Diagnostic et évaluation automatique de la qualité vocale à partir d'indicateurs hybride

Adrien Leman

N° d'ordre : 2011 ISAL 0053 Année 2011

Thèse

Présentée devant

L'Institut National des Sciences Appliquées de Lyon

Pour obtenir

Le grade de docteur

Ecole doctorale des Sciences pour l'Ingénieur de Lyon
Mécanique, Energétique, Génie Civil et Acoustique (MEGA)
Spécialité : Acoustique

Par

Adrien LEMAN

Diagnostic et évaluation automatique de la qualité vocale à partir d'indicateurs hybrides (Modèle DESQHI)

Soutenue le 7 juin 2011 devant la Commission d'examen

Membres du jury :

Christophe D'Alessandro	Directeur de Recherche CNRS (LIMSI Orsay)	Rapporteur
Régine Le Bouquin-Jeannès	Professeur (LTSI, Rennes)	Rapportrice
Alexander Raake	Docteur (Deutsche Telekom, Berlin)	Président
Etienne Parizet	Professeur (INSA Lyon)	Directeur
Julien Faure	Docteur (Orange Labs Lannion)	Encadrant

Ecoles doctorales 2010-2011

ECOLES DOCTORALES n° code national	RESPONSABLE PRINCIPAL
CHIMIE DE LYON (Chimie, Procédés, Environnement) EDA 206	M. Jean-Marc LANCELIN CPE LYON 04.72.43.13.95 lancelin@hikari.cpe.fr
HISTOIRE, GEOGRAPHIE, AMENAGEMENT, URBANISME, ARCHEOLOGIE, SCIENCE POLITIQUE, SOCIOLOGIE, ANTHROPOLOGIE (ScSo) EDA 483	M. Lionel OBADIA Lionel.obadia@univ-lyon2.fr LYON 2
ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE (E.E.A.) EDA 160	M. Alain NICOLAS ECL 04.72.18.60.96 Alain.Nicolas@ec-lyon.fr
EVOLUTION, ECOSYSTEMES, MICROBIOLOGIE , MODELISATION (E2M2) EDA 341	M. Jean-Pierre FLANDROIS 04.78.86.31.50 e2m2@univ-lyon1.fr
INTERDISCIPLINAIRE SCIENCES-SANTE (EDISS) EDA 205	M. Didier REVEL UCBL 1 04.72.35.72.32 didier.revel@creatis.univ-lyon1.fr
MATERIAUX DE LYON EDA 034	M. Jean Marc PELLETIER 83.18 Jean-Marc.Pelletier@insa-lyon.fr
INFORMATIQUE ET MATHEMATIQUES DE LYON (InfoMath) EDA 512	M. Alain MILLE UCBL 1 04.72.44.58.24 alain.mille@liris.cnrs.fr
MEGA DE LYON (MECANIQUE, ENERGETIQUE, GENIE CIVIL, ACOUSTIQUE) (MEGA) EDA 162	M. Philippe BOISSE 04.72.43.63.96 Philippe.Boisse@insa-lyon.fr

Remerciements

Je remercie vivement les rapporteurs Régine Le Bouquin-Jeannès et Christophe d'Alessandro, ainsi que le président du jury Alexander Raake d'avoir accepté de participer à mes travaux de recherche.

Je tiens à remercier Etienne Parizet qui m'a fait l'honneur d'être mon directeur de thèse, de m'avoir accueilli à L'INSA de Lyon et de m'avoir suivi de près et de loin entre Lyon et Lannion.

Un grand merci à Julien Faure de m'avoir encadré à Orange Labs de Lannion, de m'avoir soutenu dans toutes ces épreuves, pour sa présence quasi-permanente et pour tous ces bons moments.

Je remercie Pierre Henry et toute l'équipe MOV de Orange Labs pour les nombreux échanges, le soutien et l'aide dans les moments les plus durs. Je remercie particulièrement Vincent Barriac et Valérie Gautier-Turbin pour les nombreuses bases sonores mises à disposition. Je remercie aussi Noël et Christine pour leur disponibilité, et leur bonne humeur !
Je remercie Martine Apperry et Caroll Rattazzi pour leur aide précieuse lors de la réalisation des tests subjectifs et du recrutement des sujets. Merci à Laetitia Gros pour ses conseils et ses remarques avisées pour les analyses des tests subjectifs, ainsi qu'à Catherine Quinquis pour son aide.
Merci aussi au courage des nombreux sujets anonymes ou connus qui ont participés aux tests subjectifs.

Cette fabuleuse expérience m'a permis de redécouvrir la Bretagne avec ses paysages extraordinaires et sa culture musicale et festive.
Je tiens à remercier les « bretons » : les nombreux colocs de la rue Soisbault Charly, Nico, Soline, Stéphanie, Mathieu, Adrien, les voisins JB et Kristell, mais aussi les nombreux thésards et post-docs que j'ai pu rencontrés lors de cette aventure Nicolas, Emilie, Anna et Martin, Virginie, Aurélie, Vincent, Jeff, Camille, Amos, Julie et Guillaume, Lizzy, Meriem, Fatoumata, Artem, Guillaume, Aurélie, Adil, Yves...
Je pense aussi à tous mes amis avec qui j'ai partagé des moments inoubliables ! Merci à Damien, Romain, Philippe, Gaëlle, Olive et Fan pour leur gentillesse, leur accueil et leur humeur toujours joyeuse. Je remercie particulièrement Rachid de m'avoir fait découvrir Lannion, mais aussi mes compères Charly, Anaik, Jean-phi, et Olivier avec qui j'ai partagé ma passion pour la musique !!

Je remercie tous mes amis ch'ti qui ont toujours été là pour moi, en dépit de la distance : Pauline, Gauthier, Printz, Alex, Baptiste, Flo, Romain, Baron, Perrine, Gaby, Gwen, Nora, Steph, Fred et Del, Sylvain, Henry, Manu, Chrou, Clément, Jérôme et Lucie, Tomtom, Céline... Un grand merci à Antoine et à ma belle-sœur Elodie de s'être déplacés pour me soutenir dans les derniers moments !

Merci à ma famille : à mes 4 grands parents, à ma mère pour m'avoir légué sa passion pour la musique, à François pour sa gentillesse, à mon père pour m'avoir légué sa passion pour les sciences, et à Domino pour son aide très précieuse.

Un petit clin d'œil à Samuel, mon frère jumeau, qui ressent naturellement les fortes pensées que j'ai pour lui !
Je remercie au même titre ma petite sœur adorée Mathilde qui m'a sûrement le plus manqué durant ces années.

Enfin, un immense MERCI à Amandine, ma Chrousse, qui a su me supporter ces nombreuses années malgré la distance qui nous séparait, son caractère débordant d'énergie, de gaieté, et de festivité, et aussi de sa passion particulière pour les orques et les pommiers.

Je pense aussi à ces nombreux allers retours Lannion-Paris-Lille qui m'ont permis de conserver tous ces liens et d'être parvenu au bout de mes peines !

Sommaire

Liste des figures

Liste des tableaux

Abréviations et termes

ACP : Analyse en Composantes Principales
ACR : A*bsolute Category Rating* : Evaluation par catégories absolues
ADSL : *Asymmetric Digital Subscriber Line*
ANOVA : *ANalysis Of Variance* : Analyse de la variance
BER : B*it Error Rate* : taux d'erreur de bits
BDF : Bruit De Fond
C/I : C*arrier*-to-*interference ratio* : rapport signal sur interférence
CCR : C*omparison Category Rating :* Evaluation par catégories de comparaison
CCI : *Call Clarity Index* : indice de netteté des logatomes
CELP : *Code-Exited Linear Prediction*
CLP : *Conditional Loss Probability*
CMOS : C*omparison Mean Opinion Score* : Note moyenne d'opinion par comparaison
CODEC: COdeur-DECodeur
CPU : *Central Processing Unit* : Unité centrale de traitement
dB : Décibel
DAM : *Diagnostic Acceptability Measure*
DAV : Détection d'Activité Vocale
DCR : D*egradation Category Rating* : Evaluation par catégories de dégradations
DCT : *Discrete Cosinus Transform*
DECT : *Digital Enhanced Cordless Telecommunications*
DESQHI : *Diagnostic and Evaluation of Speech Quality using Hybrid Indicators*
DIAL : *Diagnostic Instrumental Assessment of Listening quality*
DMOS : D*egradation Mean Opinion Score :* Note moyenne d'opinion de la dégradation
DSP : *Digital Signal Processor*
EAM : Erreur Absolue Moyenne
ERB : *Equivalent Rectangular Bandwidth*
EVRC : *Enhanced Variable Rate Codec*
FER : *Frame Erasure Rate* : taux d'effacement de trame
FFT : *Fast Fourier Transform*
GSM : *Global System for Mobile communication*
GSM EFR : Codec mobile de type *Enhanced Full-Rate*
GSM FR : Codec mobile de type *Full-Rate*
IA : Indice d'Articulation
IP : *Internet Protocol*
IRS : *Intermediate Reference System* : Système de référence intermédiaire
ISO : *International Organisation for Standardization*
ITU : *International Telecommunication Union*
ITU-T : ITU- *Telecommunication standardization sector*
LPC : Codage par Prédiction Linéaire
MNRU : *Modulated Noise Reference Unit* : appareil de référence à bruit modulé
MOS : M*ean Opinion Score* : note moyenne d'opinion
MOS-LQS : *MOS - Listening Quality Subjective*
MOS-LQO : *MOS - Listening Quality Objective*
MUSHRA : *MUlti Stimulus test with Hidden Reference and Anchor*
NB : *Narrow-Band*
POLQA : *Objective Listening Quality Assessment*

PCA : *Principal Component Analysis*
PCM : *Pulse Code Modulation*
PESQ: *Perceptual Evaluation of Speech Quality*
PPL : the *Percentage of Packet Loss*
PLC : *Packet Loss Concealment*
RNIS : Réseau Numérique à Intégration de Services
RSB : Rapport signal sur bruit
RTC : Réseau Téléphonique Commuté
RTP : *Real-time Transport Protocol*
SD : *Semantic Differential*
SNR : S*ignal-to-Noise Ratio:* rapport signal/bruit
SVD : *Singular Value Decomposition :* Décomposition en valeurs singulière
UDP : *User Datagram Protocol :* Protocole de datagramme utilisateur
UIT : Union Internationale des Télécommunications
VoIP : *Voice over Internet Protocol*
VAD : *Voice Activity Detection*
WB : *WideBand*

Introduction

Pour communiquer, les humains utilisent souvent la voix sous la forme de parole intelligible, mais elle peut aussi être constituée d'onomatopées, de cris, de pleurs ou encore de chant. Le signal de la voix se transmet d'un individu à l'autre à travers un milieu matériel grâce à la mise en vibration des particules constituant le milieu considéré. Cette technique de communication est efficace, mais elle est rapidement limitée par la distance séparant les deux interlocuteurs. C'est seulement au 19ème siècle, avec l'invention du téléphone par Graham Bell en 1876, que le signal vocal a pu être transformé par un microphone sous la forme d'un signal électrique, afin de rendre possible la transmission de la voix sur de grandes distances. Depuis ces années, les techniques de télécommunication ne cesseront d'évoluer avec l'apparition de la transmission des signaux numériques et de la transmission mobile utilisant les ondes électromagnétiques comme porteuses du signal vocal, et plus récemment la transmission via le réseau internet (cf. §.I.1.2). Cette dernière est appelée la VoIP (Voice over Internet Protocol) et se base sur la segmentation préalable du signal en paquets qui sont ensuite envoyés indépendamment sur les réseaux IP. Il se peut que certains paquets arrivent en retard ou qu'ils se soient perdus en chemin, entraînant des coupures dans le signal vocal. Toutes les techniques de télécommunication (RTC, RNIS, GSM, VoIP) génèrent des dégradations sur le signal vocal qui leur sont propres (p. ex. coupures, bruit de fond, distorsion, grésillement, écho, délai…).
Il y a une dizaine d'années, la principale attente des utilisateurs était de communiquer avec leurs réseaux sociaux de manière simple et économique. La qualité des transmissions de la voix n'était pas un facteur primordial pour les clients, seule la bonne compréhension du message était requise. Avec la progression des technologies de télécommunication (p. ex. transmission en bande élargie (50 Hz – 7 kHz)), les transmissions téléphoniques sont presque toujours intelligibles. Les nouvelles technologies sont maintenant orientées vers l'amélioration de la qualité des services afin de les rendre plus fonctionnels et plus naturels.

La qualité vocale doit être contrôlée afin de répondre à deux principaux besoins. Le premier concerne la planification des nouvelles techniques de télécommunication en cours de développement, comme par exemple, le codage de la parole ou les algorithmes de débruitage de la parole. Le deuxième est axé sur le contrôle des télécommunications existantes en évaluant la qualité vocale des services proposés.
La manière la plus fiable et la plus performante d'évaluer la qualité vocale des transmissions téléphoniques est de demander directement l'avis aux utilisateurs. Cependant, cette méthode est très coûteuse en temps de réalisation et en nombre de personnes à interroger. Des instruments de mesure ont été développés afin d'estimer automatiquement la qualité vocale perçue par un grand nombre d'usagers. Les instruments les plus performants sont normalisés à l'UIT (Union Internationale des Télécommunications) comme les modèles PESQ [1], P.563 [2], G.107 [3]. La progression rapide des systèmes de télécommunication implique d'améliorer et de mettre à jour ces modèles existants, et de fournir de nouvelles informations sur le contrôle de la télécommunication grâce au diagnostic des dégradations perceptives.

Le but de cette thèse consiste à proposer un nouvel instrument d'évaluation de la qualité vocale, en s'intéressant principalement à deux caractéristiques.
La première concerne **le type d'indicateur** utilisé pour prédire la qualité vocale, à savoir l'utilisation des indicateurs basés sur le signal ou des indicateurs paramétriques issus des statistiques du réseau. Notre hypothèse est que l'utilisation des deux types d'indicateurs en combinaison permet d'améliorer les performances globales de la prédiction de la qualité vocale, comparativement à l'emploi de l'un ou l'autre.
La deuxième caractéristique concerne la construction du cœur du modèle à partir **d'une approche multidimensionnelle** qui permet de représenter l'évaluation de la qualité vocale (cf. §.I.1.4). Cette approche fournira potentiellement de meilleures performances lors de la prédiction de la qualité vocale comparativement aux modèles existants. L'évaluation de la qualité vocale n'est parfois pas suffisante pour le contrôle des services proposés. Cette approche multidimensionnelle permet de proposer, en plus de la note globale d'évaluation de la qualité vocale, un diagnostic de la télécommunication grâce à l'identification des principaux défauts perceptifs présents sur le signal vocal. Ce diagnostic permettra de cibler les points techniques à améliorer dans le but de proposer des services plus fidèles aux utilisateurs.
Cet instrument de mesure est appelé le modèle DESQHI ("Diagnostic and Evaluation of Speech Quality using Hybrid Indicators").

Tout d'abord, l'état de l'art décrit les principaux moyens d'évaluation de la qualité vocale (tests subjectifs et instruments de mesure) et résume les principaux attributs perceptifs influençant l'évaluation de la qualité vocale (Chapitre I). Cela permet ensuite de définir et de justifier le choix des principales caractéristiques du modèle DESQHI en fonction des besoins actuels (Chapitre II).

Le Chapitre III est consacré à la construction du cœur du modèle multidimensionnel. Plusieurs tests subjectifs sont réalisés à partir d'une même base sonore afin de déterminer les dissimilarités de toutes les paires de stimuli et d'évaluer la qualité vocale de chacun des stimuli. L'analyse des dissimilarités par la méthode d'échelonnement multidimensionnel permet de déterminer un espace perceptif constitué de trois dimensions (*bruyance*, *codage de la parole*, *continuité*). Le cœur du modèle est constitué d'une combinaison linéaire de ces trois dimensions perceptives permettant de prédire la qualité vocale.

