constantes. À un son de niveau acoustique $L = 40$ dB de fréquence 1 kHz et de durée 1 s, une sonie $N = 1$ sone est attribuée arbitrairement. Le niveau de sonie peut également s'exprimer en phones et son niveau en phones correspond alors au niveau physique du son de référence en dB SPL.

Dans la marge des intensités comprises entre 30 et 120 dB, la fonction de sonie peut être décrite par la loi de puissance de Stevens avec un exposant 0.6 : $S = kp^{0.6}$ avec S = sonie en sones, p = pression sonore en μPa et k = constante. Pour doubler la sonie, il faut augmenter l'intensité de 10 dB, c'est-à-dire que 2 sones sont égales à 50 dB à 1 kHz.

Plusieurs études ont démontré que la sonie varie en fonction de la composition spectrale. Si la sonie d'un son pur à 1 kHz est appliquée à celle d'une bande de bruit centrée sur 1 kHz, tout en gardant constante la puissance acoustique du bruit, la sonie est indépendante de la largeur de bande jusqu'à 160 Hz, ce qui correspond à la bande critique à 1 kHz. Au-delà de 160 Hz, la sonie augmente avec la largeur si le niveau global est supérieur à 20 dB SPL. La sonie dépend de la somme de l'énergie dans différentes bandes critiques.

A.4 Perception de la hauteur

La sensation de hauteur de son est appelée *tonie*. Pour un son pur, son niveau acoustique (80 dB) et sa durée (1 s) sont gardés constants : à un son pur de 131 Hz une tonie de 131 mels est attribuée. L'échelle des mels est ensuite obtenue en demandant à un auditeur d'ajuster la fréquence d'un générateur de son pur jusqu'à ce que la hauteur tonale soit le double ou la moitié d'une hauteur de référence. La tonie peut également s'exprimer en Bark. Si une fréquence f est augmentée d'une quantité égale à la bande critique, la tonie correspondante augmente de 1 Bark. Ainsi, 1 Bark = 100 mels traduisant la relation très étroite qui existe entre la tonie et l'échelle que forment les bandes critiques juxtaposées.

A.5 Échelles naturelles de la membrane basilaire

La figure A.2 montre en haut les 32 mm de la membrane basilaire, en-dessous, une échelle linéaire représente la répartition régulière des cellules ciliées le long de la membrane basilaire. La quatrième échelle donne le rapport entre la fréquence d'un son et le lieu de l'excitation maximale. Elle permet de déduire les échelles suivantes.

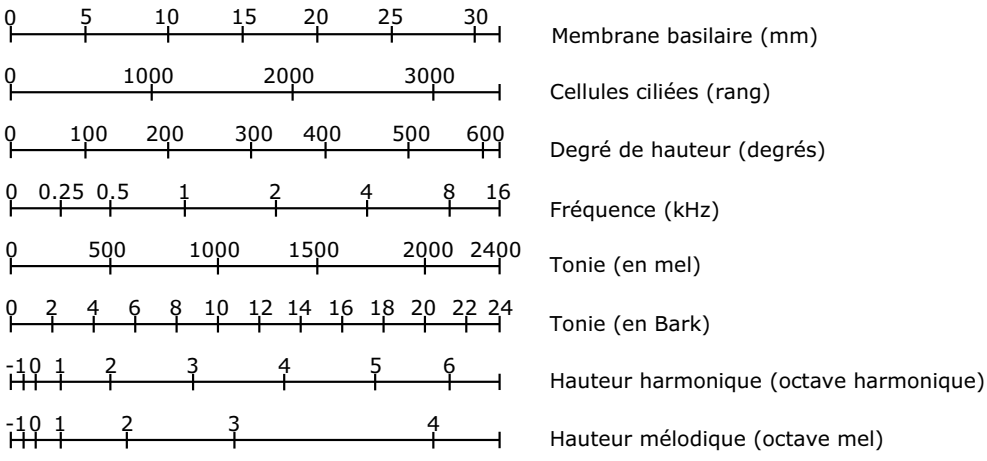


Figure A.2 : *Les échelles naturelles de la membrane basilaire (d'après [Zwicker et Feldtkeller 1981])*

Annexe B

Modèle objectif PESQ

Ce modèle objectif d'évaluation de la qualité de parole (en contexte d'écoute) résulte de la mise en commun de deux modèles : PSQM [UIT-T Rec. P.861 1998] et PAMS [Rix et Hollier 2000] et a été normalisé par l'UIT-T dans la recommandation P.862 [UIT-T Rec. P.862 2001]. Il regroupe les processus d'alignement temporel de PAMS et la modélisation perceptuelle de PSQM.

B.1 Domaine d'application et limitations de PESQ

Les tableaux B.1 à B.3 donnent les domaines d'application du modèle PESQ, ainsi que ses limitations [UIT-T Rec. P.862 2001].

Facteurs expérimentaux
Niveaux d'entrée des signaux de parole dans un codec
Erreur de transmission dans la voie
Perte de paquets et masquage de perte de paquets avec les codecs CELP
Débits lorsque le codec peut fonctionner selon plusieurs modes
Transcodages
Bruit ambiant du côté émission
Incidence de la variation du temps de propagation dans les essais d'écoute seulement
Prédistorsion à court terme du signal audio
Prédistorsion à long terme du signal audio
Techniques de codage
Codecs temporels (G.711, G.726 et G.727, par exemple)
Codecs CELP et hybrides $\geq$ 4 kbit/s (G.728, G.729 et G.723.1, par exemple)
Autres codecs : GSM-FR, GSM-HR, GSM-EFR, GSM-AMR, CDMA-EVRC, TDMA-ACELP, TDMA-VSELP, TETRA
Applications
Évaluation du codec
Sélection du codec
Essais actifs dans le réseau par connexion numérique ou analogique au réseau
Essais de réseaux émulés et prototypes

Tableau B.1 : *Facteurs pour lesquels la méthode PESQ s'est révélée d'une précision acceptable*

Facteurs expérimentaux
Niveaux d'écoute
Affaiblissement en sonie
Incidence du temps de propagation dans les essais de conversation
Écho pour le locuteur
Effet local
Techniques de codage
Remplacement de fractions continues du signal vocal constituant plus de 25% de la conversation active par un silence (écrêtage temporel extrême)
Applications
Dispositifs de mesure sans intrusion en service
Qualité des communications bidirectionnelles

Tableau B.2 : *La méthode PESQ est réputée fournir des prédictions inexactes lorsqu'elle est utilisée conjointement avec ces variables ou n'est pas censée être utilisée avec ces variables*

Facteurs expérimentaux
Perte de paquets et masquage de perte de paquets avec des codecs de type MIC
Écrêtage temporel du signal vocal
Écrêtage en amplitude du signal vocal
Variations selon le locuteur
Locuteurs multiples simultanés
Discordance de débit entre un codeur et un décodeur si un codec possède plusieurs modes de débit
Signaux d'information de couche réseau à l'entrée d'un codec
Signaux vocaux artificiels à l'entrée d'un codec
Signaux de musique à l'entrée d'un codec
Écho pour la personne qui écoute
Effets/artefacts causés par le fonctionnement des annuleurs d'écho
Effets/artefacts causés par les algorithmes de réduction de bruit
Techniques de codage
Codecs CELP et hybrides < 4 kbit/s
HVXC MPEG4
Applications
Essais acoustiques des terminaux/combinés au moyen du simulateur HATS, par exemple

Tableau B.3 : *Facteurs, techniques et applications pour lesquels la méthode d'évaluation PESQ n'a pas été encore validée à ce jour*

B.2 Principe

Le modèle PESQ intègre principalement deux modules : le module d'échelonnement et d'alignement temporel et le module psychoacoustique.

B.2.1 Échelonnement et alignement temporel

Afin d'être robuste vis-à-vis du délai pouvant exister entre le signal initial et le signal dégradé, le modèle PESQ intègre un module d'alignement temporel, présenté en détail dans [Rix *et al.* 2002]. Les signaux initial et dégradé sont alignés temporellement d'après les étapes fournies dans la figure B.1.

B.2.1.1 Échelonnement du niveau

Les niveaux des deux signaux initial $X(t)$ et dégradé $Y(t)$ sont échelonnés sur le même niveau de puissance constant. En effet, chaque système testé peut avoir un gain différent. De

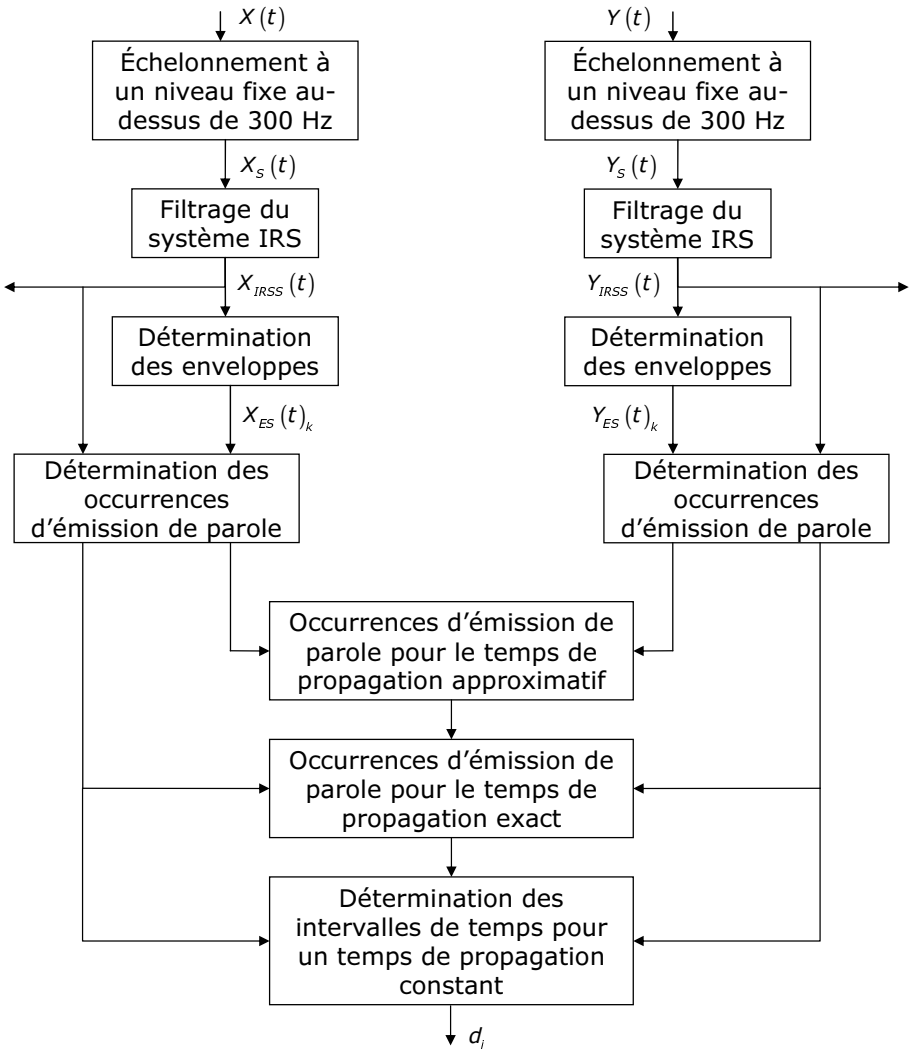


Figure B.1 : *Principe de fonctionnement du module d'alignement temporel utilisé dans le modèle PESQ (*Perceptual Evaluation of Speech Quality*) [UIT-T Rec. P.862 2001] pour déterminer le temps de propagation par intervalle de temps d_i*

plus, le niveau auquel le signal dégradé est écouté par les utilisateurs varie selon les systèmes, les téléphones, etc. Le modèle ne doit donc pas prendre en compte la différence de niveau entre le signal initial et le signal dégradé. Leurs niveaux sont donc échelonnés à une valeur constante de 79 dB SPL au point de référence oreille. De cette opération découlent les signaux échelonnés $X_S(t)$ et $Y_S(t)$.