Les deux chapitres suivants (Chapitre IV et Chapitre V) consistent à modéliser chacune des trois dimensions perceptives par différents types d'indicateurs (paramétrique, hybride ou basé sur le signal). Cela a l'avantage de proposer un modèle adaptatif en permettant à l'expérimentateur de choisir le type d'indicateur pour chacune des trois dimensions selon les informations disponibles au point de la mesure ou encore selon le domaine d'application (contrôle différé ou temps réel, planification).
Dans le cas de la modélisation de la bruyance (Chapitre IV), des tests subjectifs sont réalisés afin de démontrer qu'il existe en plus de l'influence du niveau sonore du bruit de fond, une influence du type de bruit de fond lors de l'évaluation de la qualité vocale. Un algorithme de classification automatique du type de bruit de fond est construit et intégré à la modélisation de la bruyance, afin d'améliorer les performances de la prédiction de cette dimension. Cette classification permet aussi de proposer un diagnostic avancé permettant par exemple de prévoir si le bruit provient du réseau ou bien de l'environnement du locuteur.

Enfin, le dernier chapitre (Chapitre VI) expose les performances globales du modèle DESQHI pour différentes bases sonores. Elles sont aussi comparées aux performances obtenues par les modèles existants.

Chapitre I. Etat de l'art de l'évaluation multidimensionnelle de la qualité vocale

I.1. De la production vocale vers l'évaluation de la qualité vocale

Lors d'une télécommunication, l'évaluation de la qualité vocale fait intervenir de nombreuses contributions mécaniques, électroniques et physiologiques. Pour appréhender l'ensemble des éléments qui composent la qualité vocale, il faut considérer la production de la parole d'un locuteur, la transmission du signal de la parole entre la bouche du locuteur et le microphone du terminal émetteur, puis la transmission du signal sur le réseau téléphonique. Ce signal est diffusé à l'auditeur par le terminal récepteur, faisant intervenir l'organe auditif humain. La parole est alors traitée par notre cerveau qui traduit ces informations en jugement de qualité vocale.

I.1.1. Production de la parole

Le niveau d'intensité acoustique moyenne produit par un son de parole, mesurée à 1 mètre, est compris entre 30 et 110 dB SPL, ce qui correspond respectivement à une voix chuchotée et à une voix criée. Dans une situation de conversation normale, elle se situe entre 60 et 80 dB SPL. L**e spectre fréquentiel** de la voix humaine diffère légèrement selon le sexe et l'âge des individus. Pour une voix naturelle parlée, les fréquences fondamentales sont, pour les hommes, les femmes et les enfants respectivement comprises entre 125 – 150 Hz, 220 - 300 Hz et 300 - 350 Hz. La bande fréquentielle contenant les informations utiles à la bonne compréhension de la parole humaine est comprise entre 100 Hz et 4 kHz. On remarque que le niveau d'intensité acoustique est maximal lorsque la fréquence se situe aux alentours de 500 Hz.

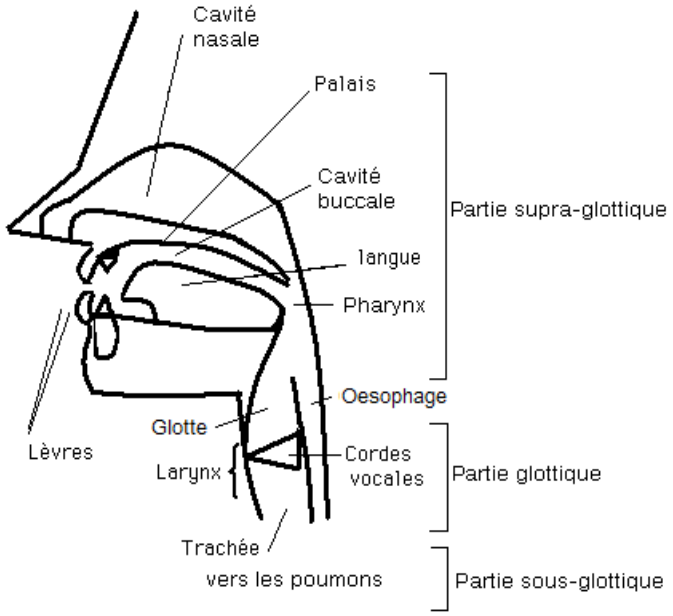


Fig. I.1 Appareil phonatoire humain

La parole est produite à partir de l'appareil phonatoire humain qui peut être décomposé en trois parties : la partie sous-glottique, la partie glottique et la partie supra-glottique (cf. Fig. I.1).

La partie sous-glottique est constituée de la soufflerie qui est gérée par les poumons et génère un flux d'air qui se propage dans la trachée, jusqu'à atteindre les cordes vocales.

La partie glottique est constituée des cordes vocales qui sont situées dans le larynx, au niveau de la pomme d'Adam. Elles sont composées de deux membranes qui se mettent à vibrer plus ou moins vite lors du passage de l'air en générant un son de fréquence fondamentale f_0.

La partie supra-glottique contient les différentes cavités telles que la cavité buccale, la cavité nasale ainsi que le pharynx. Elles sont gérées par les muscles de la bouche (langue, lèvres) afin de faire varier leurs dimensions pour produire différents types de sons.

Un locuteur adapte l'intensité et le timbre de sa voix de manière inconsciente en fonction du bruit de fond environnant, de façon à se faire mieux comprendre en environnement bruyant. C'est ce que l'on appelle l'effet Lombard. En général, l'effet de forçage de la voix entraîne une augmentation du débit de l'air expiré (augmentation de l'intensité), une augmentation de la note fondamentale, une modification du timbre de la voix, une gestuelle plus prononcée ainsi qu'une hyper-prononciation labiale (Garnier [4]). Ces paramètres évolueraient plus ou moins rapidement selon les différentes personnes en fonction de l'aptitude à mettre en avant leur voix.

I.1.2. Transmission de la parole sur les réseaux téléphoniques

Lors d'une télécommunication, le signal de la parole prononcée par un locuteur est soumis à un grand nombre de traitements qui peut perturber la qualité de la transmission téléphonique. Le signal est d'abord capté par un microphone, puis envoyé à travers le réseau de télécommunication où il est codé et décodé un certain nombre de fois, puis il est restitué par un haut-parleur à l'oreille de l'auditeur.

Le terminal joue un rôle sur la qualité de la transmission de la parole, suivant les caractéristiques du microphone et du haut-parleur utilisés, les algorithmes de codage et de décodage ou encore les algorithmes de réduction de bruit de la parole inclus dans le terminal. L'utilisation d'un kit main libre ou de la fonction "haut-parleur" influence aussi la qualité de la transmission par la variabilité de la position du microphone. Cela a pour effet de capter les diverses réflexions de la parole causées par la salle (effet de réverbération) et de recevoir les bruits environnants à un niveau sonore plus élevé, comparé à l'utilisation d'un combiné classique (Möller [5]).

La transmission du signal de la parole par les systèmes téléphoniques génère aussi toutes sortes de dégradations du signal, suivant le type de réseau utilisé dont les principales familles sont présentées dans les parties suivantes.

Les techniques de réseaux liées à la télécommunication ne cessent d'évoluer depuis l'apparition de la téléphonie classique au début du XXème siècle. La première technique utilisée est la téléphonie classique par réseau téléphonique commuté (RTC). Ensuite, de nouvelles techniques ont été développées afin de compléter et d'améliorer la transmission de la parole sur de longues distances : le Réseau Numérique à Intégration de Services (RNIS), le réseau mobile ou Global System for Mobile communication (GSM) et plus récemment la voix sur IP (VoIP).

I.1.2.1. La téléphonie classique (Réseau Téléphonique Commuté)

La première télécommunication vocale a été réalisée en transformant la parole en signal électrique afin de transmettre l'information à travers des fils de cuivre reliant les différents interlocuteurs. La bande de fréquences utilisée est alors comprise entre 300 et 3,4 kHz, représentant les fréquences les plus importantes à la bonne compréhension des informations de la

parole. Ce type de transmission était limité par la distance à cause de l'apparition du bruit de fond induit par l'utilisation de cette technique. La téléphonie classique est toujours très répandue de nos jours. Les réseaux téléphoniques commutés actuels sont utilisés afin de transmettre le signal analogique enregistré au niveau du terminal d'un locuteur jusqu'à la plus proche centrale téléphonique. Le signal analogique est alors transformé en un signal numérique grâce à un codeur (p. ex. G.711) avant d'être envoyé via un réseau numérique au commutateur situé au plus proche du récepteur. Le décodeur est alors employé afin de transmettre au terminal récepteur le signal analogique reconstitué. Les principales dégradations générées par ce type de réseau sont le bruit de fond, l'écho et les dégradations liées au codage de la parole.

I.1.2.2. Réseau Numérique à Intégration de Services (RNIS)

Le Réseau Numérique à Intégration de Services est normalisé à l'Union Internationale des Télécommunications. Il est utilisé afin de transmettre le signal de parole de bout en bout de manière numérique, afin de s'affranchir de la transmission du signal analogique qui génère du bruit de fond. Le signal de la parole est directement transformé d'un signal analogique à un signal numérique par un codeur au niveau du terminal du locuteur puis par le traitement inverse au niveau du terminal de l'auditeur. Le RNIS permet de transmettre de l'information vocale, mais aussi d'envoyer tous types de données numérisées à un débit de 64 kbits/s. Les principales dégradations engendrées par ce type de réseau sont liées au codage à bas débit de la parole (G.711) qui limite la bande passante audio à 4 kHz.

I.1.2.3. Réseau mobile ou Global System for Mobile communication (GSM)

C'est dans les années 90 que la technique de transmission mobile (GSM) s'est largement répandue. Le transport de la voix se fait au moyen des ondes électromagnétiques entre le terminal mobile et les antennes-relais qui effectuent la transformation du signal électromagnétique en signal électrique. Ensuite, le transport s'effectue par le réseau classique existant (RTC). Le téléphone mobile permet de communiquer dans n'importe quel endroit couvert par les antennes-relais. Les interlocuteurs sont parfois amenés à téléphoner dans des endroits bruyants. Ces bruits de fond d'environnement sont captés par le microphone du terminal au même titre que le signal de la parole, ce qui va produire une dégradation de la qualité vocale. La transmission utilisant les ondes électromagnétiques peut aussi engendrer des pertes et des coupures du signal vocal et présenter des problèmes d'interférences.

I.1.2.4. Voix sur IP (VoIP)

Le protocole internet (IP) permet d'échanger différents contenus tels que des sons, des images, des vidéos et de manière générale toutes données pouvant être numérisées. Le transport de la parole sur les réseaux IP est apparu en 1995 et se répand peu à peu auprès de la plupart des utilisateurs avec l'apparition de l'offre "Triple Play" regroupant les services Internet, Téléphone et Télévision, et la gratuité du service téléphonique. Aujourd'hui, la place de la VoIP dans le marché a tendance à dépasser la téléphonie classique (RTC).

La voix est numérisée, compressée puis découpée en paquets au niveau du terminal émetteur et subit le traitement inverse au niveau de l'auditeur afin de recomposer la voix du locuteur. Le transport de l'information est basé sur l'acheminement des paquets sur le réseau internet. Chaque paquet est envoyé sur un réseau différent. Ce sont les routeurs qui assurent l'acheminement de chaque paquet en empruntant le chemin le plus court. Il existe parfois des imperfections : il se peut que les paquets arrivent en retard, en avance ou de manière désor-

donnée, ou alors qu'ils soient tout simplement "perdus". Des buffers de gigues[1] sont positionnés au niveau du récepteur afin d'absorber les fluctuations de temps d'arrivée des paquets (la gigue) ou de remettre les paquets affluant dans le bon ordre. La taille des buffers étant limitée, une gigue excessive peut entraîner un débordement du buffer ce qui se traduit par des paquets perdus. Les dégradations susceptibles d'être présentes sur ce type de communication sont les pertes de paquets, le délai de bout en bout excessif, les dégradations dues aux codages/décodages successifs de la parole ainsi que les dégradations issues de la transmission RTC (écho, bruit de fond).

I.1.3. Perception de la parole

L'unité principale constituant le langage est le phonème[2]. Les informations fondamentales du langage parlé sont représentées par les syllabes composées de plusieurs phonèmes. Ces syllabes ont une durée comprise entre 100 et 300 ms (Greenberg [6]). Les fréquences inférieures à 50 Hz et supérieures à 8 kHz n'apportent pas plus d'information à la perception de la parole.

Le codage de la parole basé sur des modèles prédictifs est réalisé en découpant le signal suivant des trames de durée allant de 10 ms jusqu' à 40 ms suivant le type de codec. Les recherches en traitement et synthèse de la parole ont montré que le signal vocal est analysé par le systéme auditif comme un enchaînement de syllabes, et non comme l'analyse indépendante des syllabes. L'analyse des signaux de parole s'appuie sur ce principe en considérant un recouvrement successif des trames de 50 %.

La perception de la parole représente les capacités auditives et cognitives d'un individu à traduire un signal de parole en une information. D'après Jekosch [7], la compréhension de l'information est définie comme le résultat final de ce processus perceptif. Raake [8] définit le processus de perception de la parole comme une succession de quatre étapes correspondant à la compréhensibilité, l'intelligibilité, la communicabilité et la compréhension.

La compréhensibilité de la parole représente les capacités auditives d'un individu à identifier des sonorités composées de paroles dépourvues de signification, telles que les phonèmes ou les syllabes. Möller [5] et French et Steinberg [9] utilisent le mot "articulation" pour décrire la compréhensibilité de la parole. Un niveau de compréhensibilité élevé correspond à une reconnaissance parfaite des phonèmes constituant le signal de parole.

L'intelligibilité de la parole représente les capacités auditives et cognitives d'un individu à identifier la signification d'un mot ou d'un groupe de mots. L'intelligibilité fait intervenir des effets sémantiques, lexicaux, de syntaxe ainsi que de la prosodie de la phrase. Le niveau lexical utilisé peut aussi altérer l'intelligibilité de la parole lorsque les auditeurs questionnés ne connaissent pas certains mots. La compréhension de la parole est fortement liée à la notion d'intelligibilité.

La communicabilité de la parole nécessite que le niveau d'intelligibilité de la parole soit élevé, c'est-à-dire que l'information vocale soit parfaitement comprise par le récepteur. La communicabilité dépend de l'aspect fonctionnel de la communication en prenant en compte

[1] Les buffers de gigues sont des mémoires tampon utilisées pour contrôler la régularité du flux des paquets transmis. Cela présente l'inconvénient de rajouter un délai supplémentaire à l'ensemble de la transmission.

2 Les phonèmes sont les éléments sonores distinctifs du langage articulé. Ils représentent les plus petites unités sonores produites par l'appareil phonatoire humain.

tout le processus engendré par une communication, comme le contexte de situation des interlocuteurs, la sémantique ou encore le sujet de conversation.

La compréhension est le résultat du processus de perception de la parole. Elle représente les capacités d'un auditeur à identifier l'information transmise lorsque cet auditeur est prêt à comprendre la parole.

Ces différentes étapes correspondent grossièrement aux unités linguistiques du processus de perception de la parole, à savoir l'unité du phonème pour la compréhensibilité, l'unité du mot ou groupe de mots pour l'intelligibilité et l'unité prosodique pour la communicabilité. La bonne compréhension de la parole nécessite un certain niveau acceptable de ces trois différents mécanismes. Cependant, dans certains cas, il est possible que la compréhensibilité de la parole soit faible, mais que l'intelligibilité ou la communicabilité soit bonne grâce aux capacités auditives de l'Homme à combler les zones incomprises par rapport au contexte de la phrase, ou encore par la lecture labiale.

Lorsque la compréhension de la parole est acceptable, l'auditeur peut porter son jugement sur l'évaluation de la qualité vocale. L'auditeur juge alors d'autres aspects tels que les dégradations qui altèrent le naturel de la voix d'origine, l'effort fourni lors de la perception de la parole, ou encore le niveau lexical employé.

I.1.4. Evaluation de la qualité vocale

Selon Jekosch [7], la définition globale de la qualité est la suivante :

"*the result of [the] judgement of the perceived composition of an entity with respect to its desired composition*". P.15

Le terme "*the perceived composition"* est défini comme :
"*the totality of features of an entity*", P.16

avec "*feature*" décrit comme
"*[a] recognizable and nameable characteristic of an entity.*" P.14

et *"the desired composition"* est défini comme
"the totality of features of individual expectations and/or relevant demands and/or social requirements."

Cette définition peut être illustrée par le shéma proposé par Jekosch [7], et repris par Raake [8] et Côte [10] :

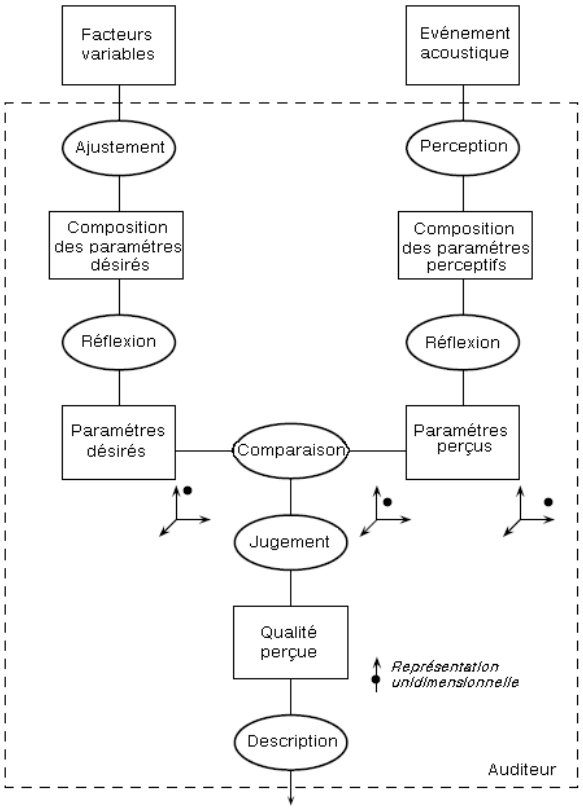


Fig. I.2 Description du phénomène de jugement de la qualité par un auditeur, basée sur une étude de Jekosch [6]. Les cercles correspondent aux procédures et les rectangles correspondent aux transformations faites par l'auditeur.

D'après cette description, le fonctionnement de l'oreille serait incapable de fournir une information précise sur les caractéristiques absolues d'un stimulus, comme par exemple l'intensité sonore. Par contre l'oreille aurait une excellente sensibilité différentielle.
L'hypothèse posée dans cette partie est qu'il en est de même pour l'évaluation de la qualité vocale : l'évaluation de la qualité vocale fait appel à la **comparaison** (cf. Fig. I.2) entre un signal vocal dégradé et un signal de référence interne.

Dans un contexte d'écoute, l'évaluation de la qualité vocale est alors définie comme le résultat du jugement de la comparaison entre **la composition des paramètres perceptifs** du signal vocal et **la composition des paramètres désirés** (cf. Fig. I.2).

- **La composition des paramètres perceptifs** du signal vocal correspond aux attributs perceptifs reconnus et identifiables des caractéristiques du signal de la parole. Ces attributs sont qualifiés par des adjectifs tels que par exemple clair/sourd, naturel/synthétique, doux, sensible, mélodieux, agressif…
- **La composition des paramètres désirés** correspond aux attributs perceptifs désirés du signal vocal non dégradé. Celui-ci est souvent appelé le signal de référence interne et correspond à la représentation de ce que l'auditeur s'attend à percevoir par rapport à ses connaissances. Cet effet est représenté dans la Fig. I.2 par les **facteurs variables** décrivant l'expérience personnelle, le contexte dans lequel la voix est diffusée, les facteurs individuels comme la motivation ou l'humeur de l'auditeur, la connaissance préalable ou non du locuteur, l'indulgence de l'auditeur face à la qualité vocale désirée, ou encore la consigne donnée aux sujets lors de l'évaluation de la qualité vocale.

Il faut retenir que l'évaluation de la qualité vocale est toujours relative ou bien à un signal de référence interne qui peut être différent suivant les sujets interrogés, ou bien à un

signal de référence imposé par l'expérimentateur qui est alors commun pour tous les sujets interrogés.

L'étape de la comparaison utilise une représentation multidimensionnelle des paramètres perçus et désirés, afin de comparer chacun des paramètres de manière indépendante (cf. Fig. I.2).

L'évaluation de la qualité vocale s'effectue ensuite en combinant les résultats issus de la comparaison de chacun des paramètres afin d'obtenir une représentation **unidimensionnelle** de la qualité vocale pouvant facilement être exprimée sur une échelle unique. Toutes ces différentes étapes de jugement de la qualité vocale sont réalisées de manière inconsciente par notre cerveau.

La mesure de la qualité vocale est généralement exprimée à l'aide de la note MOS (Mean Opinion Score en anglais ou note moyenne d'opinion) défini dans l'UIT-T P.800 [11]. Cette note peut exprimer plusieurs types de mesures selon trois critères représentés par les trois paramètres (X, Y, Z). On précise alors le type de note en la renseignant par les trois critères : **MOS – *XQYZ.*** *Q* signifie "Quality".

- *X* représente le contexte lors de la mesure de la qualité vocale avec *X* = *L* pour "Listening" (contexte d'écoute), *X* = *S* pour "Speech" (contexte de locution) ou bien *X* = *C* pour "Conversation" (contexte de conversation).
- *Y* représente la nature de la mesure avec *Y* = *O* pour "Objective" (par mesure instrumentale), *Y* = *S* pour "Subjective" (mesures issues d'un test subjectif comme le test ACR) ou bien *Y* = *E* pour "Estimated" (mesures estimées comme par exemple le modèle E).
- *Z* représente la largeur de bande spectrale utilisée par les échantillons sonores, avec *Z* = *N* pour "Narrowband" (bande étroite 300 Hz → 3,4 kHz), *Z* = *W* pour "Wideband" (bande élargie 50 Hz → 7 kHz) ou encore *Z* = *M* pour "Mixte" (Les deux bandes de fréquences étroite et large sont présentes).

L'évaluation de la qualité vocale est très variable d'un individu à un autre. Cela peut être causé par de nombreux facteurs comme l'âge du sujet, le contexte d'écoute, la voix du locuteur, la connaissance préalable du locuteur ou encore le vocabulaire employé. La note MOS représente la moyenne des jugements subjectifs pour un grand nombre de sujets. Cette note s'approche alors du jugement objectif de la qualité vocale.

Les tests subjectifs sont très coûteux et très longs à réaliser à cause du nombre important de sujets à interroger. C'est pourquoi il est nécessaire d'avoir des outils de mesure automatique qui soient aussi précis et aussi représentatifs que possible de l'ensemble des utilisateurs.

Les résultats des tests subjectifs moyennés suivant les sujets (MOS-LQS) permettent de construire **des instruments automatiques** d'évaluation de la qualité vocale appelés par abus de langage **"modèles objectifs d'évaluation de la qualité vocale".** Les notes de qualité vocale obtenues par les modèles d'évaluation de la qualité vocale en contexte d'écoute sont appelées MOS-LQO. Les principaux modèles sont décrits dans le paragraphe I.5.

I.2. Evaluation subjective unidimensionnelle de la qualité vocale

Les tests subjectifs représentent le meilleur moyen d'évaluer la qualité vocale en demandant directement l'opinion des sujets sur les échantillons vocaux testés.

La recommandation P.800 de l'UIT-T [11] intitulée "Méthodes d'évaluation subjective de la qualité de transmission" décrit trois tests subjectifs d'évaluation de la qualité vocale : le test ACR, le test DCR et le test CCR. Le test MUSHRA est aussi beaucoup utilisé pour évaluer la qualité d'un système audio. Ces quatre méthodes utilisent une représentation unidimen-

sionnelle de la qualité du signal vocal en présentant une échelle de préférence représentée par un axe linéaire.

I.2.1. Test ACR (Absolute Category Rating)

Le test subjectif le plus courant et le plus utilisé est le test appelé "Absolute Category Rating" (ACR) UIT-T P.800 [11]**.**

Les sujets expriment leurs jugements sur la qualité du signal vocal à travers une échelle catégorielle à cinq niveaux présentée dans le Tab. I.1. Pour forcer l'utilisation de toute l'échelle, il est demandé aux sujets d'utiliser au maximum les cinq catégories par rapport à l'ensemble de la base sonore. Cette consigne a pour but d'obtenir des résultats présentant le plus de variations significatives possibles entre les conditions, au détriment dans certains cas de la spontanéité et du réalisme des jugements de préférence.

La mesure de la qualité du signal vocal se fait dans l'absolu, sans le signal de référence. L'auditeur se crée sa propre référence interne de qualité vocale qui est de ce fait très variable suivant les sujets interrogés (cf. §.I.1.4). C'est pourquoi un apprentissage doit être réalisé préalablement, afin de présenter aux sujets les variations maximales des dégradations des sons proposés, afin que les sujets jugent de la qualité vocale par rapport à l'ensemble des stimuli proposés. La base sonore devra contenir des voix de sexes différents, afin de s'affranchir des effets du timbre de la voix. De plus, il est conseillé que chaque échantillon sonore dure entre 6 et 10 secondes et contienne deux ou trois phrases séparées entre elles par un intervalle d'au moins une seconde, pour permettre d'entendre le bruit de fond de la liaison. Les phrases sont égalisées à un niveau de -26 dBov[3] avant d'être restituées aux sujets en écoute monaurale à 79 dB(SPL). Le bruit d'ambiance à l'écoute doit être inférieur à 40 dB(A).

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

Tab. I.1 Echelle catégorielle de mesure pour le test ACR (UIT-T P.800 [11])

L'échelle proposée comprend cinq catégories principalement pour simplifier la tâche demandée aux sujets et parce que ces catégories sont estimées suffisantes à l'évaluation de la qualité vocale. La moyenne des résultats de tous les sujets permet d'obtenir la note de qualité représentative de l'ensemble des individus d'un groupe, avec une précision variable suivant le nombre de sujets interrogés et la difficulté du test proposé. Cette précision peut être estimée à

3 Le dBov (Over Load) représente le niveau en décibel par rapport au point de surcharge d'un système numérique. Par exemple, un échantillon sonore quantifié sur 16 bits permet de préciser son intensité grâce à 2^{16} possibilités, soit 2^{15} possibilités dans la partie positive et 2^{15} possibilités dans la partie négative du signal. Nous obtenons alors un niveau sonore négatif "N", exprimé en dBov par rapport au niveau de saturation (2^{15}), à partir du niveau numérique "Nb" représenté par un entier naturel compris entre -2^{15} et 2^{15} :

$$N(dBov) = 20.\log(|Nb| / 2^{15}).$$

travers l'intervalle de confiance entre les différents sujets. La recommandation préconise de faire participer au moins 12 sujets, et si possible 32 sujets, pour obtenir des résultats fiables.

Ce type de test est assez fatigant et peut provoquer une baisse de concentration de la part des sujets. C'est pourquoi la durée du test ne doit pas dépasser 1h30.

Le test ACR a l'avantage d'être assez rapide et permet donc de couvrir une large variété de dégradations. En effet, chaque échantillon sonore est jugé dans l'absolu sans avoir besoin de comparer avec le signal de référence. La durée d'écoute des stimuli est donc deux fois plus courte par rapport à un test utilisant la comparaison entre le signal dégradé et le signal de référence, comme pour les deux autres méthodes proposées ci-dessous.

I.2.2. Test DCR (Degradation Category Rating)

La méthode d'évaluation par catégories de dégradation (test DCR (UIT-T P.800 [11])) permet d'avoir une meilleure précision par rapport au test ACR, lors de l'évaluation de communication présentant de faibles dégradations. Le jugement se fait par comparaison entre le signal dégradé et le signal de référence non-dégradé. De ce fait, la tâche demandée aux sujets est plus simple que dans l'absolu (test ACR) et les dégradations fines de la parole sont plus facilement détectables. Par exemple, ce test est utilisé pour faire la différence entre certains codecs.

Il est demandé aux sujets d'évaluer le niveau de dégradation par rapport à un certain attribut perceptif, sur une échelle de catégories à cinq niveaux présentée sur le Tab. I.2. Cette consigne a l'avantage de pouvoir orienter le test suivant différents attributs comme par exemple le naturel de la voix, le bruit de fond ou encore l'évaluation de la qualité vocale.

Note	Niveau de dégradation
5	Dégradation inaudible
4	Dégradation audible mais pas gênante
3	Dégradation un peu gênante
2	Dégradation gênante
1	Dégradation très gênante

Tab. I.2 Echelle catégorielle de mesure pour le test DCR (UIT-T P.800 [11])

En moyennant les résultats de tous les sujets, nous obtenons la note d'appréciation moyenne de la dégradation représentée par la note DMOS (Degradation Mean Opinion Score).

I.2.3. Test CCR (Comparison Category Rating)

Le principe du test d'évaluation par catégories de comparaison (CCR (UIT-T P.800 [11])) est semblable au test DCR avec une différence par rapport à l'ordre de présentation des paires des échantillons sonores. Contrairement au test DCR, l'échantillon de référence et l'échantillon dégradé sont présentés dans un ordre aléatoire. L'échelle de mesure est adaptée à cette présentation aléatoire. Elle utilise une échelle catégorielle à sept niveaux, afin d'évaluer la qualité du deuxième son par rapport au premier :

Note	Qualité du 2ème son / au 1er son
3	Bien meilleure
2	Meilleure
1	Légèrement meilleure
0	A peu près équivalente
-1	Un peu moins bonne
-2	Moins bonne
-3	Beaucoup moins bonne

Tab. I.3 Echelle catégorielle de mesure pour le test CCR (UIT-T P.800 [11])

Grâce à cette méthode, les auditeurs formulent deux jugements avec une même réponse :

- *"Quel est l'échantillon de meilleure qualité ?"*
- *"Quelle est la différence de qualité entre les deux échantillons ?"*

Les méthodes DCR et CCR sont efficaces pour prédire des différences de qualité vocale entre les stimuli. Elles sont plus précises que lors d'une évaluation dans l'absolue (test ACR). L'avantage de la méthode CCR par rapport à la procédure DCR est qu'elle permet d'évaluer le traitement de la parole qui dégrade ou améliore la qualité de la parole.

En moyennant les résultats de tous les sujets, nous obtenons la note moyenne d'opinion par comparaison, représentée par le score CMOS (Comparison Mean Opinion Score).

I.2.4. Test MUSHRA

Le test MUSHRA "*MUlti Stimulus test with Hidden Reference and Anchor*" UIT-R BS.1534 [12] est utilisé pour évaluer la qualité de la parole avec une grande précision, en utilisant des signaux d'ancrage (aussi appelés signaux de référence). Ce test comporte deux étapes. La première consiste à identifier le signal d'ancrage (la référence cachée) proposé parmi les signaux dégradés. La deuxième consiste à demander aux sujets de donner leurs jugements de qualité vocale sur une échelle continue représentée sur le Tab. I.4, par rapport au signal d'ancrage. La première étape est évidente pour des signaux comportant de fortes dégradations, mais elle est aussi très utile dans le cas de faibles dégradations afin de vérifier les capacités des sujets à bien détecter les défauts.

Il est possible de tester un maximum de 15 signaux sur une même interface avec au moins un signal de référence caché. Le test MUSHRA se comporte alors comme un test de comparaison par paire entre chaque signal testé.