B.2.1.2 Filtrage du système IRS

Le filtrage IRS (*Intermediate Reference System*) permet de simuler les caractéristiques fréquentielles d'un signal émis par un combiné téléphonique. Le modèle PESQ applique le filtre IRS aux FFT des signaux $X_S(t)$ et $Y_S(t)$. Les signaux filtrés $X_{IRSS}(t)$ et $Y_{IRSS}(t)$ sont ensuite obtenus grâce à une FFT inverse.

B.2.1.3 Alignement temporel

Le module d'alignement temporel comprend plusieurs étapes [Rix *et al.* 2002] :

– le temps de propagation approximatif est estimé grâce à l'intercorrélation des enveloppes $X_{ES}(t)_k$ et $Y_{ES}(t)_k$ des signaux $X_S(t)$ et $Y_S(t)$. L'enveloppe est définie par la formule

$\log(\max(E(k)/Ethresh, 1))$, où $E(k)$ est l'énergie dans une trame k de 4 ms et $Ethresh$ est le seuil de conversation déterminé par un détecteur d'activité vocale.
- le signal initial est divisé en occurrences d'émission de parole (ou paroles émises).
- le temps de propagation de groupe d'après les occurrences d'émission de parole est estimé.
- le temps de propagation fin est identifié d'après les occurrences de parole, par corrélation fine ou par histogramme.
- les occurrences de parole sont coupées et les intervalles de temps sont réalignés pour repérer les variations du temps de propagation en cours de conversation.
- les longs passages comportant des erreurs importantes, identifiés par le module perceptuel de PESQ, sont réalignés.

Le calcul du temps de propagation fin permet de déterminer une valeur de temps de propagation exact par échantillon, selon les étapes suivantes :
- les signaux initial et dégradé sont divisés en trames de 64 ms (se chevauchant à 75%) par fenêtrage de Hanning.
- l'intercorrélation entre chaque trame du signal initial et chaque trame du signal dégradé est calculée, après alignement selon les enveloppes.
- le maximum de la corrélation, à la puissance 0.125, est utilisé pour mesurer le niveau de confiance de l'alignement dans chaque trame. L'indice du maximum indique la valeur estimative du temps de propagation pour chaque trame.
- un histogramme de ces valeurs estimatives du temps de propagation, pondérées par le niveau de confiance mesuré, est calculé. L'histogramme est ensuite lissé par convolution avec un noyau triangulaire symétrique d'une largeur de 1 ms.
- l'indice du maximum dans l'histogramme, conjugué à la valeur estimative du temps de propagation précédent, indique la valeur estimative du temps de propagation final.
- le maximum de l'histogramme, divisé par la somme de l'histogramme avant convolution avec le noyau, indique un niveau de confiance compris entre 0 (aucune confiance) et 1 (confiance maximale).

B.2.2 Modèle psychoacoustique

Une fois alignés, les deux signaux sont envoyés dans le modèle psychoacoustique, présenté dans les figures B.2 et B.3.

B.2.2.1 Initialisations et calibrations

Les échantillons de début et de fin du signal de référence sont déterminés. Le premier échantillon déclaré actif est celui dont l'amplitude (*i.e.* valeur absolue) plus les amplitudes des quatre échantillons précédents vaut 500 ou plus. Le dernier échantillon déclaré actif est le dernier échantillon dont l'amplitude plus les amplitudes des quatre échantillons suivants vaut 500 ou plus.

Un étalonnage entre le niveau d'écoute et la sonie comprimée est également nécessaire. Pour cela, une sinusoïde de référence avec une fréquence de 1 kHz et une amplitude de 40 dB SPL (c'est-à-dire -64 dBov, et une amplitude zéro-à-crête de 29.54) est générée[1]. Le premier étalonnage consiste à échelonner à 10^4 la valeur maximale de la représentation de puissance fondamentale contenue dans la sinusoïde d'étalonnage. Cette valeur est obtenue par une post-multiplication avec une constante, le facteur d'échelonnement de puissance S_p.

Le deuxième étalonnage fixe à la valeur de 1 Sone la sonie comprimée de la sinusoïde de référence. Cette valeur est obtenue par une post-multiplication avec une constante, le facteur d'échelonnement de la sonie S_l.

[1] Un signal de parole codé sur 16 bits varie entre $\pm 2^{16}/2 - 1 = \pm 32767$. Pour une sinusoïde, une amplitude zéro-à-crête de 32767 correspond à -3.01 dBov ou 100.9 dB SPL.

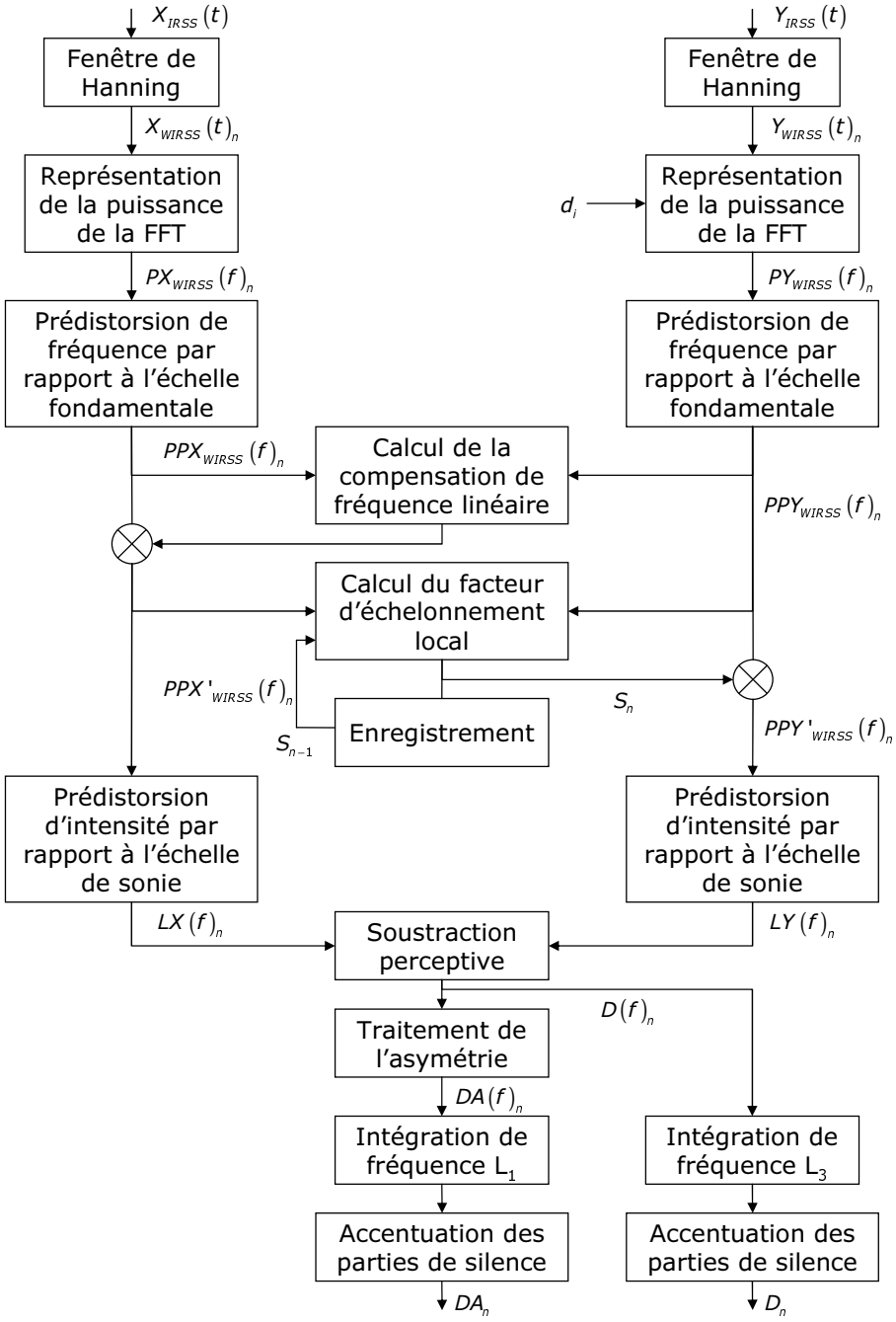


Figure B.2 : *Principe de fonctionnement du modèle psychoacoustique utilisé dans le modèle PESQ (*Perceptual Evaluation of Speech Quality*) [UIT-T Rec. P.862 2001] - Première partie*

B.2.2.2 Transformation temps-fréquence

Un fenêtrage de Hanning est appliqué aux signaux échelonnés et filtrés $X_{IRSS}(t)$ et $Y_{IRSS}(t)$. Les trames temporelles durent 32 ms et deux trames successives se chevauchent à 50%. Les signaux fenêtrés $X_{WIRSS}(t)$ et $Y_{WIRSS}(t)$ sont ensuite transformés par FFT. Leurs densités spectrales de puissance sont notées $PX_{WIRSS}(f)_n$ et $PY_{WIRSS}(f)_n$, où n est l'indice de trame et f l'indice de fréquence.

B.2.2.3 Prédistorsion et densité de puissance fondamentale

La prédistorsion permet de passer de l'échelle des fréquences en Hertz à l'échelle psychophysique des tonies dans le domaine des bandes critiques en Bark, pour obtenir une représen-