Note	Qualité de la parole
80-100	Excellente
60-79	Bonne
40-59	Passable
20-39	Médiocre
0-19	Mauvaise

Tab. I.4 Echelle catégorielle de mesure pour le test MUSHRA (UIT-R BS.1534 [12])

I.3. Evaluation subjective multidimensionnelle de la qualité vocale

La qualité vocale est un phénomène multidimensionnel faisant intervenir des attributs perceptifs (Jekosch [7], Gabrielsson et Sjogren [13], Wältermann *et al.* [14]). Les tests subjectifs présentés dans la partie précédente (§.I.2) utilisent le jugement unidimensionnel de la qualité vocale, en présentant directement une échelle de qualité vocale, afin de déterminer la note de qualité sous forme d'un scalaire. Cependant, cette tâche cognitive est complexe car elle est le résultat du jugement de la composition des paramètres perceptifs du signal vocal, par rapport à la composition des paramètres désirés (Jekosch [15] cf. §.I.1.4). Par exemple, lorsque la voix est transmise sur les différents réseaux téléphoniques, elle est soumise à des dégradations physiques, comme par exemple le codage de la parole en bande étroite qui réduit considérablement la largeur de bande (300 Hz – 3,4 kHz), par rapport à une communication face à face (20 Hz - 8 kHz pour la parole). En présence de cette simple dégradation physique, les attributs perceptifs peuvent être multiples (p. ex. le timbre, le naturel de la voix, la bruyance…).

Ces attributs perceptifs peuvent alors être représentés sous forme de dimensions perceptives d'un espace. Dans certains cas, l'espace obtenu peut être orthogonal lorsque les dimensions sont indépendantes les unes des autres ou encore lorsqu'elles ne sont pas corrélées entre elles.

La note globale unidimensionnelle de qualité vocale notée *MOS-LQS* est issue d'un test subjectif (cf. §.I.2). Elle peut être modélisée par une combinaison des dimensions perceptives d_{ni}, en minimisant l'erreur commise représentée par ε. La plupart des auteurs optent pour ce type de modèle appelé **modèle linéaire multiple** (Petersen et Hansen [16] Hall [17] Wältermann et Raake [18]). Ce modèle consiste à représenter la note globale de qualité vocale comme la somme des N dimensions d_{ni} pondérées par les coefficients c_n :

$$MOS - LQS_i = c_0 + \sum_{n=1}^{N} c_n . d_{ni} + \varepsilon \qquad \textbf{Eq. I.1}$$

Ce modèle est déterminé en effectuant une régression linéaire multiple entre les notes globales de qualité vocale et les coordonnées dans l'espace suivant les N dimensions.
Dans le cas où l'auteur considère qu'il existe des interactions entre les différentes dimensions de l'espace perceptif, un modèle linéaire multiple avec interactions peut être utilisé comme l'ont fait par exemple Mattila [19] et Wältermann *et al.* [14].

La note MOS-LQS peut aussi être estimée par un modèle utilisant les techniques d'apprentissage automatique (machine-learning en anglais) telles que les modèles "K-Nearest Neighbor", "Neural Network" ou encore "Genetic Programming" qui consistent à rechercher la meilleure représentation de la variable expliquée par les variables explicatives en utilisant toutes sortes d'outils (prise en compte des interactions, relations logarithmiques, exponentielles ou encore polynomiales). Cependant, ce genre de modèle est difficile à expliquer à cause des différentes interactions et transformations des différentes dimensions perceptives utilisées pour obtenir le score global de qualité vocale. Ces modèles sont souvent assimilés à une boîte noire. Il sera donc plus difficile d'effectuer un diagnostic de la qualité de la transmission avec ce type de modèle.

Il est aussi possible d'utiliser des modèles de classification comme une étude de Zimmer *et al.* [20] traitant de la représentation du désagrément de différents types de sons, où les auteurs mettent en évidence deux catégories correspondant aux signaux bruyants et calmes. Le désagrément est alors représenté par deux relations suivant le caractère bruyant ou calme des stimuli.

Quelle que soit la méthode utilisée, la qualité vocale est représentée comme une combinaison des dimensions de l'espace perceptif, représentative des différences de sonorité entre les différents stimuli testés. La suite de cette partie référence les deux principales méthodes de détermination de l'espace perceptif suivant que l'auteur a un *a priori* ou non sur les attributs perceptifs pertinents à la représentation des dimensions. Ces méthodes sont appelées respectivement **méthode de détermination de l'espace perceptif avec *a priori* et méthode de détermination de l'espace perceptif sans *a priori***.

I.3.1. Méthode de détermination de l'espace perceptif avec *a priori*

Cette première méthode de détermination de l'espace perceptif est utilisée lorsque l'expérimentateur a un *a priori* sur les attributs représentant les dimensions de l'espace (McGee [21], Gabrielsson et Sjogren [13], Petersen et Hansen [16], Bappert et Blauert [22]). Elle est généralement constituée de deux étapes :

- **Un test subjectif par méthode différentielle sémantique** quantifie l'impact de chaque attribut pour chaque stimulus.
- **Une Analyse en Composantes Principales (ACP)** détermine les attributs les plus pertinents à la représentation de l'espace perceptif.

I.3.1.1. Méthode par différentielle sémantique (SD)

Le test subjectif par méthode différentielle sémantique consiste à évaluer l'influence de certains attributs sur des stimuli testés (Osgood *et al.* [23]). Il est utilisé dans les études de Gabrielsson et Sjogren [13], Bappert et Blauert [22] et Petersen et Hansen [16] à partir des attributs référencés par Voiers [24]. Cette méthode consiste à proposer aux sujets un ensemble d'échelles continues qui sont nominales ou bien bipolaires, décrivant sémantiquement les dimensions perceptives. Les sujets ajustent les curseurs correspondant aux attributs perceptifs pour chaque condition. Par exemple, en ce qui concerne la brillance de la voix, les sujets auront à ajuster le curseur entre les adjectifs *sourd* et *clair*. Les curseurs sont souvent présentés sur une même interface, les uns après les autres. Une variante de cette méthode est de positionner les sons dans des espaces bidimensionnels représentant à chaque fois deux attributs.

Le principal problème de la méthode par différentielle sémantique est qu'il est nécessaire de connaître au préalable les attributs perceptifs correspondant aux dimensions subjectives.

Par ailleurs, ce test est aussi utilisé en complément à la méthode sans *a priori* de la détermination de l'espace, pour identifier les dimensions d'un espace perceptif prédéfini (p. ex. Hall [17], Mattila [19], Wältermann *et al.* [14], cf. §.I.3.2.3).

I.3.1.2. Analyse en Composantes Principales (ACP)

L'Analyse en Composantes Principales (ACP) est une méthode de la famille de la statistique multivariée, qui permet de manipuler et de synthétiser de grandes bases de données. Cette méthode consiste à sélectionner seulement les variables les plus explicatives et les plus pertinentes à l'exploitation des résultats. Ces nouvelles variables sont appelées les composantes principales des données.

Les variables les plus pertinentes seront données par ordre d'importance. Un choix doit alors être fait sur le nombre de variables à considérer, à partir du compromis entre la variance expliquée et le nombre limité de variables. Les variables non prises en compte constituent les axes bruités, considérés comme inutiles et n'apportant que très peu d'informations supplémentaires à l'analyse. Cette méthode peut être graphique en proposant un nouvel espace, ou alors

d'ordre statistique en déterminant le nombre de variables nécessaires à l'exploitation des résultats en fonction de l'erreur commise.

Une ACP est souvent employée à partir des résultats du test par la méthode différentielle sémantique, dans le but de sélectionner les dimensions perceptives les plus pertinentes pour la représentation des stimuli dans l'espace perceptif.

I.3.2. Méthode de détermination de l'espace perceptif sans *a priori*

Dans le cas où aucun *a priori* n'existe sur les attributs de l'espace perceptif, la méthode la plus courante consiste à effectuer une analyse d'échelonnement multidimensionnel (EMD) (Borg et Groenen [25]) à partir des résultats d'un test subjectif d'estimation des dissimilarités entre les différentes conditions de dégradation. Cette méthode permet de déterminer l'espace perceptif correspondant aux dissimilarités exprimées par les sujets.
Ensuite, chaque dimension de l'espace perceptif est identifiée par un attribut approprié.

I.3.2.1. Evaluation des dissimilarités

Le test de comparaison par paires est souvent utilisé afin d'évaluer les dissimilarités entre les échantillons sonores (Susini *et al.* [26], Mc Dermott [27], Mattila [19], Wältermann *et al.* [14]).

Ce test subjectif consiste à diffuser aux sujets toutes les paires d'échantillons sonores possibles. En général, pour limiter le nombre de paires à présenter aux sujets, on effectue un tirage aléatoire de l'ordre de présentation des stimuli au sein de chacune des paires afin d'annuler l'effet de l'ordre de présentation des stimuli. De plus, seulement quelques paires de stimuli identiques sont intégrées afin de vérifier la cohérence des réponses des sujets. Cela permet de réduire de moitié le nombre de paires à diffuser aux sujets. Le nombre total de paires de stimuli est de $n\cdot(n-1)/2$ paires de sons, n étant le nombre de conditions à tester. Les paires de stimuli sont diffusées aux différents sujets dans un ordre aléatoire, afin de s'affranchir de l'effet de l'ordre de présentation des paires.

Les sujets expriment leurs jugements de dissimilarité en utilisant une échelle de (dis)similarité pouvant être soit continue, soit catégorielle. Ces échelles sont souvent renseignées aux extrémités par les adjectifs *identiques* (ou *similaires*) et *très différents*.

Le test de comparaison par triade (test triadique) est utilisé entre autres par Hall [17]. Il consiste à diffuser trois stimuli à la suite et à demander à des sujets de déterminer la paire de stimuli la plus similaire ainsi que la paire de stimuli la plus différente. On attribue la valeur "0" à la paire jugée la plus similaire et "2" à la plus différente. On attribue la valeur "1" à la paire intermédiaire. L'échelle utilisée ici est donc ordinale à trois catégories, ce qui amène à traiter les données récoltées comme des données non-métriques. La matrice de dissimilarité est ensuite construite comme la somme des notes obtenues sur tous les sujets. L'utilisation de cette méthode triadique utilisant une échelle ordinale a surtout l'avantage d'être fractionnable en différentes séances, ce qui permet de considérer un plus grand nombre de conditions de dégradations. En effet, il n'est pas nécessaire que les sujets aient connaissance de l'ensemble des dégradations constituant la base sonore lors de la réalisation de la tâche de catégorisation des dissimilarités, contrairement à la procédure classique de comparaison par paires.

Le nombre de triades diffusées est de $n\cdot(n-1)\cdot(n-2)/6$. En comparant avec le nombre de paires diffusées par la méthode de comparaison par paires, il apparaît que le nombre de triades à diffuser est supérieur au nombre de paires à diffuser lorsque le nombre de conditions est supérieur ou égal à cinq. De plus, la méthode par triade nécessite d'écouter trois sons, contre

seulement deux dans le cas de la méthode de comparaison par paires. Le temps de réalisation du test triadique est donc plus important que celui du test de comparaison par paires.

La méthode de classification libre est utilisée par Lavandier [28] et Bernex et Barriac [29]. Cette méthode est proposée afin de pouvoir tester un nombre plus important de stimuli. La tâche demandée aux sujets consiste à regrouper les stimuli suivant leurs ressemblances. Le nombre de groupes n'est pas imposé, de sorte que les sujets constituent eux-mêmes les groupes de stimuli correspondant aux principaux attributs identifiés. Les dissimilarités sont ensuite calculées en considérant que les stimuli d'un même groupe sont similaires (0) et deux stimuli de deux groupes différents sont différents (1). La matrice des dissimilarités est ensuite calculée comme la moyenne de ces valeurs sur tous les sujets.

Bernex et Barriac [29] utilisent la matrice de dissimilarité pour déterminer l'espace perceptif par deux analyses distinctes : l'analyse par arbre et l'analyse par méthode d'échelonnement multidimensionnel (EMD). Il apparaît que ces deux analyses donnent des catégories similaires, mais avec des résultats qualifiés de "complémentaires". La méthode de l'arbre a l'avantage de départager les groupes caractérisant les principaux attributs de manière très précise et assez complète, tandis que l'analyse EMD représente de manière globale et quantitative les échantillons sonores dans l'espace perceptif.

I.3.2.2. Méthode d'échelonnement multidimensionnel (EMD)

Les méthodes généralement utilisées lors de l'analyse des données de dissimilarités appliquées aux sciences humaines sont les méthodes d'échelonnement multidimensionnel (EMD) encore appelées MultiDimensional Scaling (MDS). Ces méthodes consistent à modéliser les dissimilarités par des distances (euclidiennes la plupart du temps) pouvant être représentées dans un espace perceptif constitué d'un nombre minimal de dimensions, en ayant une erreur d'approximation minimale des dissimilarités par les distances (Borg et Groenen [25] et Escoufier [30]). Les méthodes de Torgeson [31], de Kruskal [32] puis de Caroll et Chang [33] sont brièvement présentées dans cette partie. Elles sont détaillées dans l'Annexe A.

A. Méthode EMD métrique de Torgeson

La première des méthodes EMD a été développée par Torgeson [31] en 1958. Elle est nommée la méthode EMD métrique ou Classical MDS (CMDS). Cette méthode utilise des données métriques de dissimilarités obtenues par le jugement moyen des sujets à l'aide d'une échelle métrique continue. Ces dissimilarités sont modélisées par des distances euclidiennes qui peuvent être représentées dans un espace euclidien à N dimensions. Plus le nombre de dimensions de l'espace euclidien est élevé, plus l'erreur de modélisation des dissimilarités sera faible, mais certaines dimensions ne participent que très peu à la diminution de l'erreur commise et peuvent être exclues de l'espace perceptif final. La méthode de Torgeson utilise une décomposition en valeurs singulières (SVD) afin de limiter le nombre de dimensions en considérant une limite acceptable de l'erreur commise de la modélisation des dissimilarités.

B. Méthode EMD non-métrique de Kruskal (NMDS)

Kruskal [32] a développé en 1964 une méthode EMD permettant d'échelonner des données non-métriques de dissimilarités. Cette méthode est utilisée entre autres par Mc Dermott [27].

Contrairement à Torgeson, Kruskal cherche directement une représentation de l'espace euclidien pour un nombre défini de dimensions qui vérifie la propriété de l'ordre des distances euclidiennes par rapport à l'ordre des dissimilarités.

C. Méthode EMD pondérée de Caroll et Chang (1970) [33]

Cette méthode, aussi appelée "Individual Difference Scaling" (INDSCAL) ou "weighted MDS" (WMDS), est une extension de la méthode de Kruskal en prenant en compte les différences interindividuelles des sujets. Elle est utilisée par Bappert et Blauert [22], Mattila [19], Wältermann *et al.* [14], Etame [34] ainsi que Gabrielsson et Sjogren [13]. Les données d'entrée utilisées par cette méthode peuvent être des matrices de dissimilarités métriques ou non-métriques, déterminées pour chaque sujet.

La configuration d'origine est tout d'abord déterminée par la méthode de Torgeson ou de Kruskal sur l'ensemble des résultats de dissimilarités moyennées sur les sujets.

La méthode INDSCAL considère que tous les sujets ont jugé les dissimilarités en considérant les mêmes dimensions perceptives, mais en leur attribuant des pondérations différentes selon les sujets. INDSCAL évalue donc les pondérations suivant chaque dimension pour chacun des sujets.

Les méthodes d'échelonnement multidimensionnel présentées dans cette partie sont utilisées lorsque le chercheur n'a aucun *a priori* sur le nombre de dimensions à prendre en compte pour représenter l'espace perceptif. Dans ce cas, le nombre de dimensions à considérer est choisi en analysant l'erreur commise lors de la reconstruction de la matrice des distances perceptives (cf. Annexe A). En pratique, le nombre optimal de dimensions est déterminé lorsque l'ajout d'une dimension n'ajoute que peu d'information supplémentaire à la reconstruction de la matrice des distances.

I.3.2.3. Méthode d'identification des dimensions perceptives

Lorsque les dimensions perceptives ont été déterminées, il est profitable de les identifier par des attributs afin de mieux comprendre leurs significations. Le test de verbalisation permet d'identifier sémantiquement les dimensions perceptives de l'espace. Il peut être réalisé de manière libre ou de manière forcée.

Dans le cas du test de verbalisations libres, il est demandé à des sujets de décrire les différences de sonorité perçues en utilisant leurs propres mots. Ce test permet de rassembler les attributs perceptifs les plus utilisés et les plus pertinents à la description des dimensions lorsqu' aucun *a priori* n'est admis. Cependant, les réponses données par les sujets peuvent être dans certains cas très diversifiées à cause de la nature subjective du jugement sensoriel des sons (p. ex. le niveau lexical, l'habitude des sujets à effectuer cette tâche, sujets experts ou musiciens ou encore naïfs). Néanmoins, les redondances des résultats permettent dans la plupart des cas d'obtenir de bonnes descriptions des dimensions perceptives.

Le test de verbalisations forcées consiste à proposer une liste d'adjectifs caractérisant les différences de sonorité de la voix. Ainsi, les sujets doivent choisir l'adjectif le plus pertinent à la sensation auditive perçue. Cette méthode de verbalisation forcée permet d'orienter les réponses des sujets sur des attributs prédéfinis afin de faire converger les résultats sémantiques, et de simplifier la tâche demandée. Cependant, il faut s'assurer que les attributs proposés sont interprétés de la même manière pour tous les auditeurs.

Ce dernier test peut être réalisé seulement si l'expérimentateur a un *a priori* sur les attributs perceptifs composant l'espace. Dans la plupart des études traitant de la détermination de l'espace perceptif avec *a priori* (Gabrielsson et Sjogren [13], Petersen *et al.* [16] ainsi que Bappert et Blauert [22]), les auteurs se réfèrent aux attributs perceptifs identifiés dans l'étude de Voiers [24]. Ce dernier a développé la méthode appelée DAM "Diagnostic Acceptability Measure" qui évalue l'acceptabilité de la transmission du signal de la parole dégradé par des

systèmes de communication. Cette méthode consiste à demander à des sujets de juger 20 attributs perceptifs à l'aide de 20 échelles continues à 100 points. Ces échelles correspondent à 10 attributs perceptifs liés à la qualité du signal de parole, 7 attributs perceptifs sont liés à la qualité du bruit de fond et 3 échelles concernent le jugement global de l'intelligibilité de la parole, de l'agrément d'écoute et de l'acceptabilité globale du signal. Les résultats montrent alors que les réponses font intervenir six attributs perceptifs relatifs au signal de la parole (battement, grincement, étroit, sourd, interruption, nasal) et quatre attributs perceptifs concernant le bruit de fond (bruissement, sifflement, buzz, bourdonnement).

Dans un cas idéal, les dimensions peuvent être identifiées de manière intuitive en écoutant les différences de sonorités suivant chacun des axes (Mc Dermott [27]). Dans certains cas, l'identification des dimensions n'est pas toujours évidente. Les méthodes classiques de Torgesson et de Kruskal utilisent les dissimilarités moyennées suivant les sujets, ce qui conduit à définir arbitrairement les axes de l'espace perceptif. De ce fait la correspondance entre les attributs perceptifs et les axes de l'espace perceptif n'est pas directe car il est possible que l'espace perceptif soit défini à une transformation près (Mc Dermott [27]). Les espaces perceptifs obtenus par les méthodes d'EMD peuvent être ajustés en utilisant certaines transformations comme par exemple la rotation varimax ou encore la transformation procrustéenne[4], afin de mieux faire correspondre chaque axe à un attribut perceptif.

La méthode INDSCAL présente la particularité d'utiliser les dissimilarités pour tous les sujets en prenant en compte les différences individuelles dont l'hypothèse de départ est que tous les sujets ont jugé les dégradations sur les mêmes axes perceptifs. Grâce à cette propriété, les axes déterminés par cette méthode sont directement corrélés aux attributs perceptifs.

Lorsque l'espace perceptif est déterminé par la méthode sans *a priori* employant l'EMD, les auteurs Hall [17], Mattilla [19] ainsi que Wältermann *et al.* [14] utilisent une analyse complémentaire afin d'identifier les dimensions de l'espace par des attributs perceptifs à l'aide du test différentiel sémantique (cf. §.I.3.1). Les échelles différentielles sont renseignées par les attributs relevés par un test de verbalisation ou alors par des attributs couramment utilisés (Voiers [24] cf. §.I.3.1). Les sujets ajustent ensuite les curseurs de chaque échelle sémantique correspondant aux conditions sonores. Les corrélations entre ces données perceptives et chaque dimension perceptive permettent de définir le(s) attribut(s) le(s) plus pertinent(s) à l'identification de chaque dimension. Cette technique est efficace lorsque plusieurs attributs perceptifs sont envisageables pour une dimension.

I.3.3. Conclusions sur les méthodes de détermination de l'espace perceptif

Bappert et Blauert [22] et Gabrielsson et Sjogren [13] utilisent et comparent les deux méthodes de détermination de l'espace perceptif (avec ou sans *a priori*).

Bappert et Blauert [22] utilisent les deux méthodes (avec et sans *a priori*) et concluent que celle avec *a priori* utilisant la méthode par différentielle sémantique fournit un espace perceptif pertinent, composé de deux dimensions. La méthode sans *a priori* utilisant l'EMD donne un espace tridimensionnel peu pertinent. Ce résultat est justifié, selon eux, par la sélection particulière des codecs.

[4] La transformation procrustéenne est une transformation linéaire incluant les translations, les réflexions, les changements d'échelles et les rotations orthogonales. La transformation procrustéenne des axes n'altère en rien le rapport entre les distances euclidiennes des différentes conditions.

Gabrielsson and Sjogren [13] traitent de l'évaluation de la qualité de plusieurs types de contenus sonores (musique, parole, et d'autres sons de la vie de tous les jours). Les échantillons sonores sont restitués par différents systèmes de diffusion (HP, casque ou encore par des appareils d'aide auditive). Pour chacun des échantillons sonores, les systèmes de restitution sont évalués en réalisant un test par méthode différentielle sémantique, un test d'évaluation des similarités et un test de verbalisation libre. Une analyse d'échelonnement multidimensionnel est ensuite utilisée. Les principaux attributs relevés sont la clarté, la précision, la brillance, la bruyance, la sonie et l'espace. Ces attributs ne sont pas toujours présents dans les espaces perceptifs déterminés selon les méthodes employées. Gabrielsson et Sjogren [13] concluent qu'il est profitable d'utiliser des méthodes de jugement variées (jugement de similarité, test par méthode différentielle sémantique, test de verbalisation libre), en combinaison, afin d'explorer par différents points de vue l'évaluation de la qualité vocale.

Il semble donc que la méthode sans *a priori* utilisant le jugement des dissimilarités soit adaptée à l'extraction de l'espace perceptif. Cependant, l'identification des dimensions perceptives de cet espace est parfois difficile.

La méthode avec *a priori* utilisant le test par méthode différentielle sémantique semble bien adaptée à l'identification des dimensions perceptives, cependant l'élaboration de l'espace perceptif dépend directement de la bonne sélection des attributs. Ce type de méthode produit souvent un biais causé par la sélection des attributs. Cet effet de biais est notamment relevé par Gabrielsson et Sjogren [13] qui ont constaté que les sujets avaient du mal à qualifier les différences de perception. Certains même disent qu'il est impossible de décrire sémantiquement certaines dégradations perceptives.

La méthode avec *a priori* est utilisée comme une méthode de détermination de l'espace perceptif à part entière, mais peut aussi être utilisée en analyse complémentaire à la méthode sans *a priori* pour identifier les dimensions perceptives.

I.4. Attributs perceptifs relatifs à la qualité vocale

L'identification verbale des dimensions d'un espace perceptif n'est pas obligatoire pour évaluer la qualité vocale. Cependant, elle constitue une aide non négligeable pour la compréhension du phénomène de l'évaluation de la qualité vocale et de la recherche d'indicateurs pertinents à la représentation des dimensions. Le vocabulaire utilisé pour décrire les différences de sonorité est complexe car il n'est pas toujours clairement défini, et peut varier suivant les individus interrogés. En général, les sujets experts ou les musiciens ont un vocabulaire plus riche, plus précis et souvent plus représentatif des différences de sonorité par rapport à des sujets naïfs (Gabrielsson ***et al.*** [35]). Les attributs perceptifs relatifs à l'évaluation de la qualité vocale sont référencés à partir des études réalisées de 1965 à nos jours. Ils sont traduits en français par les adjectifs les plus courants afin de les regrouper selon leurs significations. Chaque attribut est défini d'un point de vue sensoriel puis les dégradations physiques responsables des différences de sonorités sont présentées. Des indicateurs automatiques (aussi appelés descripteurs audio) sont parfois disponibles pour prédire les dimensions perceptives. Le paragraphe I.4.10 récapitule les principaux attributs perceptifs utilisés pour évaluer la qualité vocale.

I.4.1. Bruyance

La *bruyance* est définie comme un bruit de fond présent dans un signal de parole. A l'issue des tests de verbalisations libres (cf. §.I.3.1), les termes connexes à la bruyance sont craquement [13, 19, 29], souffle [29], grésillement [34], grincement [17, 19], sifflement [13, 29, 34, 36], bourdonnement [17], [19] ou encore bruissement [19].

La bruyance représente l'une des dimensions utiles à l'évaluation de la qualité vocale (Hall [17], Mattila [19], Bernex et Barriac [29], Wältermann *et al.* [14, 18], Etame [34]). Cet attribut est identifié quelle que soit la technique de communication utilisée (RTC, RNIS, GSM, VoIP).

La bruyance peut être causée par différents phénomènes physiques :

- Le bruit issu du réseau peut être induit par différents types de dégradations. Le codage et le décodage de la voix génèrent des bruits de quantification ressemblant à du bruit rose. Les transmissions analogiques peuvent aussi générer du bruit, suivant la distance entre les deux locuteurs. Le bruit provenant du signal électrique classique est constitué de la fréquence de 50 Hz et de ses harmoniques. Le terminal peut aussi générer du bruit de fond.
- Le bruit d'environnement est, comme son nom l'indique, le bruit environnant les interlocuteurs. Il peut être de nature très variée suivant le lieu, l'ambiance et le contexte où est situé le locuteur. Le bruit présent du côté du locuteur est capté par le microphone au même titre que le signal de parole, et transmis jusqu'aux oreilles de l'auditeur.

Dans un contexte d'écoute, il a été remarqué que le bruit d'environnement présent du côté de l'auditeur n'est pas gênant grâce aux facultés auditives de discrimination entre les deux oreilles (libre et au combiné) de cet auditeur (Möller [5]). Cela peut être justifié par le phénomène appelé BMLD « Binaural Masking Level Difference » (Zwicker [37]). Cet effet peut aider l'auditeur dans la tâche de ségrégation des sources sonores afin de faire abstraction du bruit de son environnement pour se focaliser exclusivement sur le message de son correspondant. L'auditeur peut aussi limiter l'influence du bruit de son environnement en augmentant le volume de son haut-parleur ou en s'isolant dans un endroit plus calme. Il jouera ainsi sur le rapport signal sur bruit.

Le bruit de fond provenant du côté du locuteur va, quant à lui, être capté par le microphone du terminal d'entrée au même titre que le signal de la parole, et sera donc soumis aux mêmes dégradations, celles liées à la transmission du signal (codecs, pertes de paquets).

La présence de bruyance sur un signal vocal influence la perception de la parole et fait diminuer le jugement de la qualité vocale avec l'augmentation du niveau sonore du bruit. Lorsque le bruit de fond a un niveau sonore suffisamment élevé par rapport à celui du signal vocal, la bruyance engendre **les phénomènes de masquage fréquentiel et temporel** du signal vocal (Zwicker et Fastl [38]). Cela provoque une baisse de la qualité vocale et parfois même, une diminution de l'intelligibilité de la parole. Les phénomènes de masquage dépendent principalement du rapport de niveau sonore entre le signal de parole et le bruit de fond, mais dépendent aussi de leurs caractéristiques fréquentielles et temporelles.

Il existe plusieurs études mettant en évidence, en plus du masquage énergétique de la parole, une **influence du masquage informationnel du bruit de fond, lors de l'estimation de l'intelligibilité de la parole** (Festen et Plomp [39], Bronkhorst et Plomp [40], Brungart *et al.* [41], Grataloup *et al.* [42], Hoen *et al.* [43], Durlach *et al.* [44] et Rhebergen *et al.* [45]).

Par exemple l'étude présentée par Hoen *et al.* [43] utilise des bruits de cocktail party[5] constitués de 4, 6 ou 8 signaux de parole, masquant un signal cible constitué de mots isolés. Lorsque le nombre de voix du bruit de cocktail party est élevé, il est difficile, voire impossible de discerner les mots de ce bruit. Les bruits de cocktail sont aussi présentés de manière inversée afin de supprimer le contenu lexical tout en conservant le contenu phonétique. Il est aussi présenté par des bruits modulés par les signaux de parole constituant le bruit de cocktail, le rendant ainsi dépourvu de contenu lexical et phonétique.

Les auteurs remarquent alors que l'intelligibilité de la parole est meilleure lorsque les conditions sont présentées avec du bruit modulé qu'avec du bruit de cocktail diffusé à l'endroit ou à l'envers. Cela révèle qu'il existe une influence du masquage informationnel lors de l'estimation de l'intelligibilité de la parole.

Les auteurs montrent ensuite que les conditions de bruit de cocktail constitué de 6 voix obtiennent de meilleurs résultats d'intelligibilité que lorsque le bruit est constitué de 4 voix. Il semblerait que le bruit constitué de 4 voix ait un effet de masquage informationnel supérieur à cause de son contenu lexical plus important que dans le cas du bruit constitué de 6 voix. Cet effet est vérifié en comparant les résultats d'intelligibilité selon les bruits constitués de 4 voix et celui constitué de 4 voix inversées. Il existe donc plusieurs types de masquage informationnel causé par le contenu phonétique et le contenu lexical.

Par ailleurs, l'étude de Grataloup *et al.* [42] montre qu'il existe une influence de la fréquence des mots constituant le bruit masquant lors de la reconstruction de mots cibles. Lorsque les mots constituant le bruit de cocktail party sont de basse fréquence (mots peu courants), le pourcentage de reconstruction du signal cible est plus élevé que lorsque les mots du bruit sont de forte fréquence (mots courants). Les auteurs proposent deux hypothèses pour expliquer ce phénomène. La première concerne l'effet attentionnel de l'auditeur qui serait perturbé lors de mots courants et moins perturbé lors de mots de fréquences basses, lors de la reconstruction du signal cible. La deuxième hypothèse concerne la différence de réactivité de l'auditeur pour les mots des deux catégories. Les mots de basse fréquence du cocktail créeraient une réactivité moindre de la part de l'auditeur et influenceraient donc moins la reconstruction des mots cibles.

En d'autres termes, il suffit qu'un message contenu dans le bruit de fond soit entendu et reconnu comme étant un élément important pour l'auditeur interrogé, pour que celui-ci diminue ses capacités attentionnelles pour le message cible et focalise aussi son attention sur le message du bruit de fond. Cela crée, en général, une diminution de la compréhension du message cible.

L'influence du masquage informationnel du bruit de fond lors de l'estimation de l'intelligibilité est observée pour des rapports signal sur bruit négatifs. Lorsque le RSB atteint des valeurs de 0 dB, l'intelligibilité de la parole est en général acceptable.

D'autres études montrent **une influence du contenu informationnel du bruit de fond** lors de différentes analyses psychoacoustiques.