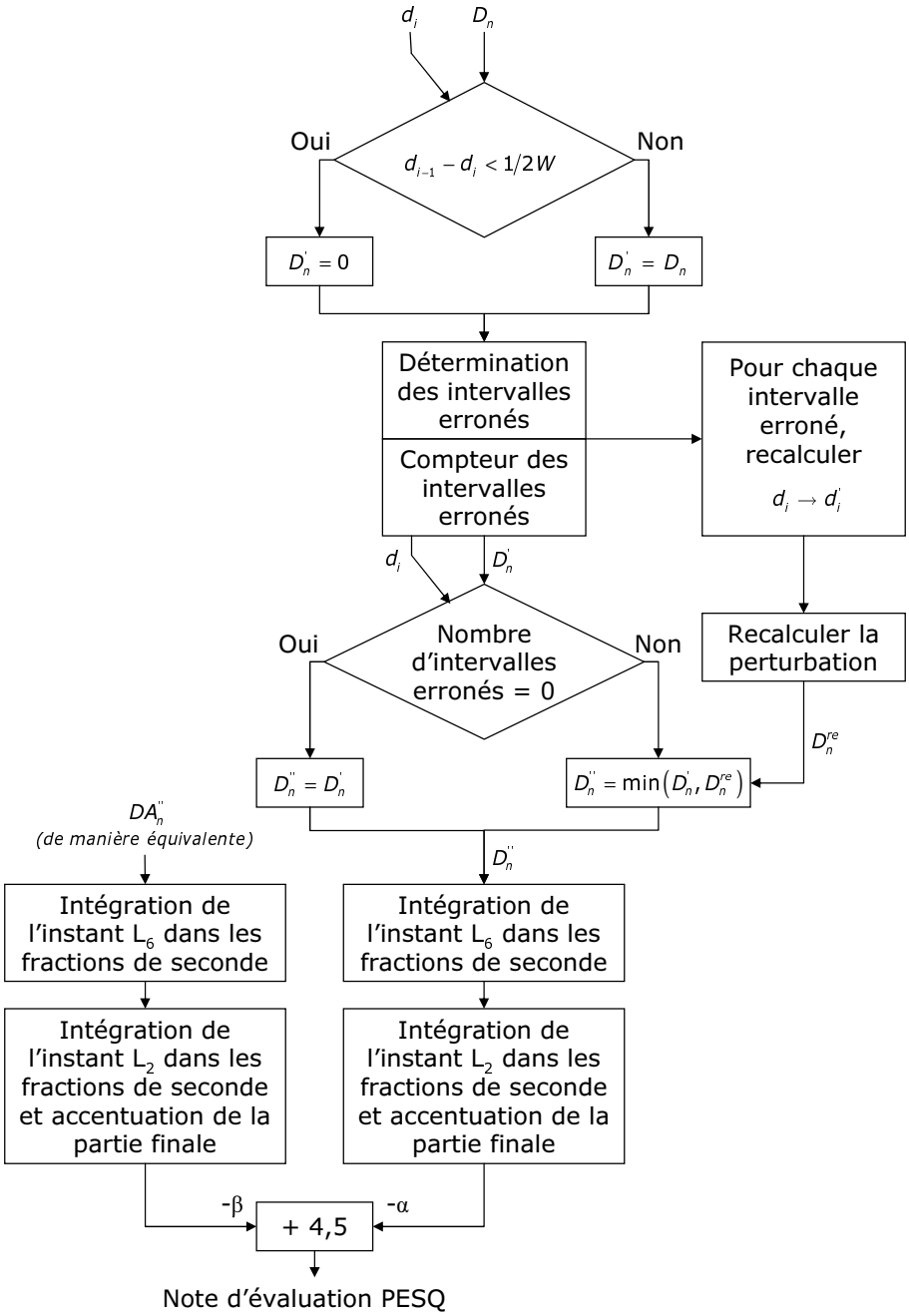


Figure B.3 : *Principe de fonctionnement du modèle psychoacoustique utilisé dans le modèle PESQ (*Perceptual Evaluation of Speech Quality*) [UIT-T Rec. P.862 2001] - Seconde partie.* NB : W = longueur d'une trame FFT en nombre d'échantillons

tation de la densité de puissance fondamentale trame par trame. L'échelle des bandes critiques est d'abord subdivisée en bandes d'intervalle égal et, pour chaque bande, une valeur (échantillon) de densité de puissance fondamentale est calculée à partir des échantillons de densité de puissance spectrale dans la bande correspondante sur l'échelle en Hertz. Les densités de puissance fondamentale sont notées $PPX_{WIRSS}(f)_n$ et $PPY_{WIRSS}(f)_n$.

B.2.2.4 Compensations

L'effet du filtrage et des variations de gain de courte durée sont partiellement compensées par le traitement des densités de puissance fondamentale trame par trame. Les densités de puissance fondamentale avec compensation partielle $PPX'_{WIRSS}(f)_n$ et $PPY'_{WIRSS}(f)_n$ sont

alors obtenues.

B.2.2.5 Densité de sonie

Les densités de sonie $LX(f)_n$ et $LY(f)_n$ sont calculées à partir d'une fonction de compression donnée par Zwicker [Zwicker et Feldtkeller 1981]

$$LX(f)_n = S_l \left(\frac{P_0(f)}{0.5} \right)^{\gamma} \left[\left(0.5 + 0.5 \frac{PPX'_{WIRSS}(f)_n}{P_0(f)} \right)^{\gamma} - 1 \right] \tag{B.1}$$

où $P_0(f)$ est le seuil absolu d'audition et S_l est le facteur d'échelonnement en sonie. Au-dessus de 4 Bark, la puissance de Zwicker, γ, est de 0.23. Au-dessous de 4 Bark, la puissance de Zwicker croît légèrement pour tenir compte de l'effet dit de recrutement.

B.2.2.6 Densité de perturbation

La différence entre les densités de sonie $LX(f)_n$ et $LY(f)_n$ est ensuite calculée. Celle-ci est corrigée de telle sorte que les petites différences de sonie soient moins perçues en présence de signaux ayant une valeur élevée de sonie (masquage). Il en résulte une densité de perturbation en fonction du temps (nombre de fenêtres n) et de la fréquence f, $D(f)_n$.

B.2.2.7 Traitement de l'asymétrie

Le phénomène d'asymétrie traduit le fait qu'une dégradation additive est plus audible et gênante qu'une dégradation soustractive pour le cerveau humain. Le modèle PESQ prend en compte cette asymétrie en donnant plus de poids aux dégradations additives qu'aux dégradations soustractives. Cet effet est modélisé en calculant la densité de perturbation asymétrique $DA(f)_n$ par trame et en multipliant la densité de perturbation $D(f)_n$ par un facteur d'asymétrie. Ce facteur d'asymétrie est égal au rapport des densités de puissance fondamentale du signal dégradé et du signal initial élevé à la puissance de 1.2. S'il est inférieur à 3, le facteur d'asymétrie est mis à zéro. S'il est supérieur à 12, il est écrêté à cette valeur. Ainsi, seules demeurent les cellules temps-fréquence, sous forme de valeurs non nulles, pour lesquelles la densité de puissance fondamentale du signal dégradé est supérieure à la densité de puissance fondamentale du signal initial.

B.2.2.8 Accentuation des parties de silence

Les dégradations intervenant dans les segments de non-activité vocale sont également prises en compte mais avec une pondération moindre.

B.2.2.9 Intégration en temps et fréquence

Enfin, l'intégration en temps et fréquence des « différences audibles » est non linéaire afin de modéliser le fait que des erreurs isolées (en temps et/ou fréquence) ont un impact perceptif plus fort que des erreurs uniformément réparties.

B.2.2.10 Calcul du score PESQ

La note d'évaluation PESQ finale est une combinaison linéaire de la valeur de perturbation moyenne et de la valeur de perturbation asymétrique moyenne.

B.2.3 Performances

Pour les 22 tests subjectifs utilisés à l'UIT pour évaluer la performance de PESQ, le coefficient de corrélation moyen s'est établi à 0.935 et la distribution moyenne des erreurs résiduelles a montré que l'erreur absolue était inférieure à une note MOS de 0.25 (0.25 sur une échelle de 5 points) pour 69.2% des conditions et inférieure à une note MOS de 0.5. Pour les 8 tests utilisés pour la validation finale (inconnus pendant l'élaboration de PESQ), le coefficient de corrélation moyen s'est également établi à 0.935 et l'erreur absolue était inférieure à une note MOS de 0.25 pour 72.3% des conditions et inférieure à une note MOS de 0.5 pour 91.1% des conditions.

Annexe C

Modèle objectif PESQM

La majorité des méthodes objectives d'évaluation de la qualité vocale portent sur la qualité en contexte d'écoute [UIT-T Rec. P.862 2001] [UIT-T Rec. P.563 2004], *i.e.* sur la façon dont nous percevons la parole d'une autre personne qui parle. Cependant, en situation de conversation, nous ne percevons pas seulement la voix de notre interlocuteur, mais aussi notre propre voix. Ce lien « bouche - oreille » constitue une boucle de rétroaction qui nous permet d'adapter notre voix à notre environnement. Ainsi, lorsque nous parlons dans un environnement bruité, nous avons tendance à élever la voix pour compenser la perte de niveau due au bruit (phénomène connu sous le nom d'effet Lombard). Cette boucle de rétroaction joue donc un rôle primordial dans notre production de parole et dans notre confort de locution. Ainsi, en se plaçant dans le contexte des communications téléphoniques, des phénomènes altérant la perception de notre propre voix, tels que l'écho, peuvent s'avérer très gênants et dégrader la qualité de la conversation.

Appel et Beerends [Appel et Beerends 2002] ont développé un modèle perceptuel de la qualité de locution, qui n'est pas normalisé à l'heure actuelle. Ce modèle, nommé PESQM (*Perceptual Echo and Sidetone Quality Measure*), a pour but d'évaluer la qualité de locution d'un lien téléphonique avec des dégradations causées par des distorsions d'effet local, de l'écho et des combinaisons de ces deux dégradations.

C.1 Principe

Le modèle PESQM est basé, comme PSQM [UIT-T Rec. P.861 1998] et PESQ [UIT-T Rec. P.862 2001], sur la différence des représentations internes d'un signal de référence et d'un signal dégradé. Cette transformation en représentation interne, présentée dans le paragraphe 1.3.2.1 du chapitre 1, consiste à transformer la représentation physique d'un signal (mesurée en décibels, secondes, Hertz) en une représentation psychophysique (mesurée en sones, secondes, Bark). Elle est effectuée sur les deux signaux de référence $x(t)$ et dégradé $y(t)$, puis leurs représentations internes sont comparées et une note de qualité est déduite à partir de leur différence.

Dans le cas de la qualité d'écoute, le signal de référence est le signal entré dans le système et le signal dégradé est le signal qui en sort. Dans le cas de la qualité de locution, le signal de référence est beaucoup plus difficile à définir. En effet, dans le contexte de la qualité d'écoute, les sujets peuvent tous écouter le même signal de référence qu'ils comparent au signal dégradé, alors que dans le contexte de la qualité de locution, *chaque sujet évalue le système en se référant à sa propre voix*, ce qui nécessite d'avoir un signal de référence par sujet en entrée du modèle et ne permet pas de donner un signal de référence unique au modèle. Pour remédier à ce problème, Appel et Beerends proposent d'utiliser le même enregistrement pour mesurer les différents systèmes et comparer les résultats du modèle ainsi obtenus avec les notes données par les sujets. Deux approches sont alors proposées :

- la première approche consiste à envoyer un signal de parole dans le système à un niveau *électrique* et à enregistrer le signal retourné par le système,
- la seconde approche consiste à utiliser un « simulateur de tête et torse » (HATS, *Head And Torso Simulator*) [UIT-T Rec. P.58 1996] pour simuler les caractéristiques d'un locuteur. Un signal de parole enregistré auparavant est appliqué à la bouche du HATS, ce qui permet d'obtenir une voix naturelle. En plaçant un combiné téléphonique sur le HATS, les signaux à son oreille peuvent être enregistrés au niveau *acoustique* : ils contiennent ainsi les signaux d'effet local et d'écho.

Selon l'approche choisie, les étapes du modèle seront différentes. Avec la première approche, le signal d'écho est enregistré séparément, et mélangé au signal de référence pour simuler le signal dégradé perçu par le sujet. Avec la seconde approche, le signal à l'entrée de l'oreille du HATS est un mélange de l'effet local acoustique avec un délai quasi nul et du signal de parole retourné par le système testé, qui peut contenir de l'écho et/ou de l'effet local distordu. Cependant, il n'est pas possible ici d'utiliser comme signal de référence le signal entré dans la bouche du HATS. En effet, cette approche étant acoustique, les enregistrements effectués à l'entrée de l'oreille (au niveau acoustique) sont très différents des signaux envoyés à l'entrée de la bouche (au niveau électrique). Cet effet de filtrage de l'effet local entre la bouche et l'oreille est pris en compte en enregistrant un signal de référence idéal (effet local normal, pas d'écho) à l'entrée de l'oreille du HATS à un niveau acoustique.

Pour l'approche acoustique, le fonctionnement du modèle est donné dans la figure C.1. Pour l'approche électrique, seules les premières étapes du fonctionnement diffèrent de l'approche acoustique et sont données dans la figure C.2.

C.2 Équations

C.2.1 Initialisations et calibrations

Avant de commencer, les échantillons de début et de fin du signal de référence sont déterminés. Le premier échantillon déclaré actif est celui dont l'amplitude (*i.e.* valeur absolue) plus les amplitudes des quatre échantillons précédents vaut 200 ou plus. Le dernier échantillon déclaré actif est le dernier échantillon dont l'amplitude plus les amplitudes des quatre échantillons suivants vaut 200 ou plus.