Ellermeier *et al.* [46] présentent une expérience visant à évaluer l'influence du contenu informationnel de son lors de l'estimation de la sonie. Le test consiste à diffuser 40 sons dans leurs versions originales (avec du contenu informationnel) et leurs correspondants dépourvus d'informations grâce à la méthode développée par Fastl [47]. Les résultats montrent pour certains types de sons (alarme, cloche et sonnette) une influence du contenu informationnel du son lors de l'évaluation de la sonie. Dans le cas du son d'alarme et de sonnette, l'étude montre

[5] Le bruit cocktail party est constitué d'un mélange de plusieurs voix superposées. Ce bruit est caractéristique des réunions festives arrosées, d'où son nom.

que la sonie des sons originaux est plus importante que la sonie des sons dépourvus d'information. L'effet inverse est observé dans le cas du son de cloche.

Une deuxième étude d'Ellermeier *et al.* [48] est réalisée sur les mêmes 40 sons afin d'évaluer l'influence du contenu informationnel du son lors de l'estimation de la gêne. Les résultats montrent qu'en général, les sons originaux présentant du contenu informationnel sont moins gênants que leurs correspondants dépourvus d'informations.

Scholz *et al.* [36] étudient la dimension bruyance lors de l'évaluation de la qualité vocale appliquée à la téléphonie. Ils remarquent qu'il existe une influence du type de bruit de fond lors de l'évaluation de la qualité vocale entre les bruits de fond d'environnement (cocktail party, marteau), les bruits de circuits et les bruits de terminaux.

Bernex et Barriac [29] observent aussi une influence entre les différents types de bruits de fond lors de la catégorisation libre de signaux. Pour un rapport signal sur bruit assez faible, les résultats de la classification libre des échantillons sonores font apparaître deux nouveaux groupes pour les conditions jugées non-dérangeantes. Le premier groupe englobe des bruits d'environnement (rue, Train) et le deuxième groupe rassemble du bruit de souffle (bruit de ventilateur d'ordinateur).

Certains des auteurs cités précédemment utilisent des descripteurs afin de mesurer automatiquement le degré de bruyance perçu d'un échantillon vocal.

Le descripteur le plus utilisé est le niveau sonore du bruit de fond calculé par rapport à l'énergie moyenne du signal. Il est souvent déterminé par le rapport signal sur bruit (RSB).
Le RSB s'exprime en dB par la formule suivante :

$$RSB = 10.\log_{10}\left(\frac{S}{B}\right), \qquad \textbf{Eq. I.2}$$

avec, *S* et *B* représentant respectivement la puissance moyenne du signal de la parole et du bruit de fond.

Il existe aussi d'autres indicateurs représentatifs du niveau sonore perçu par l'Homme, comme par exemple les modèles de sonie de Zwicker [38] et de Moore [49], et les diagrammes de sonie spécifique du bruit et de la parole.

I.4.2. Bruit sur la parole

La dimension *bruit sur la parole* est définie comme un bruit ayant une enveloppe temporelle similaire au signal de parole cible. Le *bruit sur la parole* est présent uniquement sur les zones actives de la parole, contrairement aux bruits présentés dans la partie précédente qui sont présents sur l'ensemble du signal (cf. §.I.4.1). Le *bruit sur la parole* peut causer un phénomène de masquage de la parole. Dans certains cas, il peut aussi être perçu comme une déformation ou une distorsion de la voix et influencer le naturel de la voix (cf. §.I.4.6).

Etame [34] propose un espace perceptif à quatre dimensions, composé en plus de la deuxième dimension *bruyance*, d'une troisième dimension *bruit sur la parole*. De nombreuses études ont aussi relevé l'attribut perceptif *bruit sur la parole* lors de tests de verbalisation libre. Cependant, cet attribut est souvent admis dans la dimension perceptive de la bruyance (Mattila [19], Hall [17], Scholz *et al.* [36], Wältermann *et al.* [14]). Par exemple, dans l'étude de Scholz *et al.* [36], l'espace perceptif est déterminé par les deux méthodes avec et sans *a priori* (cf. §.I.3.2) pour diverses conditions de dégradations bruitées (terminal, cocktail party,

marteau, MNRU[6]) diffusées à plusieurs niveaux sonores. Les résultats montrent que la dimension bruyance peut elle même s'exprimer par un sous-espace perceptif à trois dimensions représentées par le niveau sonore du bruit de fond additif, la composition spectrale du bruit et la quantité de bruit corrélé à la parole.

Cette dégradation est causée par l'approximation faite lors du codage de la parole. Le codage G.726 utilisant la technique ADPCM (Adaptive Differential Pulse Code Modulation) est particulièrement touché par ce type de dégradation. D'autres codecs sont susceptibles de générer du bruit sur la parole, comme les codages GSM (GSM EFR, GSM HR, AMR). Etame [34] suggère que cette dimension est liée au débit et à la famille du codage utilisé : *"On observe suivant la dimension 3 que les objets sonores sont rangés* ***par débit croissant au sein de chaque famille de codecs****. En écoutant les stimuli suivant cette dimension, l'attribut qui caractériserait la dimension 3 serait "bruit sur la parole", comme l'attestent également les résultats de la verbalisation."*

Cette dégradation peut aussi provenir d'une détérioration du haut-parleur du terminal utilisé, provoquant de la distorsion sur les zones du signal de forte intensité. Elle est alors perçue comme un bruit de grésillement sur les zones de parole.

Etame [34] propose de modéliser cette dimension perceptive en considérant qu'elle est causée par un phénomène additif au signal original. L'indicateur calcule la différence entre le signal non-dégradé et le signal dégradé afin de représenter cette dimension (modèle intrusif).

Le modèle P.563 [2] estime la dimension *bruit sur la parole* en évaluant les statistiques spectrales du signal pendant les zones actives du signal, en supposant que le bruit ajouté forme un "plancher de bruit" dans le domaine spectral.

Falk et Chan [50] évaluent la dimension *bruit sur la parole* en utilisant une technique similaire à celle du modèle P.563. Cette technique repose sur l'utilisation des paramètres PLP (Perceptual Linear Prediction) et de la moyenne des écarts des distances cepstrales (cf. §.I.4.6). En comparant des spectres fréquentiels de signal de parole non dégradé et dégradé par du bruit ajouté au signal de parole, il apparaît que le spectre devient plus lisse lorsqu'il est bruité que lorsqu'il est non dégradé.

Le signal présentant du bruit sur la parole *y(n)* peut être représenté par l'équation suivante :

$$y(n) = s(n) + s(n).10^{-Q/20}.N(n) \qquad \textbf{Eq. I.3}$$

où *s(n)* est le signal non dégradé et *N(n)* le bruit blanc gaussien. Le niveau de bruit sur la parole est contrôlé par le paramètre *Q* exprimé en dB qui représente le rapport entre la puissance du signal d'entrée et la puissance du signal du bruit sur la parole.

I.4.3. Continuité - discontinuité

La dimension *continuité* est assimilée à la présence d'artefact temporel sur le signal vocal, comme par exemple des coupures dans le signal. Cette dimension est identifiée à partir de 1997, avec l'apparition de la transmission de la voix par le réseau mobile et surtout par le protocole IP (Petersen *et al.* [16], Mattila [19, 51], Bernex et Barriac [29], Wältermann *et al.* [14], [18], Etame [34]). Des tests de verbalisation libre ont relevé des attributs connexes à la continuité comme interruption [14, 16, 18, 19], fluctuant [19], craquement [19], métallique [19], lisse [18, 19], coupures [29], clips [29], silence [14, 29], grincement [16].

[6] MNRU : Le bruit « Modulated Noise Reference Unit » est un bruit de quantification généré par le codage de la parole.

L'origine la plus courante de la présence de discontinuité est la transmission de l'information sur le réseau IP basée sur la paquétisation des informations du signal de la parole (cf. §.I.1.2.4). Les paquets sont envoyés sur le réseau Internet et reçus par un récepteur qui décode les paquets et restitue leur contenu. Il se peut que certains de ces paquets arrivent trop tard ou qu'ils soient perdus en route, générant des coupures du signal vocal. Ce phénomène peut survenir de manière aléatoire, mais aussi de manière consécutive. On parle alors de pertes de paquets par rafale ou encore de "burst".
Certains codecs, comme le G.711 par exemple, permettent de réduire cette perception de discontinuité en utilisant un algorithme de PLC (Packet Loss Concealment "UIT G.711 App.1" [52]). Cet algorithme masque les zones de discontinuité en les remplaçant par un signal de synthèse généré à partir des paquets précédents, ou encore par un bruit. Cette technique limite la dégradation provoquée par les discontinuités. Il subsiste parfois des résidus à l'emplacement des anciennes discontinuités, mais ceux-ci sont plus légers, plus subtils et moins dérangeants.

Voran [53] propose une étude perceptive des dégradations de pertes de paquets en les analysant suivant trois états : **pauses**, **pertes** ou encore **sauts de paquets**. Dans le premier cas, aucun paquet n'est perdu, mais un silence de la taille d'un paquet est ajouté. La deuxième configuration représente une perte de paquet remplacée par un silence, et la troisième représente un saut de paquets. Dans ce cas, il n'y a pas de silence. Les résultats d'un test subjectif de type "ACR" (cf. §.I.2.1) montrent qu'il n'y a pas de différences significatives lors de l'évaluation de la qualité vocale. Voran [53] propose alors une relation linéaire suivant la durée des pertes ainsi que la moyenne du taux de pertes par trames afin de déterminer la note de qualité vocale.

Huo *et al.* [54] suggèrent que la dimension *discontinuité* peut elle-même s'exprimer par un sous-espace perceptif composé de trois dimensions. L'étude utilise des conditions de dégradation relatives aux codages de la parole (G.711 / G.728 / G.729A), aux pertes de paquets (0 à 20 %), aux clips, puis aux bruits musicaux (simulant les imperfections des algorithmes de réduction de bruit). Les trois sous-dimensions de la discontinuité sont déterminées par la méthode sans *a priori* (cf. §.I.3.2), puis identifiées comme interruption (causée par la perte de paquets), artefact ajouté (utilisation de PLC) et classification du bruit musical (imperfections des algorithmes de réduction de bruit).

Wältermann *et al.* [14] montrent aussi que les dégradations dites "musical tones" provoquées par l'utilisation de certains algorithmes de réduction de bruit peuvent générer une perception de discontinuité.

Dans l'étude de Mattila [19], la dimension continuité est causée par les erreurs de bits présentes lors d'une transmission radio. Ces dégradations peuvent être décrites par le taux "BER" (Bit Errors Ratio) dans le cas d'une transmission mobile, ou encore le "FER" (Frame Errors Ratio). La dégradation de type BER est générée lors d'une inversion du contenu d'un bit : lorsqu'un "0" est transmis à la place d'un "1" et inversement. La dégradation de type FER est générée par des pertes de trames contenant un certain nombre de bits suivant le débit du codeur utilisé. Ces pertes sont perçues comme des interruptions du signal comme dans le cas des pertes de paquets lors d'une transmission IP.

Les modèles actuels d'évaluation de la qualité vocale utilisent principalement le taux de pertes de paquets. Cet indicateur paramétrique est disponible à partir des informations issues du buffer de gigues qui est positionné au niveau du récepteur afin d'absorber les fluctuations

de temps d'arrivée des paquets (la gigue). La taille des buffers étant limitée, une gigue excessive ou une absence de paquets peut entraîner un débordement du buffer, ce qui se traduit par des paquets perdus.
Le taux de pertes peut être pondéré par le codage utilisé suivant sa résistance aux pertes (p. ex. utilisation d'un PLC ou non). Le modèle E [3] représente cette dégradation de pertes de paquets aléatoires par le facteur appelé *Ie,eff* :

$$Ie,eff = Ie + (95 - Ie).\frac{Ppl}{Ppl+Bpl}, \quad \textbf{Eq. I.4}$$

avec *Ie* "equipment Impairment factor" le facteur de dégradation du codage utilisé et *Bpl* "packet-loss robustness factor" le facteur qui représente la capacité du codage à résister aux pertes de paquets. Ces deux indicateurs sont dépendants du codage utilisé et disponibles dans le rapport de l'ITU-T G.113 [55].

Lijin Ding *et al.* [56] utilisent aussi les indicateurs paramétriques issus des statistiques réseaux pour représenter les dégradations de pertes de paquets. Les informations récupérées sont les coefficients *Bpl*, le type de codec, la taille du paquet, le taux de pertes ainsi que l'information concernant le type de pertes (aléatoires ou par rafale). De plus, une classification Silence / Unvoiced / Voiced est proposée pour évaluer l'effet des pertes de paquets selon l'emplacement des pertes grâce à l'outil développé par Rabiner et Sambur [57]. Le facteur proposé pourrait aussi être utilisé par le modèle E [3] à cause de la forme de la structure globale proposée. L'effet des pertes de paquets est déterminé à partir du facteur de dégradation "$DMOS_{PL}$".

$$\mathrm{DMOS}_{\mathrm{PL}} = W_U.\mathrm{DMOS}_{\mathrm{PLu}} + W_v.\mathrm{DMOS}_{\mathrm{PLv}} \quad \textbf{Eq. I.5}$$

W_u et W_v sont les coefficients de pondération suivant l'emplacement des pertes "*u*" pour Unvoiced *et "v"* pour Voice, correspondant aux deux facteurs de pertes $DMOS_{Plu}$ et $DMOS_{PLv}$. Ces deux facteurs de pertes sont déterminés par des relations polynomiales d'ordre 3 déterminées par

$$DMOS_I = C_0 + C_{1i}.ulp + C_{2i}.ulp^2 + C_{3i}.ulp^3, \quad \textbf{Eq. I.6}$$

avec *ulp* représentant le pourcentage de pertes de paquets. Les coefficients C_i dépendent du codec et du type de PLC utilisé.

La dimension continuité peut aussi être représentée par des indicateurs basés sur le signal comme dans le cas du modèle P.563, du modèle de Kim *et al.* [58] et le modèle de Falk et Chan [50].

Le modèle P.563 [2] détecte les discontinuités à partir des chutes du signal temporel, caractérisées par un effondrement profond avec des bords abrupts définis par des modifications importantes de niveau sonore. Quatre descripteurs sont ainsi définis caractérisant la durée des silences, les chutes abruptes, les silences non-naturels et la détection d'interruption de signal :

- Afin d'estimer la durée des silences, un profil de niveau est tout d'abord réalisé pour chacune des trames de 40 ms. Les zones de silence sont alors détectées lorsque l'écart de niveau entre la trame *n* et *n+1* est supérieur à 30 dB. La durée des silences est déterminée dans la limite supérieure de 1 seconde.
- Afin de localiser les chutes abruptes du signal, le modèle part de l'hypothèse qu'un signal vocal naturel ne peut jamais baisser abruptement, aussi bien au début qu'à la fin de

l'émission de la parole. Pour cela, les puissances du signal sont mesurées et comparées entre elles toutes les 50 ms. Lorsque l'écart des puissances entre deux trames successives atteint une certaine valeur limite une deuxième procédure consiste à calculer deux indicateurs de mesure de périodicité fondés sur deux corrélations croisées, afin de définir s'il s'agit d'une interruption brute ou seulement de la fin d'une voyelle.

- Les silences non-naturels sont déterminés à partir de la différence de niveau entre toutes les trames successives de 320 ms. Si le niveau varie plus de quatre fois, la trame est considérée non naturelle.
- L'algorithme de détection des interruptions d'un signal fournit plusieurs informations comme la position de l'interruption, sa durée, sa puissance estimée et la fréquence fondamentale estimée juste avant l'endroit de l'interruption.