Il est également nécessaire de faire un étalonnage entre le niveau d'écoute et la sonie comprimée. Pour cela, une sinusoïde de référence avec une fréquence de 1 kHz et une amplitude de 40 dB SPL (c'est-à-dire -64 dBov, et une amplitude zéro-à-crête de 29.54) est générée[1]. Le premier étalonnage consiste à échelonner à 10^4 la valeur maximale de la représentation de puissance fondamentale contenue dans la sinusoïde d'étalonnage (c'est-à-dire, si la valeur $\max_j(P_{x_i}[j]) = 1$ pour un niveau acoustique de 0 dB, cette valeur $\max_j(P_{x_i}[j]) = 10000$ pour un niveau acoustique de 40 dB). Ce facteur d'échelonnement en puissance est calculé comme suit

$$S_p = \frac{10000}{\max_j(P_{x_i}[j])} \tag{C.1}$$

lorsque la puissance $P_{x_i}[j]$ est calculée pour la sinusoïde de référence.

Le deuxième étalonnage fixe à la valeur de 1 Sone la sonie comprimée de la sinusoïde de référence. Ce facteur est calculé comme suit

$$S_l = \frac{1}{L_{x_i}} \tag{C.2}$$

[1] Un signal de parole codé sur 16 bits varie entre $\pm 2^{16}/2 - 1 = \pm 32767$. Pour une sinusoïde, une amplitude zéro-à-crête de 32767 correspond à -3.01 dBov ou 100.9 dB SPL.

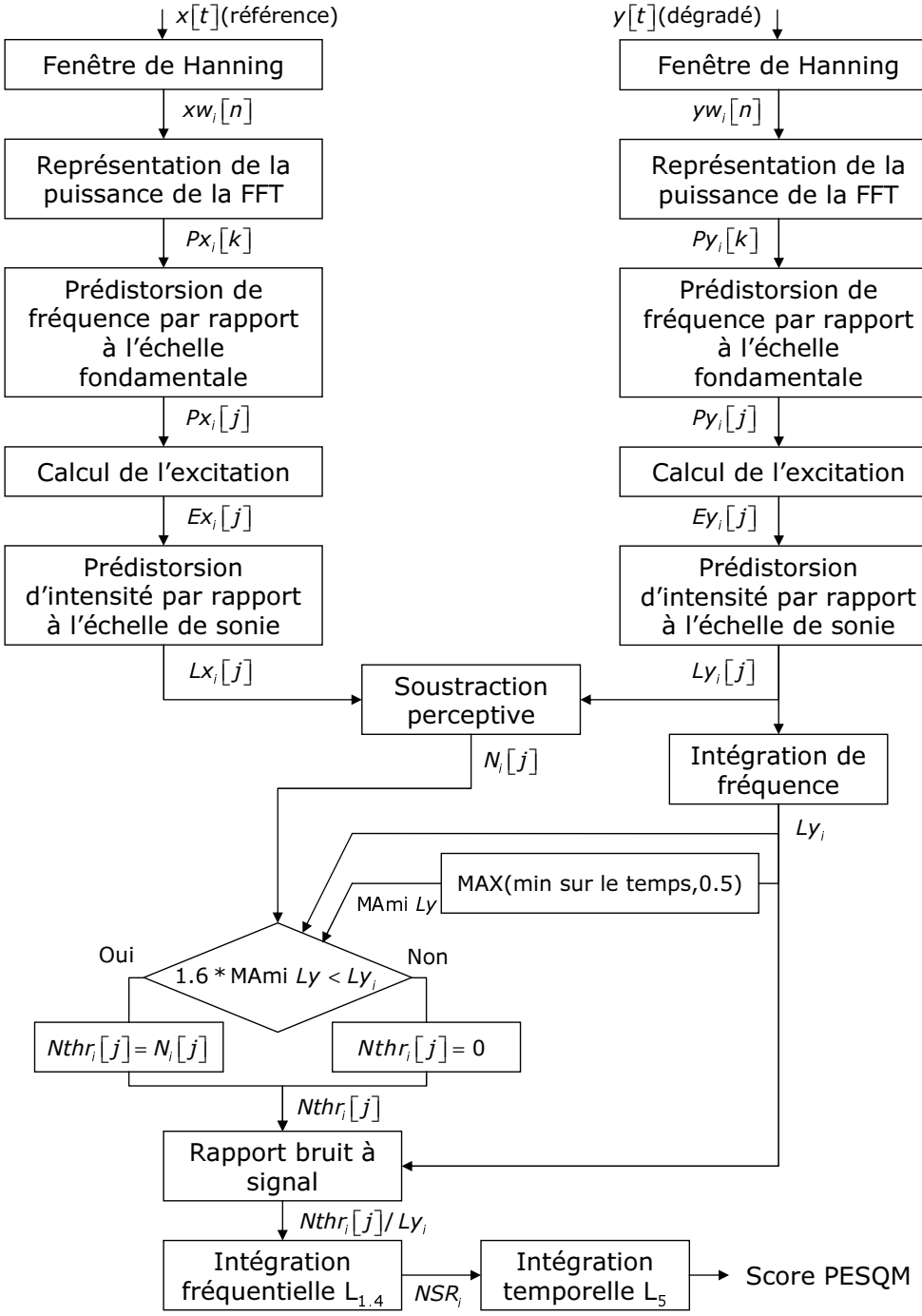


Figure C.1 : *Principe de fonctionnement du modèle PESQM (*Perceptual Echo and Sidetone Quality Measure*) en version acoustique [Appel et Beerends 2002]*

lorsque la valeur L_{x_i} est calculée pour la sinusoïde de référence.

Il est aussi nécessaire d'échelonner les données d'entrée à un niveau de conversation active de -26 dBov (ou au niveau acoustique d'écoute de 79 dB SPL), grâce à un détecteur d'activité vocale [UIT-T Rec. P.56 1993].

C.2.2 Fenêtrage et densité spectrale de puissance

Le point de départ de l'algorithme est une transformation temps-fréquence des deux signaux de référence $x(t)$ et dégradé $y(t)$. Les signaux $x(t)$ et $y(t)$ dans la trame i sont pondérés

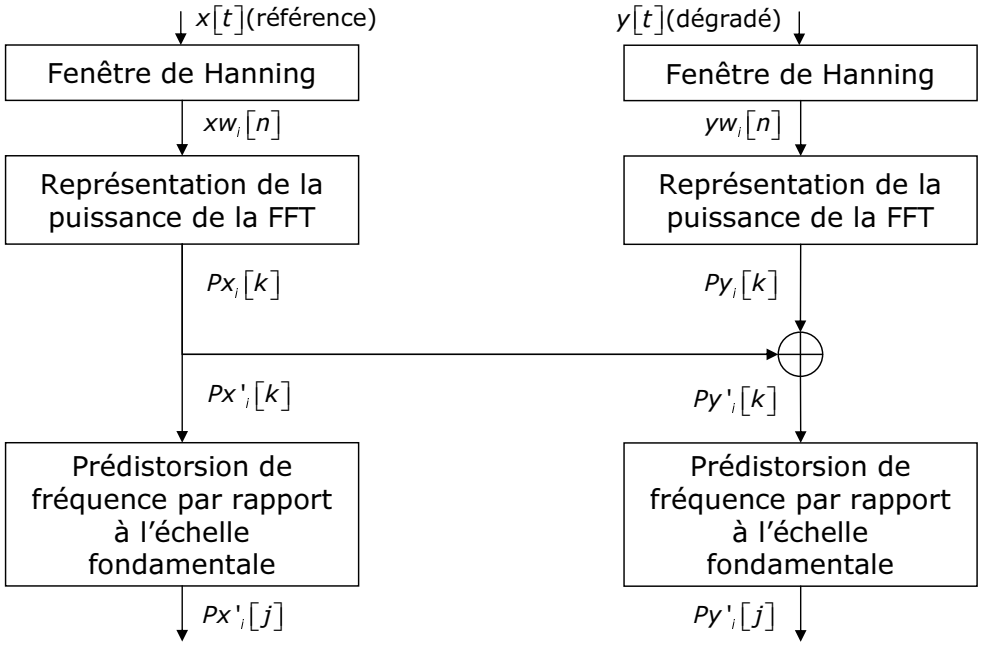


Figure C.2 : *Principe de fonctionnement du modèle PESQM (*Perceptual Echo and Sidetone Quality Measure*) en version électrique [Appel et Beerends 2002] - Premières étapes différant de la version acoustique*

par une fenêtre de Hanning de 32 ms :

$$\begin{aligned} x_{w_i}[n] &= x_i[n]w[n] \\ y_{w_i}[n] &= y_i[n]w[n] \end{aligned}$$

où $w[n] = 0.5\left(1 - \cos\left(\frac{2\pi i}{N_f}\right)\right)$ pour $0 \leq i \leq N_f - 1$. Le recouvrement entre deux trames temporelles successives est de 50%.

Les densités spectrales de puissance des signaux $x_{w_i}[n]$ et $y_{w_i}[n]$, représentées par $P_{x_i}[k]$ et $P_{y_i}[k]$ sont calculées au moyen de la FFT :

$$\begin{aligned} P_{x_i}[k] &= (\Re(X_i[k]))^2 + (\Im(X_i[k]))^2 \\ P_{y_i}[k] &= (\Re(Y_i[k]))^2 + (\Im(Y_i[k]))^2 \end{aligned}$$

où :

$$\begin{aligned} X_i[k] &= FFT(x_{w_i}[n]) \\ Y_i[k] &= FFT(y_{w_i}[n]) \end{aligned}$$

Les représentations en puissance des deux signaux, notées $P_{x_i}[k]$ et $P_{y_i}[k]$, sont alors obtenues.

C.2.3 Prédistorsion et densité de puissance fondamentale

La prédistorsion permet de passer de l'échelle des fréquences en Hertz (indice k) à l'échelle psychophysique des tonies dans le domaine des bandes critiques (indice j), pour obtenir une représentation de la densité de puissance fondamentale trame par trame. L'échelle des bandes critiques est d'abord subdivisée en bandes d'intervalle égal et, pour chaque bande, une valeur (échantillon) de densité de puissance fondamentale est calculée à partir des échantillons de densité de puissance spectrale dans la bande correspondante sur l'échelle en Hertz. Les densités

de puissance fondamentale, $P_{x_i}[j]$ et $P_{y_i}[j]$ pour la bande j dans la trame i, sont données par la formule suivante :

$$\begin{aligned} P_{x_i}[j] &= S_p \frac{\Delta f_j}{\Delta z} \frac{1}{I_l[j] - I_f[j] + 1} \sum_{k=I_f[j]}^{I_l[j]} P_{x_i}[k] \\ P_{y_i}[j] &= S_p \frac{\Delta f_j}{\Delta z} \frac{1}{I_l[j] - I_f[j] + 1} \sum_{k=I_f[j]}^{I_l[j]} P_{y_i}[k] \end{aligned}$$

où $I_f[j]$ est l'indice du premier échantillon et $I_l[j]$ celui du dernier sur l'échelle hertzienne pour la bande j, Δf_j étant la largeur de la bande j en Hertz, Δz étant la largeur de chaque sous-bande dans le domaine des bandes critiques, et S_p étant le facteur d'échelonnement en puissance, calculé précédemment. Les densités de puissance fondamentale correspondantes, notées $P_{x_i}[j]$ et $P_{y_i}[j]$, sont obtenues.

Le tableau C.1 donne les caractéristiques de chaque bande j (fréquence haute, premier et dernier échantillons de transformée FFT dans la bande j $I_f[j]$ et $I_l[j]$, seuil d'audition P_0).

C.2.4 Étalement dans le domaine fréquentiel

L'excitation provoquée par le stimulus sonore sur la membrane basilaire est déterminée grâce à une convolution de la densité de puissance fondamentale avec une fonction d'étalement fréquentiel. La forme de la fonction d'étalement fréquentiel dépend de l'intensité et de la fréquence. Cette opération n'est pas toujours nécessaire, elle n'est pas utilisée dans PSQM et PESQ. Ici, l'étalement fréquentiel est effectué en multipliant les densités de puissance fondamentale par une fonction d'étalement fréquentiel s

$$s = 22 + \frac{230}{f_m} - 0.2P_i[j], f_t > f_m \tag{C.3}$$

avec f_t la fréquence cible, f_m la fréquence du masquant, et $P_i[j]$ le niveau d'excitation à la fréquence f_m. L'excitation à la tonie j résultant de la puissance $P_i[\mu]$ à la tonie μ dans la trame i, notée $PE_i[j, \mu]$, est alors calculée

$$PE_i[j, \mu] = P_i[\mu] 10^{-s(j-\mu)\Delta z/10}, j > \mu. \tag{C.4}$$

En utilisant une addition non linéaire des composantes d'excitation individuelles, l'excitation suivante à la tonie j dans la trame i est obtenue

$$E_i[j] = \left[\sum_{\mu=\max(1,j-40)}^{j-1} (PE_i[j, \mu])^{0.8} \right]^{1.25}. \tag{C.5}$$

Les excitations $E_{x_i}[j]$ et $E_{y_i}[j]$ en sont alors déduites.

C.2.5 Densité de sonie

Les densités de sonie $L_{x_i}[j]$ et $L_{y_i}[j]$ sont calculées à partir d'une fonction de compression donnée par Zwicker [Zwicker et Feldtkeller 1981] :

$$\begin{aligned} L_{x_i}[j] &= S_l \left(\frac{P_0[j]}{0.5} \right)^{0.23} \left(\left(0.5 + 0.5 \frac{E_{x_i}[j]}{P_0[j]} \right)^{0.23} - 1 \right) \\ L_{y_i}[j] &= S_l \left(\frac{P_0[j]}{0.5} \right)^{0.23} \left(\left(0.5 + 0.5 \frac{E_{y_i}[j]}{P_0[j]} \right)^{0.23} - 1 \right) \end{aligned}$$

Les valeurs négatives de $L_{x_i}[j]$ et $L_{y_i}[j]$ sont mises à zéro. Les sonies instantanées (totales) des sonies comprimées L_{x_i} et L_{y_i} (exprimées en sones comprimés) sont calculées :

$$\begin{aligned} L_{x_i} &= \sum_{j=1}^{N_b} L_{x_i}[j]\Delta z \\ L_{y_i} &= \sum_{j=1}^{N_b} L_{y_i}[j]\Delta z \end{aligned}$$

C.2.6 Densité de perturbation due au bruit

La densité de perturbation due au bruit $N_i[j]$ dans la bande de tonie j et la trame i est calculée

$$N_i[j] = |L_{y_i}[j] - L_{x_i}[j]| - 0.01 \tag{C.6}$$

où 0.01 sone représente le bruit interne. Les valeurs négatives de $N_i[j]$ sont mises à zéro.

C.2.7 Suppression du bruit

L'étape suivante du modèle concerne l'influence du masquage par un bruit de fond. Dans l'évaluation de la qualité de locution, le bruit peut être, contrairement à ce qui se passe dans l'évaluation de la qualité d'écoute, bénéfique. En effet, un bruit de fond peut masquer un éventuel écho et ainsi améliorer la qualité de locution. Il est donc nécessaire de déterminer un seuil à partir duquel le bruit de fond est gênant. L'idée clé utilisée dans PESQM est de dire que, puisque les sujets auront toujours des intervalles de silence dans leur parole, la sonie minimale du signal dégradé dans le temps est quasiment complètement causée par le bruit de fond. Ce minimum peut donc servir pour obtenir un seuil pour lequel toutes les trames ayant une sonie inférieure à ce seuil sont mises à zéro. Si ce minimum est inférieur à 0.5 sone, le seuil est pris à 0.5 sone. Le résultat est une fonction de densité de dégradation $Nthr_i[j]$.

C.2.8 Calcul du score PESQM

Afin d'obtenir une corrélation élevée entre les notes subjectives et objectives, un rapport densité de sonie du bruit à sonie du signal doit être utilisé au lieu de la sonie du bruit. De plus, il a été noté que les sujets attribuaient un poids plus fort aux dégradations locales fortes, qui déterminent apparemment la qualité globale. Un tel poids des dégradations fortes peut être obtenu en utilisant une norme L_p, équivalant à l'addition non linéaire utilisée dans le calcul de l'excitation. Alors

$$NSR_i = \left[\frac{1}{N_b} \sum_{j=1}^{N_b} \left(\frac{Nthr_i[j]}{Ly_i} \right)^p \right]^{\frac{1}{p}} . \tag{C.7}$$

Pour obtenir le score PESQM final le NSR_i est cumulé sur toutes les trames

$$PESQM = \left(\frac{1}{N} \sum_{i=1}^{N} NSR_i^q \right)^{\frac{1}{q}} . \tag{C.8}$$

Les paramètres p et q du modèle PESQM sont optimisés afin que le score PESQM soit le mieux corrélé possible aux données subjectives. Avec leur base de données, Appel et Beerends ont obtenu $p = 1.4$ et $q = 5$ [Appel et Beerends 2002].

C.3 Performances

Le modèle PESQM n'étant pas normalisé à l'UIT-T, nous ne disposons pas de performances officielles, contrairement au modèle PESQ. Dans [Appel et Beerends 2002], les auteurs rapportent une corrélation moyenne de 0.97 pour des conditions avec écho, distorsion de l'effet local, bruit de fond et écho, écho et distorsion de l'effet local.
Dans la contribution à l'UIT-T [van Vugt et Beerends 2003], il est mentionné une corrélation moyenne de 0.87 pour des conditions avec écho, distorsion de l'effet local, bruit de fond, et avec différents terminaux et différents réseaux (VoIP, RTC, GSM).

Numéro de la bande j	**Fréquence haute (Hz)**	**Échantillon** $I_f[j]$	**Échantillon** $I_l[j]$	**Seuil d'audition** P_0
0	15.6	0	0	-
1	46.9	1	1	3.89e+07
2	78.1	2	2	1.12e+06
3	109.4	3	3	1.26e+05
4	140.6	4	4	1.86e+04
5	171.9	5	5	6.17e+03
6	203.1	6	6	2.29e+03
7	234.4	7	7	9.33e+02
8	265.6	8	8	4.37e+02
9	296.9	9	9	2.29e+02
10	328.1	10	10	1.29e+02
11	359.4	11	11	7.76e+01
12	390.6	12	12	4.27e+01
13	421.9	13	13	3.02e+01
14	453.1	14	14	2.19e+01
15	484.8	15	15	1.66e+01
16	519.2	16	16	1.32e+01
17	553.6	17	17	1.07e+01
18	590.8	18	18	8.91e+00
19	631.2	19	20	7.59e+00
20	672.9	21	21	6.31e+00
21	716.6	22	22	5.62e+00
22	760.4	23	24	5.13e+00
23	804.6	25	25	4.68e+00
24	851.4	26	27	4.37e+00
25	898.3	28	28	4.17e+00
26	947.0	29	30	4.07e+00
27	997.0	31	31	3.98e+00
28	1051	32	33	3.98e+00
29	1108	34	35	3.98e+00
30	1168	36	37	3.98e+00
31	1231	38	39	3.98e+00
32	1297	40	41	4.07e+00
33	1366	42	43	4.27e+00
34	1437	44	45	4.47e+00
35	1509	46	48	4.68e+00
36	1582	49	50	5.01e+00
37	1658	51	53	5.37e+00
38	1736	54	55	5.62e+00
39	1817	56	58	5.89e+00
40	1902	59	60	6.31e+00
41	1991	61	63	6.61e+00
42	2084	64	66	6.92e+00
43	2184	67	69	7.24e+00
44	2289	70	73	7.59e+00
45	2401	74	76	7.76e+00
46	2520	77	80	7.94e+00
47	2647	81	84	7.94e+00
48	2781	85	88	7.94e+00
49	2922	89	93	7.94e+00
50	3069	94	98	8.13e+00
51	3225	99	103	8.13e+00
52	3392	104	108	8.32e+00
53	3572	109	114	8.32e+00
54	3765	115	120	8.32e+00
55	3971	121	127	8.32e+00
56	4193	128	134	8.32e+00

Tableau C.1 : *Caractéristiques de chaque bande j (avec une fréquence d'échantillonnage de 16 kHz)*

Annexe D

Régression linéaire

D.1 Modèle

La régression linéaire multiple consiste à expliquer au mieux une grandeur y (la réponse) en fonction d'autres grandeurs x (les régresseurs). Un ensemble de n données (les individus) constituées de paires (x_i, y_i), $1 \leq i \leq n$, est disponible. La réponse et les p régresseurs sont arrangés dans des vecteurs

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \dots & x_{np} \end{pmatrix}. \tag{D.1}$$

Le modèle de régression linéaire multiple est

$$y = X\beta + u \tag{D.2}$$

où β est le vecteur des coefficients du modèle et $u = (u_1, \dots, u_n)$ est le vecteur des résidus (variables aléatoires) modélisant l'inadéquation des mesures au modèle.

Le but de la régression linéaire multiple est de trouver la meilleure estimation des $k = p-1$ coefficients β, notée $\hat{\beta}$, qui minimise le vecteur des résidus u, soit la différence entre les réponses observées y_i et les prédictions correspondantes du modèle $\hat{y_i}$. L'estimateur de β aux moindres carrés ordinaires est tel que

$$\hat{\beta} = \text{argmin}_\beta \|u_i\|^2 = \text{argmin}_\beta \sum_i u_i^2 = \text{argmin}_\beta \sum_i (y_i - x_i\beta)^2 . \tag{D.3}$$

L'estimateur solution est $\hat{\beta} = (X^T X)^{-1} X^T y$.

D.2 Suppositions

Le modèle linéaire multiple est basé sur plusieurs suppositions :

1. Linéarité : la relation entre les régresseurs et la réponse est linéaire.
2. X non stochastique : $E[u_i X_{i,k}] = 0$. Les erreurs u_i sont décorrélées avec les régresseurs individuels.
3. Moyenne nulle : $E[u_i] = 0$. L'espérance des résidus est nulle.
4. Variance constante : $E[u_i^2] = \sigma^2$. La variance des résidus est constante.
5. Non autoregressive : $E[u_i u_{i-m}] = 0$, $m \neq 0$. Les résidus sont aléatoires, ou décorrélés dans le temps.
6. Normalité : l'erreur a une distribution normale. Cette supposition doit être vérifiée pour que les tests de significativité des coefficients soient valides.

D.3 Statistiques

D.3.1 Sommes des carrés (SC)

Les carrés des erreurs entre les réponses observées et prédites sont utilisés pour décrire la qualité du modèle de régression. La somme des carrés totale peut être partitionnée en composantes dues à la régression et aux résidus

$$\sum_i^n (y_i - \bar{y})^2 = \sum_i^n (y_i - \hat{y}_i)^2 + \sum_i^n (\hat{y}_i - \bar{y})^2 \quad \text{(D.4)}$$

$$SCT = SCE + SCR \quad \text{(D.5)}$$

où $\bar{y}$ est la moyenne des y_i, $1 \leq i \leq n$. Cette partition signifie que la variabilité totale dans les réponses observées (SCT) est égale à la variabilité aléatoire non expliquée par le modèle (SCE) plus la variabilité systématique expliquée par le modèle de régression (SCR). Ajustés avec les degrés de liberté (dl) correspondants, les trois termes SCR, SCT et SCE donnent :

- $CMT = \frac{SCT}{n-1}$ = carré moyen (MC) total
- $CME = \frac{SCE}{n-p}$ = carré moyen d'erreur
- $CMR = \frac{SCR}{k}$ = carré moyen de régression.