Le modèle de Kim *et al.* [58] prend en compte les interruptions sur les zones actives du signal (utilisation d'un outil de DAV "Détection d'Activité Vocale") en recherchant les discontinuités abruptes non-naturelles de départ et de fin. Les auteurs utilisent 12 indicateurs MFCC ainsi que leurs dérivées par rapport au temps, la puissance sonore de chacune des trames et la fonction d'auto corrélation de la voix afin de distinguer les silences non-naturels abrupts des sons naturels de fin comme /p/ ou /t/. Ils proposent aussi un modèle de récence suivant la localisation des pertes en début ou en fin du signal à évaluer.
Cette technique, aussi utilisée dans le modèle P.563 [2], atteint ses limites lorsque les pertes de paquets du signal vocal ne sont pas constituées de silence. C'est le cas lors de l'utilisation d'algorithmes de PLC et de la présence de discontinuité par saut (sans zones de silences). Le signal est soumis à une discontinuité, cependant il n'y a pas d'interruption du signal et le niveau sonore reste constant entre les trames successives. Ces différents types de discontinuité ne sont donc pas pris en compte par ces modèles.

Falk et Chan [50] estiment les discontinuités du signal à partir des dérivées temporelles simples et doubles des distances cepstrales du signal (cf. §.I.4.6), afin de représenter respectivement la vitesse et l'accélération du changement du cepstre. Ils utilisent, comme le modèle de Kim, la détection des interruptions non-naturelles abruptes du signal de début ou de fin de parole, en proposant de détecter ces chutes de signal par la différence de niveau entre les trames *n* et *n+1*. Ensuite, une classification de type "Support Vector Classifier" SVC est réalisée à partir des paramètres définis préalablement afin d'identifier le type de discontinuité.

Křenek et Holub [59] proposent une mesure des discontinuités du réseau à partir de la visualisation des histogrammes, pour une application non-intrusive. En comparant les histogrammes d'un signal vocal original avec ce même signal soumis à des pertes de paquets, il apparaît une augmentation de l'énergie acoustique dans les zones de basses fréquences (inférieures à 300 Hz). Des tests subjectifs montrent que cette augmentation d'énergie dans les basses fréquences est corrélée à la baisse de l'évaluation de la qualité vocale.

I.4.4. Distorsion

La distorsion peut être perçue comme une caractéristique non-naturelle ou métallique du signal vocal.

La distorsion peut être générée par une déformation du signal sonore engendrée par l'utilisation des techniques de codage à bas débit (McDermott [27]). L'utilisation de filtres étroits est aussi la cause de cette dégradation perceptive.
Les indicateurs paramétriques issus des statistiques réseaux pourraient être une piste pour déterminer le codage utilisé et donner une information sur le niveau de distorsion.

Le modèle de Kim [58] évalue la distorsion à partir d'indicateurs basés sur le signal en séparant les zones actives et non-actives de la parole. Les descripteurs utilisés sont basés sur l'analyse de l'articulation de la parole.

La distorsion peut être assimilée à l'attribut voix naturelle-synthétique car elle génère une déformation de la voix naturelle.

I.4.5. Sifflement - bulleux

L'attribut bulleux peut être décrit comme le son engendré par une goutte d'eau tombant dans un contenu d'eau. Ce son d'impact est composé d'une fréquence fondamentale qui va augmenter de manière continue tout en s'estompant rapidement.

L'attribut sifflement a été identifié par Petersen et Hansen [16] et Etame [34] pour identifier une dimension utile à l'évaluation de la qualité vocale. Mattila [19] identifie quant à lui les attributs sifflement et bulleux pour représenter sa quatrième dimension perceptive.

Selon Etame [34], l'attribut sifflement peut être différencié suivant les zones actives ou non-actives du signal de la parole. Dans le cas des zones non-actives du signal de parole, le sifflement est perçu comme un son pur harmonique de haute fréquence, tandis que sur les zones actives du signal de la parole, le sifflement est perçu comme un effet de réverbération ou d'écho. Ce phénomène est confirmé dans l'étude de Mattila [19] qui considère que l'attribut bulleux est perçu comme de l'écho très faible de la parole.

Petersen et Hansen [16] identifient les attributs sifflement et craquement mais uniquement pour les zones comportant uniquement du bruit de fond. L'attribut sifflement concernerait donc principalement le bruit de fond tandis que l'attribut bulleux caractériserait les déformations sifflantes et/ou bulleuses de la voix.

Les dégradations à l'origine des attributs sifflement et bulleux peuvent être provoquées par certains codecs à bas débit, par l'apparition d'écho sur la parole ou encore par la présence d'un bruit de fond sifflant comme du vent. Etame [34] explique cette dégradation par l'utilisation du "codage descripteur d'insertion de silence" (SID, silence description) G.729 [60]. Il propose alors de représenter la dimension sifflement à partir d'indicateurs calculant les maxima des estimées des intercorrélations entre les spectres de puissance du signal original et de chaque version codée.

I.4.6. Voix naturelle-synthétique / voix de robot

Cet attribut révèle le niveau du naturel de la voix afin de quantifier les différences de sonorité entre des voix réelles et des voix synthétiques. Hall [17], Mattila [19] et Bernex et Barriac [29] ont identifié l'attribut *naturel de la voix* et *voix de robot* dans leurs espaces perceptifs respectifs. Les adjectifs employés généralement dans la littérature pour décrire une voix non naturelle sont métallique, synthétique, voix de robot, voix électronique. Cette dimension est corrélée aux attributs distance, gémissement nasal, crispé, mécanique, métallique, rugueux ou encore étouffé (Mattila [19]).

La dimension *voix naturelle* est fortement corrélée aux notes d'évaluation de la qualité vocale (Hall [17]). Ces deux grandeurs semblent être très proches l'une de l'autre.

Une voix considérée comme parfaitement naturelle peut être représentée par une transmission ne comportant aucune compression et approximation du signal vocal. Les techniques associées à la transmission téléphonique correspondent à la diminution de ces caractéristiques avec la diminution de la largeur de bande fréquentielle et l'approximation réalisée par le codage de la parole. Hall [17] et Mattila [19] montrent que les échantillons sonores codés à bas

débit comme par exemple par la technique CELP (Code Excited Linear Predictive), génèrent une sensation de voix synthétique (UIT-T G.729 [60]).
La présence de pertes de paquets lors d'une transmission de type VoIP peut aussi générer une perception de voix de robot à cause des coupures brèves et saccadées du signal vocal.

La simple représentation dans le domaine spectral ne suffit pas à distinguer les différentes caractéristiques de la production de la parole. Les indicateurs physiques souvent utilisés afin de quantifier le niveau de naturel de la voix sont basés sur les indicateurs cepstraux ainsi que les coefficients déterminés par la technique LPC[7] (Codage par Prédiction Linéaire). La représentation cepstrale souvent utilisée sous la forme des coefficients MFCC (Mel Frequency Cepstrum Coefficients) ainsi que les techniques LPC sont basées sur le modèle de production de voix. Ce modèle considère que la voix peut être reconstituée à partir d'une source (bruit ou son pur) représentant les cordes vocales et d'un filtre se comportant comme un résonateur afin de simuler le conduit vocal. Ces techniques décrivent les caractéristiques sonores de la parole par des indicateurs plus globaux et sont largement employées dans les domaines de la reconnaissance de la parole, du codage de la parole, du traitement et de la modification du signal et de la synthèse de la parole (D'Alessandro et Demars [61], Didiot [62] et Istrate [63]).

Le modèle P.563 [2] évalue le niveau du naturel de la voix à partir d'une analyse statistique faisant intervenir une analyse cepstrale ainsi que les coefficients LPC. En ce qui concerne la présence d'une voix de robot très prononcée, le modèle P.563 suggère que *"la robotisation est causée par un signal vocal qui contient une trop grande périodicité…*". Une technique est développée afin d'estimer le niveau de "robotisation" à partir de la périodicité du signal de la parole sur une bande de fréquences comprise entre 2,2 et 3,3 Hz. La périodicité est calculée par corrélation croisée des trames du signal vocal adjacentes sur des durées de 32 ms. Le modèle P.563 repère aussi les répétitions de trames dans le signal à l'aide de la corrélation croisée entre les trames, à l'origine de la voix de robot.

I.4.7. Clarté - intelligibilité

La clarté représente la facilité à comprendre l'information contenue dans la parole. Elle dépend principalement du contenu informationnel de la parole, des effets de la réverbération de la salle, mais aussi de la coloration de la parole (cf. §.I.4.9).

La clarté et l'intelligibilité de la parole sont attribuées à une dimension qui semble pertinente pour différencier des signaux de parole (Gabrielsson et Sjogren [13], Bappert et Blauert [22], McDermott [27], Wältermann *et al.* [18]).

La dimension clarté a aussi été identifiée par les attributs franc, précis, direct, clair, distant (Wältermann [18]) et les attributs clair, transparent, brillant, net par opposition aux adjectifs diffus, vague (Bappert et Blauert [22]). Les résultats de cette dernière expérience montrent que l'attribut clarté est corrélé aux attributs naturel, proche et surtout à l'intelligibilité (à 99%).

L'intelligibilité de la parole dépend directement de la clarté du signal émis et de la capacité cérébrale à traiter l'information issue de ce signal. L'intelligibilité de la parole est définie comme les capacités auditives et cognitives d'un individu à identifier la signification d'un mot ou d'un groupe de mots (cf. §.I.1.3).

[7] Les coefficients LPC (Linear Predictive Coding) sont utilisés afin de représenter un signal sonore sous forme de coefficients représentant les caractéristiques de la production de la parole (corde vocale + conduit vocal). Ces coefficients permettent de transmettre le signal sonore sous une forme compressée jusqu'en bout de la télécommunication où les coefficients sont utilisés afin de reconstruire le signal sonore.

Wältermann [18] suggère que cette dimension est liée à la perception de distance de la parole qui est principalement caractérisée par l'environnement et les équipements des interlocuteurs. Dans un contexte de télécommunication, la clarté ou l'intelligibilité est provoquée par les différents terminaux utilisés ainsi que l'emploi du kit main libre ou du haut-parleur.

Les indicateurs automatiques utilisés pour quantifier l'intelligibilité de la parole peuvent être utilisés, comme l'indice d'articulation (IA) et le score d'intelligibilité.

Indice d'articulation (IA) :

On peut déterminer physiquement l'effet de masquage à l'origine (entre autres) des problèmes d'intelligibilité en calculant l'indice d'articulation French et Steinberg [9] :

$$IA = \sum \alpha_i \cdot W_i \,, \qquad \textbf{Eq. I.7}$$

avec : i : bande de tiers d'octave pour 160 Hz < f < 8 kHz
W_i : Quantité d'information qui, dans la bande de fréquence 'i', parvient à l'auditeur malgré le bruit : $W_i = (Si + 12 - Bi) / 30$
Si : niveau moyenné temporellement du signal en dB
Bi : niveau moyenné temporellement du bruit en dB
α_i : importance de l'information sur la bande de fréquence
On considère que le masquage est total lorsque Si = Bi - 12

Remarques :

- l'IA ne doit pas être employé s'il existe plusieurs chemins de propagation de la parole bien marqués.
- Cet indice ne tient pas compte de la réverbération.

Cet indice d'articulation est relié à la qualité des communications (cf. Tab. I.5). Les adjectifs utilisés pour qualifier la qualité des communications sont similaires à l'échelle à cinq catégories employée pour obtenir la note MOS.

I.A.	Qualité des communications
1.0	Excellent
0.9	
0.8	
0.7	
0.6	Bon
0.5	Satisfaisant
0.4	
0.3	Mauvais
0.2	
0.1	Très Mauvais
0	Néant

Tab. I.5 Relation entre l'IA et la qualité des communications (French et Steinberg [9])

Le score d'intelligibilité est exprimé en pourcentage (Kryter [64]).

$$S = 100 \cdot (1 - 10^{-(IA/Q)}) \qquad \textbf{Eq. I.8}$$

Q : constante qui dépend du type de stimuli utilisés (logatomes, mots phrases... ex : $Q = 0{,}6$ pour des logatomes) (cf. Fig. I.3).

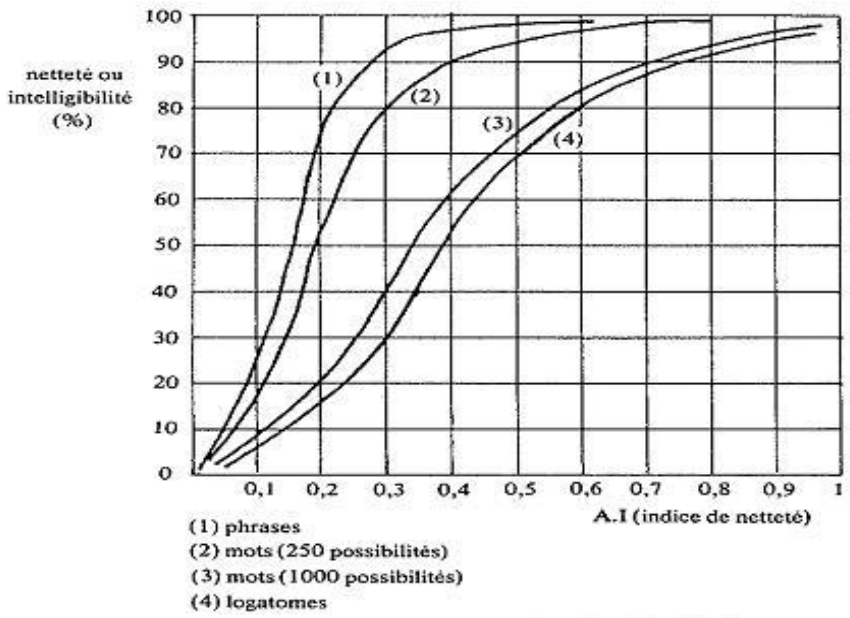


Fig. I.3 Pourcentage de netteté et d'intelligibilité en fonction de l'indice d'articulation pour 4 contenus différents (phrases, 250 mots, 1000 mots, logatomes) (Kryter [64])

La Fig. I.3 montre que pour le même indice d'articulation, les phrases se comprennent mieux que les mots et les mots sont plus compréhensibles que les logatomes.

I.4.8. Sonie de la parole

La sonie de la parole représente le niveau sonore de la parole perçu par l'utilisateur. McDermott [27] a identifié entre autres la sonie de la parole comme étant un attribut utile à l'évaluation de la qualité vocale. Cette dimension est aussi utilisée par Côté [10] dans le modèle DIAL "Diagnostic Instrumental Assessment of Listening quality".

Cet attribut perceptif peut être causé par les caractéristiques du terminal (microphone et haut-parleur) ou encore par le niveau sonore du bruit de fond ambiant, présent du côté du locuteur. Ce phénomène appelé effet Lombard montre qu'un locuteur adapte sa manière de parler lorsqu'il est situé dans un environnement bruyant, et principalement en augmentant le niveau sonore de sa voix afin de mieux se faire entendre (Garnier [4]).

Dans une application téléphonique, le niveau sonore de la voix peut être gênant lorsque ce niveau est trop faible, mais il peut être tout aussi gênant lorsqu'il devient trop fort. Le niveau sonore optimal fixé par l'UIT ("Handbook on Telephonometry" [65]) correspond à un niveau de 79 dB SPL en écoute monaurale. Dans le cas d'une écoute diotique, le niveau optimal est fixé entre 69 et 73 dB SPL.

Les modèles de Zwicker [38] ou de Moore [49] sont efficaces pour évaluer la sonie de sons stationnaires. Dans le cas de la parole qui est un signal non-stationnaire, d'autres indicateurs sont proposés comme STLmax, LTLmax, N5, N30 et LMIS (Loudness Model for Impulsive Sounds) dans le cas de sons impulsionnels. Une revue de ces indicateurs est présentée par Molla *et al.* [66].

I.4.9. Coloration-brillance

La coloration et la brillance représentent la perception de l'équilibrage des niveaux fréquentiels des échantillons sonores. La plupart des auteurs ont identifié cet attribut comme correspondant à une dimension perceptive utile à l'évaluation de la qualité vocale (Mc Gee [21], Gabrielsson et Sjogren [13], Bappert and blauert [22], Petersen et Hansen [16], Hall [17], Mattila [19], Wältermann *et al.* [18], Etame [34], Wältermann *et al.* [14]). Certains le désignent comme la brillance, le timbre ou le contenu de hautes ou basses fréquences. Il est aussi souvent identifié par les adjectifs sourd / clair, grave / aigu ou encore mat / brillant par analogie à l'image. Les adjectifs sourd / clair montrent qu'il y a un lien entre l'attribut coloration et l'attribut clarté présenté dans la partie I.4.7. La clarté (ou l'intelligibilité) d'un signal vocal pourrait être influencée par la répartition du contenu spectral de la parole, et réciproquement.