D.3.2 Coefficient de détermination

La proportion de variance expliquée par le modèle de régression est résumée par le coefficient R^2

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}. \quad \text{(D.6)}$$

D.3.3 Coefficient de détermination ajusté

Le coefficient R^2 pour une régression peut être artificiellement élevé simplement en ajoutant de plus en plus de régresseurs dans le modèle. Le coefficient ajusté $\bar{R}^2$ permet de compenser cette augmentation artificielle de la performance du modèle

$$\bar{R}^2 = 1 - \frac{(n-1)}{(n-p)}(1 - R^2). \quad \text{(D.7)}$$

D.3.4 Test de Fisher

Le test de Fisher (F-test) de niveau α permet de tester l'hypothèse H_0 : « $\beta_0 = \beta_1 = \ldots = \beta_k = 0$ », c'est-à-dire d'indiquer si au moins un des p régresseurs est significatif. La statistique de Fisher de la régression est donnée par $F = \frac{CMR}{CME}$ et est à comparer avec $F_{k,n-p}$ de la distribution F. L'hypothèse H_0 est rejetée si $P = 1 - F_{k,n-p}$ est inférieur au niveau α (par exemple, $\alpha = 5\%$).

D.3.5 Table d'analyse de variance (ANOVA)

La table d'analyse de variance résume les statistiques du modèle de régression sous forme de tableau :

Facteur	SC	dl	CM	F	P>F
Modèle	SCE	$p-1$	CME	CMR/CME	$1-F_{k,n-p}$
Résidu	SCR	$n-p$	CMR		
Total	SCT	$n-1$	CMT		

D.3.6 Intervalles de confiance des coefficients estimés

Si les suppositions de la régression sur les résidus sont satisfaites, la distribution des coefficients de regression estimés est normale avec une variance proportionnelle au carré moyen d'erreur (CME). La variance des estimateurs dépend aussi des variances et covariances des régresseurs

$$\hat{\sigma}(\hat{\beta}_j)^2 = CME[(X^T X)^{-1}]_{jj},\ 1 \leq j \leq k. \tag{D.8}$$

D.3.7 Test de nullité d'un coefficient

Le test de niveau α permet de décider si β_j est différent de zéro

$$\frac{|\hat{\beta}_j|}{\hat{\sigma}(\hat{\beta}_j)} \geq t_{n-p}(1-\alpha/2) \tag{D.9}$$

où $t_{n-p}(.)$ désigne la fonction quantile de la loi de Student de paramètre $n-p$, c'est-à-dire $\alpha = P(|T_{n-p}| > t_{n-p}(1-\alpha/2))$.

Les résultats sont résumés dans une table d'analyse des coefficients :

Facteur	Estimée ($\hat{\beta}_j$)	Écart-type ($\hat{\sigma}(\hat{\beta}_j)$)	Pr($>|t|$)
Facteur 1	...	...	...
⋮	⋮	⋮	⋮
Facteur p	...	...	...

où la dernière colonne contient α_j solution de

$$|\hat{\beta}_j| = \hat{\sigma}(\hat{\beta}_j)\, t_{n-p}(1-\alpha_j/2). \tag{D.10}$$

L'hypothèse H_0 : « $\beta_j = 0$ » est refusée si α_j est inférieur au niveau α.

D.4 Analyse des résidus

L'analyse des résidus permet de vérifier que les résidus remplissent les suppositions faites dans le paragraphe D.2, à savoir que les résidus :

1. Sont normalement distribués.
2. Sont non autocorrélés.
3. Sont corrélés avec aucun des régresseurs.
4. Ont une variance constante (hétéroscedasticité).

La normalité des résidus se vérifie en traçant leur histogramme. L'autocorrélation des résidus se vérifie en traçant leur fonction d'autocorrélation. La non corrélation avec les régresseurs se vérifie en traçant les résidus en fonction de chacun des régresseurs. L'hétéroscedasticité des résidus se vérifie en traçant les résidus en fonction de la réponse et des régresseurs.

D.5 Multicolinéarité

Dans une régression, les régresseurs sont aussi appelés des « variables indépendantes », cependant ceci est rarement le cas en pratique. Le terme de « multicolinéarité » est donc réservé pour les cas où l'intercorrélation entre les régresseurs est élevée. La multicolinéarité n'invalide pas le modèle de régression, mais entraîne des effets négatifs.

Tout d'abord la variance des coefficients de régression (proportionnelle à l'intercorrélation des régresseurs) peut être augmentée au point que les coefficients individuels ne soient plus statistiquement significatifs. Les indications de la multicolinéarité peuvent être :

- des coefficients de régression ayant le « mauvais » signe,
- l'ajout/le retrait d'un régresseur entraînant de grands changements des coefficients de régression,
- des coefficients de régression n'étant pas significativement différents de zéro.

La multicolinéarité peut être détectée par l'étude de la matrice de corrélation des régresseurs, ainsi que par le calcul des facteurs d'inflation de la variance (VIF). Le facteur d'inflation de la variance associé au régresseur x_j est calculé avec

$$VIF_j = \frac{1}{(1 - R_j^2)} \tag{D.11}$$

où R_j^2 est le coefficient de détermination (R^2) de la régression de x_j par les $k-1$ autres régresseurs du modèle. Ainsi, plus la variation de x_j peut être expliquée par les autres régresseurs, plus R_j^2 sera proche de 1 et plus VIF_j va augmenter. Ce problème peut être réglé en éliminant un ou plusieurs régresseurs. Les méthodes de sélection des régresseurs permettent de choisir quel(s) régresseur(s) éliminer ou garder.

D.6 Sélection des régresseurs

Les méthodes de sélection des régresseurs sont utiles pour réduire le nombre de régresseurs et également pour remédier à des problèmes de multicolinéarité.

D.6.1 Backward elimination

Les étapes de la méthode « backward elimination » sont les suivantes :

1. Construire le modèle complet de régression avec tous les prédicteurs.
2. Calculer les corrélations partielles et les statistiques de Fisher F associées à chaque prédicteur avec la variable dépendante y.
3. Éliminer le prédicteur ayant la corrélation partielle non significative la plus faible avec y.
4. Réestimer l'équation de régression avec les prédicteurs restants.
5. Répéter les étapes 3 et 4 jusqu'à ce que tous les prédicteurs dans la régression soient significatifs et que tous ceux en dehors de la régression soient non significatifs.

D.6.2 Forward selection

Les étapes de la méthode « forward selection » sont les suivantes :

1. Calculer la matrice d'intercorrélation incluant y et tous les prédicteurs.
2. Sélectionner le prédicteur ayant la corrélation la plus significative avec y.
3. Estimer la régression de y avec ce prédicteur.
4. Calculer les corrélations partielles et les statistiques de Fisher F associées à chaque prédicteur restant avec la variable dépendante y.

5. Inclure dans l'équation estimée à l'étape 3 le prédicteur ayant la corrélation la plus significative avec y.
6. Répéter les étapes 4 et 5 jusqu'à ce que tous les prédicteurs dans la régression soient significatifs et que tous ceux en dehors de la régression soient non significatifs.

D.6.3 Stepwise regression

Les étapes de la méthode « stepwise regression » sont les suivantes :

1. Calculer la matrice d'intercorrélation incluant y et tous les prédicteurs.
2. Sélectionner le prédicteur ayant la corrélation la plus significative avec y.
3. Estimer la régression de y avec ce prédicteur.
4. Calculer les corrélations partielles et les statistiques de Fisher F associées à chaque prédicteur restant avec la variable dépendante y.
5. Inclure dans l'équation estimée à l'étape 3 le prédicteur ayant la corrélation la plus significative avec y.
6. Vérifier si tous les prédicteurs dans l'équation à ce niveau sont toujours significatifs.
7. Répéter les étapes 4 à 6 jusqu'à ce que tous les prédicteurs dans la régression soient significatifs et que tous ceux en dehors de la régression soient non significatifs.

D.7 Méthode de bootstrap

Les étapes de la méthode de bootstrap d'estimation de l'intervalle de confiance de la moyenne sont les suivantes [Zoubir et Boashash 1998] :

1. *Expérience.* Faire l'expérience et collecter l'échantillon $X = \{X_1, X_2, \ldots, X_N\}$. Calculer la moyenne μ_X de X.
2. *Rééchantillonnage.* Utiliser un générateur de nombre pseudo-aléatoire, tirer un échantillon X^* de N valeurs, avec remise, à partir de X.
3. *Estimation.* Calculer la moyenne μ_{X^*} de X^*.
4. *Répétition.* Répéter les étapes 1 et 2 un grand nombre de fois pour obtenir un total de K estimées μ_{X^*}.
5. *Approximation des distributions.* Ranger μ_{X^*} dans l'ordre croissant.
6. *Calcul de l'intervalle de confiance.* L'intervalle de confiance de la moyenne μ_X de X à (1-α)100% est donné par μ_{X^*} aux positions K$\alpha/2$ et K(1-$\alpha/2$).

Annexe E

Matériel de test

Agence de voyage

Votre nom: ***Chevallier***

Voyage pour une semaine en Méditerranée

Voyage de dernière minute : à partir de demain,
pas en Espagne,
le moins cher possible

Prix : _____ €

Jour de départ : _____ à ____h_____

Destination : _____

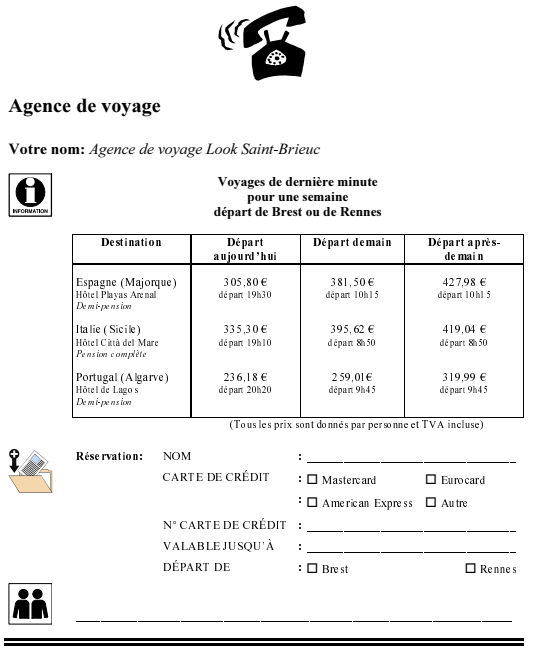
Hôtel : _____

Réservation : Départ de Brest,
Mastercard,
N° de la carte de crédit: 9685 4712 0951 2781,
Valable jusqu'en 10/07

Combien de temps avant le départ dois-je être à l'aéroport?

_____

Figure E.1 : *Exemple de scénario de conversation (*Short Conversation Test*, SCT) - Participant 1*



Agence de voyage

Votre nom: *Agence de voyage Look Saint-Brieuc*

Voyages de dernière minute
pour une semaine
départ de Brest ou de Rennes

Destination	Départ aujourd'hui	Départ demain	Départ après-demain
Espagne (Majorque) Hôtel Playas Arenal *Demi-pension*	305,80 € départ 19h30	381,50 € départ 10h15	427,98 € départ 10h15
Italie (Sicile) Hôtel Città del Mare *Pension complète*	335,30 € départ 19h10	395,62 € départ 8h50	419,04 € départ 8h50
Portugal (Algarve) Hôtel de Lagos *Demi-pension*	236,18 € départ 20h20	259,01 € départ 9h45	319,99 € départ 9h45

(Tous les prix sont donnés par personne et TVA incluse)

Réservation: NOM : _____

CARTE DE CRÉDIT : ☐ Mastercard ☐ Eurocard

: ☐ American Express ☐ Autre

N° CARTE DE CRÉDIT : _____

VALABLE JUSQU'À : _____

DÉPART DE : ☐ Brest ☐ Rennes

_____

Figure E.2 : *Exemple de scénario de conversation (*Short Conversation Test*, SCT) - Participant 2*

Critère testé	Bruit de fond	Défauts	Écho	Voix	Niveau vocal	Possibilité d'interruption	Double parole	Effet local	Qualité globale
Question posée	Dans la conversation que vous venez d'avoir, un bruit de fond (dans votre téléphone) était-il :	Dans la conversation que vous venez d'avoir, des défauts, de type coupures, craquements, etc. (dans votre téléphone) étaient-ils :	Comment jugez-vous la dégradation due à l'écho de votre propre voix ?	Comment trouvez-vous la voix de votre interlocuteur ?	Comment trouvez-vous le niveau de la voix de votre interlocuteur ?	Comment qualifiez-vous l'effort nécessaire pour interrompre votre interlocuteur ?	Comment jugez-vous la qualité de la communication quand vous et votre interlocuteur parliez en même temps ?	Pendant que vous parliez, comment qualifieriez-vous votre confort d'écoute avec le combiné ?	Comment évaluez-vous la qualité globale de la communication ?
Libellés des réponses possibles	imperceptible	imperceptibles	imperceptible	fidèle, naturelle	beaucoup plus fort que préféré	sans aucun effort	excellente	excellent	excellente
	perceptible mais non gênant	perceptibles mais non gênants	perceptible mais non gênante	presque naturelle	plus fort que préféré	sans effort sensible	bonne	bon	bonne
	légèrement gênant	légèrement gênants	légèrement gênante	un peu déformée	selon préférence	avec un effort modéré	moyenne	moyen	moyenne
	gênant	gênants	gênante	assez déformée	plus faible que préféré	avec un effort important	médiocre	médiocre	médiocre
	très gênant	très gênants	très gênante	déformée, dénaturée	bien plus faible que préféré	interruption impossible	mauvaise	mauvais	mauvaise

Tableau E.1 : *Questions posées lors des tests subjectifs*

Annexe F

Liste de publications associées

Conférences internationales avec comité de lecture et actes

Guéguin M, Gautier-Turbin V, Gros L, Barriac V, Le Bouquin Jeannès R, Faucon G. "Study of the relationship between subjective conversational quality, and talking, listening and interaction qualities : towards an objective model of the conversational quality". In *4th International Conference Measurement of Audio and Video Quality in Networks (MESAQIN)*, p. 47-50, Prague, République Tchèque, Juin 2005.

Guéguin M, Le Bouquin Jeannès R, Faucon G, Barriac V. "Towards an objective model of the conversational speech quality". In *31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, p. 1229-1232, Toulouse, France, Mai 2006.

Guéguin M, Le Bouquin Jeannès R, Faucon G, Gautier-Turbin V, Barriac V. "A step further to objective modeling of conversational speech quality". In *14th European Signal Processing Conference (EUSIPCO)*, Florence, Italie, Septembre 2006.

Conférence nationale avec comité de lecture et actes

Guéguin M, Gautier-Turbin V, Barriac V, Le Bouquin Jeannès R, Faucon G. "Évaluation de la qualité vocale dans les télécommunications". In *XXVIèmes Journées d'Études de la Parole (JEP)*, p. 537-540, Dinard, France, Juin 2006.

Contributions à l'Union Internationale des Télécommunications

Guéguin M, Gros L, Gautier-Turbin V. "Report on a new subjective test on the relationships between listening, talking and conversational qualities when facing delay and echo". Contribution UIT-T COM 12-D45-E, Genève, Suisse, Janvier 2005.

Barriac V, Guéguin M. "Survey of existing subjective test results useful for the development and the training of objective models for the evaluation of speech quality in a conversational context". Contribution UIT-T COM 12-D133-E, Genève, Suisse, Juin 2006.

Guéguin M, Barriac V. "Towards an objective model of the conversational speech quality". Contribution UIT-T COM 12-D134-E, Genève, Suisse, Juin 2006.

Brevet

Barriac V, Guéguin M. "Modèle d'évaluation de qualité vocale en conversation". Demande de brevet français n^{o} 0653145 déposée le 27 juillet 2006.

Article

Guéguin M, Le Bouquin Jeannès R, Gautier-Turbin V, Faucon G, Barriac V. "Building a model for the objective evaluation of speech quality in the conversational context". *IEEE Transactions on Communications*, soumis.

Bibliographie

[Appel et Beerends 2002] Appel, R. et Beerends, J. G. (2002). On the quality of hearing one's own voice. *Journal of Audio Engineering Society*, 50(4):237–248.

[Barnwell 1980] Barnwell, T. P. (1980). Correlation analysis of subjective and objective measures for speech quality. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'80)*, 5:706–709.

[Barriac 1997] Barriac, V. (1997). Détermination de facteurs de dégradation à partir des résultats de tests d'écoute pour des codeurs mis en cascade. Contribution UIT-T COM 12-40.

[Barriac et Gilloire 1996] Barriac, V. et Gilloire, A. (1996). Procédé et dispositif de mesure sans intrusion de la qualité de transmission d'une ligne téléphonique. Brevet européen n° EP 0 741 471 B1.

[Barriac et Guéguin 2006] Barriac, V. et Guéguin, M. (2006). Survey of existing subjective test results useful for the development and the training of objective models for the evaluation of speech quality in a conversational context. Contribution UIT-T COM 12-D.133.

[Beerends *et al.* 2002] Beerends, J. G., Hekstra, A. P., Rix, A. W. et Hollier, M. P. (2002). Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment - Part II : Psychoacoustic model. *Journal of Audio Engineering Society*, 50(10):765–778.

[Beerends et Stemerdink 1992] Beerends, J. G. et Stemerdink, J. A. (1992). A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of Audio Engineering Society*, 40:963–978.

[Beerends et Stemerdink 1994] Beerends, J. G. et Stemerdink, J. A. (1994). A perceptual speech-quality measure based on a psychoacoustic sound representation. *Journal of Audio Engineering Society*, 42(3):115–123.

[Berger 2005] Berger, J. (2005). Requirements for a new model for objective speech quality assessment P.OLQA. Contribution UIT-T COM 12-D.75.

[BharrathSingh et Britt 2001] BharrathSingh, K. et Britt, R. (2001). Proposed Ie values for G.711, G.729 and G.729E under conditions of bursty packet loss. Contribution UIT-T COM 12-D.36.

[Blauert 1997] Blauert, J. (1997). *Spatial Hearing : the psychophysics of human sound localization*. The MIT Press, Cambridge MA, USA.

[Brady 1968] Brady, P. T. (1968). A statistical study of on-off patterns in 16 conversations. *Bell System Technical Journal*, 47:73–91.

[Brandenburg 1987] Brandenburg, K. (1987). Evaluation of quality for audio encoding at low bit rates. In *82nd AES Convention*. Papier n° 2433.

[Cermak 2002] Cermak, G. W. (2002). Subjective quality of speech over packet networks as a function of packet loss, delay and delay variation. *International Journal of Speech Technology*, 5:65–84.

[Cheung 1998] Cheung, K. (1998). Derivation of impairment factors for G.729-B, G.723.1 (6.3 kb/s & 5.3 kb/s). Contribution UIT-T COM 12-D.040.

[Clark 2001] Clark, A. D. (2001). Modeling the effects of burst packet loss and recency on subjective voice quality. In *IP Telephony Workshop (IPTEL'01)*.

[Colomes *et al.* 1995] Colomes, C., Lever, M., Rault, J. B. et Dehery, Y. F. (1995). A perceptual model applied to audio bit-rate reduction. *Journal of Audio Engineering Society*, 43:233–240.

[Coverdale et Cheung 1997] Coverdale, P. et Cheung, K. (1997). Impairment factor of GSM-EFR codec. Contribution UIT-T COM 12-D.007.

[Dvorak et James 2004] Dvorak, C. et James, J. (2004). Echo-free delay, VoIP speech quality and the E-model. Contribution UIT-T COM 12-D.214.

[ETSI 3GPP TR 26.935 2004] ETSI 3GPP TR 26.935 (2004). 3rd Generation Partnership Project ; Technical Specification Group Services and System Aspects ; Packet-switched conversational multimedia applications ; Performance characterisation of default codecs (Release 6).

[ETSI ETR 250 1996] ETSI ETR 250 (1996). Transmission and multiplexing (TM) ; Speech communication quality from mouth to ear for 3.1 kHz handset telephony across networks.

[ETSI Guide 201 377-1 2002] ETSI Guide 201 377-1 (2002). Speech Processing, Transmission and Quality Aspects (STQ) ; Specification and measurement of speech transmission quality ; Part 1 : Introduction to objective comparison measurement methods for one-way speech quality across networks.

[ETSI Guide EG 201 377-3 2003] ETSI Guide EG 201 377-3 (2003). Speech Processing, Transmission and Quality Aspects (STQ) ; Specification and measurement of speech transmission quality ; Part 3 : Non-intrusive objective measurement methods applicable to networks and links with classes of services.

[Gierlich et Diedrich 2000] Gierlich, H. W. et Diedrich, E. (2000). Auditory judgement of echo : talking and listening test in comparison to third party listening test. Contribution UIT-T COM 12-16.

[Gros 2001] Gros, L. (2001). *Évaluation subjective de la qualité vocale fluctuante.* Thèse de doctorat, Université d'Aix-Marseille II.

[Hammer 2006] Hammer, F. (2006). *Quality aspects of packet-based interactive speech communication.* Thèse de doctorat, Graz University of Technology.

[Hammer *et al.* 2005] Hammer, F., Reichl, P. et Raake, A. (2005). The well-tempered conversation : interactivity, delay and perceptual VoIP quality. In *IEEE International Conference on Communications (ICC)*.

[Holub 2006] Holub, J. (2006). Subjective experiments supporting conversational quality objective estimator design. Rapport technique confidentiel non publié.

[Hoth 1941] Hoth, D. F. (1941). Room noise spectra at subscribers' telephone locations. *Journal Acoustical Society of America*, 12:499–504.

[IEEE 1969] IEEE (1969). Recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246.

[Jekosch 1993] Jekosch, U. (1993). Speech quality assessment and evaluation. In *Eurospeech'93*, pages 1387–1394.

[Jekosch 2000] Jekosch, U. (2000). *Sprache hören und beurteilen. Qualitätsbeurteilung von Sprechtechnologien als Forschung- und Dienstleistungsaufgabe.* Thèse d'habilitation, Universität/Gesamthochschule Essen.

[Jekosch et Klaus 1997] Jekosch, U. et Klaus, H. (1997). Verification of the E-model : results of a pilot study. Contribution UIT-T COM 12-13.

[Jones et Atkinson 1998] Jones, C. et Atkinson, D. (1998). Development of opinion-based audiovisual quality models for desktop video-teleconferencing. In *6th IEEE International Workshop on Quality of Service*.

[Jones et Munhall 2002] Jones, J. A. et Munhall, K. G. (2002). The role of auditory feedback during phonation : studies of Mandarin tone production. *Journal of Phonetics*, 30:303–320.

[Juric 2001] Juric, P. (2001). Non-intrusive speech quality measurement. Contribution UIT-T COM 12-27.

[Karlsson 2001] Karlsson, C. (2001). The effect of packet losses on speech quality. Contribution UIT-T COM 12-D.15.

[Kitawaki *et al.* 2005] Kitawaki, N., Arayama, Y. et Yamada, T. (2005). Multimedia opinion model based on media interaction of audio-visual communications. In *Measurement of Audio and Video Quality in Networks (MESAQIN'05)*, pages 5–10.

[Kitawaki et Itoh 1991] Kitawaki, N. et Itoh, K. (1991). Pure delay effects on speech quality in telecommunications. *IEEE Journal on Selected Areas in Communications*, 9(4):586–593.

[Le Faucheur 2004] Le Faucheur, N. (2004). Caractérisation de l'efficacité des fonctions de réduction de bruit mises en œuvre dans les réseaux de télécommunications. Mémoire de master, Conservatoire National des Arts et Métiers.

[Lombard 1911] Lombard, E. (1911). Le signe de l'élévation de la voix. *Annales des maladies de l'oreille et du larynx*, 37(2):101–119.

[McAdams 1994] McAdams, S. (1994). *Audition : physiologie, perception et cognition*, volume 1, pages 283–344. Presses Universitaires de France, Paris, France.

[Möller 1997a] Möller, S. (1997a). Development of scenarios for a short conversation test. Contribution UIT-T COM 12-35.

[Möller 1997b] Möller, S. (1997b). The E-model : an analysis of the sources and comparison with published and new test results. Contribution UIT-T COM 12-37.

[Möller 1998] Möller, S. (1998). E-model predictions and the impairment factor principle for low-bitrate codecs and quantizing distorsion : analysis of test results. Contribution UIT-T COM 12-69.

[Möller 2000a] Möller, S. (2000a). Application of the new methodology for derivation of equipment impairment factors (P.833) : further results. Contribution UIT-T COM 12-15.

[Möller 2000b] Möller, S. (2000b). *Assessment and prediction of speech quality in telecommunications*. Kluwer Academic Publishers, Boston, USA.

[Möller 2001] Möller, S. (2001). Application of the new methodology for the derivation of equipment impairment factors (P.833) : additional results. Contribution UIT-T COM 12-D.14.

[Möller et Raake 2002] Möller, S. et Raake, A. (2002). Telephone speech quality prediction : Towards network planning and monitoring models for modern network scenarios. *Speech Communication*, 38:47–75.

[Mohamed 2003] Mohamed, S. A. (2003). *Évaluation automatique de la qualité des flux multimédias en temps réel : une approche par réseaux de neurones*. Thèse de doctorat, Université de Rennes 1.

[Osaka *et al.* 1992] Osaka, N., Kakehi, K., Iai, S. et Kitawaki, N. (1992). A model for evaluating talker echo and sidetone in a telephone transmission network. *IEEE Transactions on Communications*, 40(11):1684–1692.

[Paillard *et al.* 1992] Paillard, B., Mabilleau, P., Morisette, S. et Soumagne, J. (1992). Perceval : Perceptual evaluation of the quality of audio signals. *Journal of Audio Engineering Society*, 40:21–31.

[Pastrana-Vidal *et al.* 2003] Pastrana-Vidal, R., Colomes, C., Gicquel, J. C. et Cherifi, H. (2003). Caractérisation perceptuelle des interactions audiovisuelles : revue. In *COmpression et REprésentation des Signaux Audiovisuels (CORESA03)*.

[Pomy 2005] Pomy, J. (2005). Proposed Scope for P.CQO. TD27 (WP 2/12), UIT-T SG12 Q.20.

[Pomy 2006] Pomy, J. (2006). Status Report of Question 20/12. TD36rev2 (WP 2/12), UIT-T SG12 Q.20.

[Preminger et van Tasell 1995] Preminger, J. E. et van Tasell, D. J. (1995). Quantifying the relation between speech quality and speech intelligibility. *Journal of Speech and Hearing Research*, 38:714–725.

[Puglia 2005] Puglia, R. (2005). Influence of audio and video quality on subjective audiovisual quality - H.263 and Adaptive Multi Rate (AMR) coding. Mémoire de master, Technische Universität Wien - Politecnico Di Milano - Institut für Nachrichtentechnik und Hochfrequenztechnik.

[Raake 2001] Raake, A. (2001). Modelling impairment due to packet loss for application in the E-model. Contribution UIT-T COM 12-D.44.

[Raake 2004a] Raake, A. (2004a). E-Model : additivity of burst packet loss impairment with other impairment types. Contribution UIT-T COM 12-D.221.

[Raake 2004b] Raake, A. (2004b). Predicting speech quality under random packet loss : individual impairment and additivity with other network impairments. *Acta Acustica united with Acustica*, 90:1061–1083.

[Richards 1973] Richards, D. L. (1973). *Telecommunication by Speech : The Transmission Performance of Telephone Networks*. Butterworths, Londres.

[Rix *et al.* 2001] Rix, A., Broom, S. et Reynolds, R. (2001). Non-intrusive monitoring of speech quality in voice over IP networks. Contribution UIT-T COM 12-D.49.

[Rix et Gray 2001] Rix, A. et Gray, P. (2001). NiQA - Non-intrusive speech Quality Assessment. Contribution UIT-T COM 12-D.48.

[Rix 2004] Rix, A. W. (2004). Perceptual speech quality assessment - a review. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, volume 3, pages 1056–1059.

[Rix et Hollier 2000] Rix, A. W. et Hollier, M. P. (2000). The perceptual analysis measurement system for robust end-to-end speech quality assessment. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, volume 3, pages 1515–1518.

[Rix *et al.* 2002] Rix, A. W., Hollier, M. P., Hekstra, A. P. et Beerends, J. G. (2002). Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment - Part I : time-delay compensation. *Journal of Audio Engineering Society*, 50(10):755–764.

[Rix *et al.* 1999] Rix, A. W., Reynolds, R. et Hollier, M. P. (1999). Perceptual measurement of end-to-end speech quality over audio and packet-based networks. In *106th AES Convention*. Papier n° 4873.

[Schroeder *et al.* 1979] Schroeder, M. R., Atal, B. S. et Hall, J. L. (1979). Optimizing digital speech coders by exploiting the masking properties of the human ear. *Journal of the Acoustical Society of America*, 66(6):1647–1652.

[Sporer 1997] Sporer, T. (1997). Objective audio signal evaluation - Applied psychoacoustics for modeling the perceived quality of digital audio. In *103rd AES Convention*. Papier n° 4512.

[Tebaldi 2005] Tebaldi, T. (2005). Influence of audio and video quality on subjective audiovisual quality - MPEG-4 and AAC coding. Mémoire de master, Technische Universität Wien - Politecnico Di Milano - Institut für Nachrichtentechnik und Hochfrequenztechnik.

[Thiede et Kabot 1996] Thiede, T. et Kabot, E. (1996). A new perceptual quality measure for bit rate reduced audio. In *100th AES Convention*. Papier n° 4280.

[UIT-R Rec. BS.1387 1998] UIT-R Rec. BS.1387 (1998). Méthode de mesure objective de la qualité du son perçu.

[UIT-T Rec. G.107 2005] UIT-T Rec. G.107 (2005). Le modèle E : modèle de calcul utilisé pour la planification de la transmission.

[UIT-T Rec. G.114 2003] UIT-T Rec. G.114 (2003). Temps de transmission dans un sens.

[UIT-T Rec. G.131 1996] UIT-T Rec. G.131 (1996). Réduction de l'écho pour le locuteur.

[UIT-T Rec. P.11 1993] UIT-T Rec. P.11 (1993). Effet des dégradations de la transmission.

[UIT-T Rec. P.56 1993] UIT-T Rec. P.56 (1993). Mesure objective du niveau vocal actif.

[UIT-T Rec. P.561 2002] UIT-T Rec. P.561 (2002). Dispositif de mesure en service et sans intrusion - mesures pour les services vocaux.

[UIT-T Rec. P.562 2004] UIT-T Rec. P.562 (2004). Analyse et interprétation des mesures en service sans intrusion dans les services vocaux.

[UIT-T Rec. P.563 2004] UIT-T Rec. P.563 (2004). Méthode mono-extrémité pour l'évaluation objective de la qualité vocale dans les applications de la téléphonie à bande étroite.

[UIT-T Rec. P.564 2006] UIT-T Rec. P.564 (2006). Tests de conformité des modèles d'évaluation de la qualité de transmission de la voix sur IP à bande étroite.

[UIT-T Rec. P.58 1996] UIT-T Rec. P.58 (1996). Simulateur de tête et de torse pour la téléphonométrie.

[UIT-T Rec. P.59 1993] UIT-T Rec. P.59 (1993). Voix conversationnelle artificielle.

[UIT-T Rec. P.800 1996] UIT-T Rec. P.800 (1996). Méthodes d'évaluation subjective de la qualité de transmission.

[UIT-T Rec. P.831 1998] UIT-T Rec. P.831 (1998). Évaluation subjective de la qualité de fonctionnement des annuleurs d'écho de réseau.

[UIT-T Rec. P.861 1998] UIT-T Rec. P.861 (1998). Mesure objective de la qualité des codecs vocaux fonctionnant en bande téléphonique (300-3400 Hz).

[UIT-T Rec. P.862 2001] UIT-T Rec. P.862 (2001). Évaluation de la qualité vocale perçue : méthode objective d'évaluation de la qualité vocale de bout en bout des codecs vocaux et des réseaux téléphoniques à bande étroite.

[UIT-T Rec. P.862.2 2005] UIT-T Rec. P.862.2 (2005). Extension bande élargie de la Recommandation P.862 pour l'évaluation des codeurs vocaux et des réseaux téléphoniques à bande élargie.

[UIT-T Rec. P.862.3 2005] UIT-T Rec. P.862.3 (2005). Guide applicatif concernant les mesures objectives de la qualité fondées sur les Recommandations P.862, P.862.1 et P.862.2.

[UIT-T Supp. 11 (Séries P) 1995] UIT-T Supp. 11 (Séries P) (1995). Some effects of sidetone.

[van Vugt 2005] van Vugt, J. (2005). Measuring a talking quality of a communication link in a network. Brevet international n° WO 2005/022876 A1.

[van Vugt et Beerends 2003] van Vugt, J. M. et Beerends, J. G. (2003). Updates to the Perceptual Echo and Sidetone Quality Measure (PESQM), an objective method for talking quality assessment. Contribution UIT-T COM 12-D.89.

[Yang 1999] Yang, W. (1999). *Enhanced Modified Bark Spectral Distortion (EMBSD) : an objective speech quality measure based on audible distortion and cognition model.* Thèse de doctorat, Temple University.

[Zoubir et Boashash 1998] Zoubir, A. M. et Boashash, B. (1998). The bootstrap and its application in signal processing. *IEEE Signal Processing Magazine*, 15(1):56–76.

[Zwicker et Feldtkeller 1981] Zwicker, E. et Feldtkeller, R. (1981). *Psychoacoustique : l'oreille récepteur d'information*. Collection technique et scientifique des télécommunications - CNET - ENST. Masson, Paris, France.

Résumé

La qualité vocale des systèmes de télécommunications est évaluée par les opérateurs qui se doivent de la maîtriser. Les méthodes subjectives permettent de connaître le jugement humain mais sont coûteuses : les méthodes objectives représentent une alternative. Un modèle objectif est proposé pour évaluer la qualité en contexte de conversation à partir des qualités d'écoute, de locution et d'interaction. Il est divisé en deux parties : la partie intégration combine les notes de qualité d'écoute, de locution et d'interaction pour estimer une note de qualité de conversation et la partie mesure fournit les notes objectives de qualité à la partie intégration en se basant sur les modèles existants de qualité vocale dans les différents contextes. Quatre tests subjectifs étudiant différentes dégradations de la qualité de conversation sont utilisés pour construire et valider la partie intégration du modèle. Les performances du modèle sont vérifiées en l'appliquant à des signaux réels.

Abstract

Speech quality is a huge issue for telecommunications operators : they have to evaluate it to satisfy their customers. Subjective methods assess human perception directly but they are expensive to conduct. Objective methods have been conceived to replace them. An objective model is proposed to assess conversational speech quality from the listening, talking and interaction qualities. This model comprises two parts : the integration function combines the listening, talking and interaction quality scores to estimate conversational quality score and the measurement function provides to the integration function the objective quality scores obtained from the existing models in the various contexts. Four subjective tests studying several degradations of the conversational speech quality have been designed to build and validate the integration function. The model is applied to real signals and achieves high performance.