La coloration est principalement causée par le timbre original de la voix du locuteur et surtout entre les voix d'homme, de femme ou d'enfant.
La coloration de la voix transmise peut aussi être influencée par le type de codage utilisé. Par exemple, le codage G.711 utilisant la technique PCM reconstruit la voix de manière plus brillante, plus claire que le codage G.729 basé sur la technique ACELP (moins coûteux en débit).
La présence de bruit de fond dans l'environnement du locuteur peut dans certains cas provoquer un changement de coloration de la parole. Pour des timbres de voix et de bruit de fond similaires, l'effet de masquage fréquentiel peut être augmenté et provoquer une forte diminution de la qualité vocale. Il a été constaté dans de telles conditions qu'un locuteur s'adapte inconsciemment à son environnement bruité en ajustant sa voix par rapport au bruit de fond afin de mieux se faire comprendre. Cet effet est connu sous le nom de l'effet Lombard et peut se traduire par une augmentation du centre de gravité spectral de la voix, en plus d'une augmentation globale du niveau sonore de la voix (Garnier [4]).

Le timbre de la voix peut parfois influencer la clarté de la voix, le naturel de la voix et l'évaluation de la qualité vocale.

Dans le cas de la connaissance préalable du locuteur, l'auditeur a une référence interne de la voix et peut juger de la dégradation causée par le timbre. Cependant, dans l'absolu, il est parfois difficile de caractériser une voix claire ou sourde sans référence. Il en est de même lors de la détermination d'indicateurs automatiques : la plupart des modèles intégrant la dimension coloration utilise la comparaison des indicateurs entre le signal de référence et le signal dégradé. Les indicateurs de la coloration les plus utilisés sont présentés par la suite.

Le centre de gravité spectral du signal de la parole représente le centre de gravité de la distribution d'énergie du spectre fréquentiel du signal. On peut le considérer comme le point d'équilibre du spectre. La hauteur spectrale donnée par la position du centre de gravité des composantes du spectre définit ce que l'on appelle la brillance du son. Le centre de gravité spectral peut être calculé de la manière suivante :

$$SC = \frac{\sum_k f_k a_k}{\sum_k a_k}, \qquad \textbf{Eq. I.9}$$

où a_k correspond à l'amplitude de la composante spectrale de fréquence f_k.

La brillance perceptive "Sharpness" définie par Aures est un indicateur perceptif équivalent à l'indicateur centre de gravité spectral. Il est calculé à partir de la sonie spécifique par bande de bark définie par Zwicker [38] :

$$A = 0{,}11\frac{\sum_{z=1}^{nband} z \cdot g(z) \cdot N'(z)}{N}, \qquad \textbf{Eq. I.10}$$

avec z l'index de la bande de bark considérée, *N'(z)* la sonie spécifique de la z^iéme^ bande de bark, *N* la sonie totale déterminée comme la somme des sonies spécifiques sur toutes les bandes de bark et *g(z)* la fonction déterminée par

$$\begin{cases} g(z) = 1 & si\ z < 15 \\ g(z) = 0{,}066 \cdot \exp(0{,}171 \cdot z) & si\ z > 15 \end{cases} \qquad \textbf{Eq. I.11}$$

Le point spectral de coupure ou "Spectral Rolloff Point" développé par Scheirer [67] est la position de l'échantillon fréquentiel en dessous duquel est contenu 95 % de la puissance du spectre.

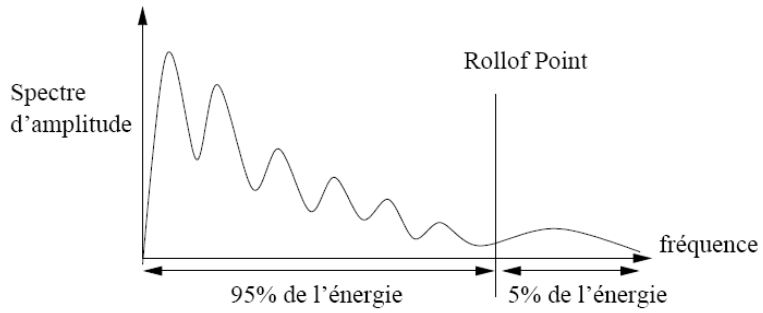


Fig. I.4 Définition du "Spectral Rolloff Point"

Le point spectral de coupure peut être défini pour chaque trame. L'indicateur utilisé est ensuite défini comme étant l'écart-type des composantes obtenues sur toutes les trames du signal.

La pente de l'enveloppe du spectre fréquentiel peut aussi représenter la dimension coloration.

I.4.10. Synthèse des dimensions perceptives

Le Tab. I.6 synthétise les espaces perceptifs relatifs à l'évaluation de la qualité vocale, selon différents auteurs (Mc Gee [21], Mc Dermott [27], Gabrielsson et Sjogren [13], Bappert and Blauert [22], Petersen et Hansen [16], Hall [17], Mattila [19, 51], Bernex et Barriac [29], Wältermann *et al.* [14, 18], Etame [34]). L'ordre d'importance des dimensions de chaque espace perceptif est obtenu par les analyses statistiques (MDS ou ACP). Il révèle l'importance des dimensions par rapport aux différences de sonorité entre les stimuli, et non par rapport au poids accordé lors de l'évaluation de la qualité vocale. Cet ordre dépend donc principalement du poids des dégradations physiques imposées aux stimuli.

Dimensions / Auteurs	Conditions & nb de stimuli	Méthode	Bruyance	Bruit_sur_la_parole	Continuité	Naturel	Clarté	Sifflement, bulleux	Sonie	Coloration	Basse fréquence	Haute fréquence	Distorsion	Précision, finesse
McGee (1965)	Filtre (15)	SA									X_1	X_1		
McDermott (1969)	RTC (22)	SA					X_1		X_3				X_2	
Gabrielsson (1979)	Haut-parleur (9)	AA	X_4				X_1			X_3				X_2
Bappert (1994)	Codecs NB (12)	AA					X_1			X_2				
Petersen (1997)	Codecs NB (16)	AA			X_1			X_3			X_4	X_2		
Hall (2001)	Codecs NB (10)	SA	X_2			X_1					X_3			
Mattila (2002a)	GSM (41)	SA	X_5		X_3	X_1		X_4		X_2				
Mattila (2002b)	GSM & Bruit (41)	SA	X_4		X_3	X_1				X_2				
Bernex (2002)	VoIP (ppl) (6)	SA	X_1		X_2	X_3								
Wältemann (2006)	VoIP main-libre (14)	SA	X_4		X_2		X_1				X_3			
Etame (2007)	Codecs NB & WB (20)	SA	X_2	X_3				X_4		X_1				
Wältermann (2008)	VoIP&GSM & bruit (50)	SA	X_2		X_1					X_3				
Nombre de récurrences			**7**	**1**	**5**	**3**	**4**	**3**	**1**	**12**			**1**	**1**

Tab. I.6 Synthèse des attributs perceptifs représentant les dimensions des espaces perceptifs obtenus par différents auteurs et leur ordre d'importance par rapport aux différences de sonorité. SA→ méthode sans *a priori* / AA → méthode avec *a priori*

Les attributs les plus utilisés sont **bruyance**, **clarté**, **continuité** ainsi que **coloration**. Ainsi, ces attributs seraient les plus pertinents pour expliquer les différences de sonorité du signal de la parole. On remarque que l'espace obtenu par Wältermann *et al.* [18] utilise ces quatre dimensions en considérant que coloration et contenu basse fréquence sont des attributs similaires. Il est cependant délicat de fixer de telles conclusions à cause de la variabilité des résultats obtenus selon le type de transmission testée ou bien la méthode utilisée.

Certains des attributs présentés dans cette partie révèlent certaines similarités sensorielles (p. ex. naturel, clarté, distorsion). L'attribut distorsion semblerait être pertinent à la différenciation des sonorités, cependant il a été remarqué qu'il regroupe de nombreux attributs, le rendant imprécis et difficile à interpréter.
Les dimensions perceptives peuvent être décomposées en d'autres sous-dimensions plus précises comme dans le cas des dimensions bruyance et continuité (Huo [54, 68]). Dans certains cas, il est possible qu'une dimension perceptive regroupe une combinaison d'attributs ou alors qu'elle utilise plusieurs attributs distincts suivant les extrémités de la dimension. Parfois, l'identification d'une dimension par un attribut n'est tout simplement pas possible.

Les espaces perceptifs doivent être déterminés en rapport avec leurs domaines d'application. Plus le domaine d'application est ciblé, plus l'espace perceptif correspondant sera précis. Un espace perceptif pertinent pour l'évaluation de la qualité vocale est déterminé par le bon compromis entre la précision de l'espace perceptif et le domaine d'application le plus global possible.

Il existe une grande variabilité des espaces perceptifs suivant le moyen de diffusion, le choix des sujets (expert ou naïf), le contenu des échantillons sonores (musique, parole ou bruit) et la méthode de détermination de l'espace perceptif (Gabrielsson et Sjogren [13]). D'après les différentes études référencées dans cette partie, il est judicieux de déterminer l'espace perceptif par la méthode sans *a priori*, en préférant l'analyse multidimensionnelle des dissimilarités.

Suivant le domaine d'application et les besoins de l'expérimentateur, les dimensions perceptives peuvent être identifiées par un test de verbalisation et/ou par un test de sémantique différentielle. Ces tests peuvent être utilisés en complément de la détermination de l'espace perceptif afin d'approfondir et de clarifier la signification des dimensions par les attributs correspondants.

I.5. Modèles objectifs d'évaluation de la qualité vocale

Les modèles objectifs d'évaluation de la qualité vocale les plus performants et les plus utilisés sont normalisés par l'Union Internationale des Télécommunications (UIT). Cet organisme est chargé de la normalisation et de la planification des télécommunications dans le monde. Elle établit les normes de ce secteur et diffuse toutes les informations techniques nécessaires pour permettre l'exploitation des services mondiaux de télécommunications. D'autres modèles sont aussi proposés par des laboratoires universitaires ou encore par des sociétés privées, mais ils ne sont pas validés par les normes de l'UIT.

Ces modèles objectifs sont classés selon trois critères fondamentaux (cf. Fig. I.5 d'après Guéguin [69]) :

- le contexte d'évaluation de la qualité vocale (**écoute, locution ou conversation**),
- le type d'indicateurs utilisés lors de la modélisation (**paramétriques** ou **basés sur le signal**)
- le type de mesure (**avec ou sans référence** dans le cas d'indicateurs basés sur le signal ou encore **mono-extrémité ou de bout en bout** dans le cas d'indicateurs paramétriques).

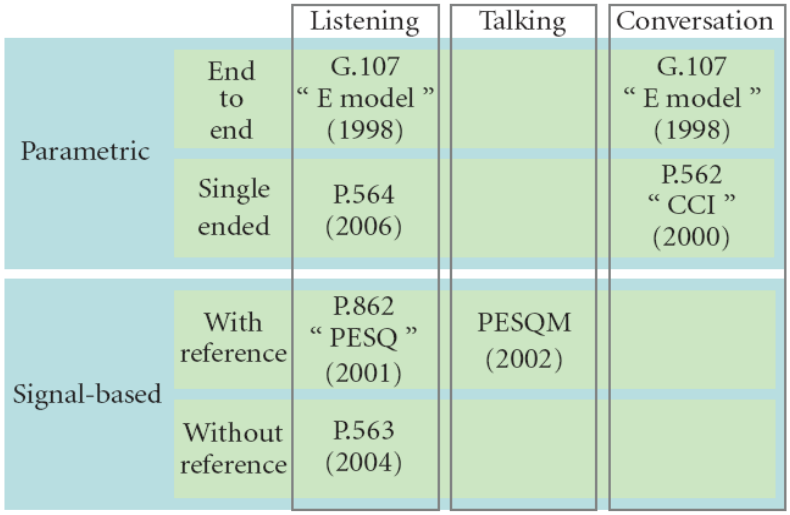


Fig. I.5 Présentation des différents modèles existants d'évaluation de la qualité vocale, selon Guéguin [69]

I.5.1. Les trois contextes d'évaluation de la qualité vocale

Lors d'une télécommunication, l'évaluation de la qualité vocale peut être réalisée suivant trois contextes différents : le contexte d'écoute, le contexte de locution et le contexte de conversation.

Le contexte d'écoute représente la configuration où un utilisateur écoute un message parlé, puis donne son jugement sur la qualité vocale. Dans ce contexte, seule la transmission allant du locuteur vers l'auditeur est évaluée. De ce fait, certaines dégradations liées au temps de transmission de l'information de la parole ne sont pas prises en compte, comme le délai de la transmission ou encore l'écho de la voix d'un locuteur. Les tests subjectifs décrits dans la partie I.2 sont bien adaptés à ce contexte, et notamment le test ACR (Absolute Category Rating) qui est souvent utilisé. La majorité des modèles existants utilisent ce contexte d'écoute, comme les modèles PESQ [1], P.563 [2], et, dans certains modes d'application, le modèle E G.107[8] [3].

Le contexte de locution représente la configuration où un locuteur évalue la qualité vocale de sa propre voix. Dans ce cas, les principales dégradations testées sont l'écho de la voix et l'effet local du terminal. Un test subjectif correspondant à ce contexte est décrit par Appel [70]. Il consiste à demander aux sujets de prononcer une phrase et d'évaluer simultanément la qualité vocale du retour de leur propre voix rediffusée par le haut-parleur du même terminal en utilisant le test DCR décrit dans la partie I.2.

Le modèle PESQM développé par Appel et Beerends [70] correspond à ce contexte d'évaluation de la qualité vocale. Cependant, il n'a pas été normalisé par l'UIT.

Le contexte de conversation allie les deux configurations d'écoute et de locution en considérant la totalité de la transmission. Les interactions entre deux interlocuteurs sont alors prises en compte lors de l'évaluation de la qualité vocale, ce qui représente un contexte plus réaliste mais plus complexe à déterminer (Guéguin [71]). La mesure subjective de la qualité vocale dans un tel contexte peut être réalisée par une expérience détaillée dans Guéguin [69]. Deux participants sont placés dans deux salles différentes et ont une conversation à travers un

[8] Le modèle E est un modèle de planification de réseau dans un contexte de conversation, cependant en fixant certains paramètres non disponibles dans ce contexte (le délai, écho), il est souvent utilisé en contexte d'écoute.

système de télécommunication présentant des dégradations présentes en contexte d'écoute, ainsi que des dégradations présentes en contexte de locution. Le test consiste à simuler une conversation téléphonique réelle en proposant aux sujets différents scénarios. Le jugement se fait alors en trois étapes : les deux sujets (deux interlocuteurs) évaluent tout d'abord la qualité globale de la conversation. Puis, dans une deuxième étape, l'un des sujets (le locuteur) prononce une phrase au cours de laquelle il juge la qualité vocale en condition de locution, tandis que l'autre sujet (auditeur) juge la qualité en condition d'écoute. La troisième étape est la réciproque de la deuxième étape. Les jugements sont effectués par le test ACR détaillé dans le paragraphe I.2. Ce type de test est très coûteux et très lourd à réaliser et n'est donc pas souvent effectué.

Les modèles d'écoute PESQ [1] et de locution PESQM [70] sont combinés en intégrant en plus la mesure du délai afin de proposer un modèle d'évaluation de la qualité vocale en contexte de conversation. D'autre part, les modèles normalisés par l'UIT sont le modèle E [3] et le modèle "Call Clarity Index" (indice de netteté des logatomes) P.562 [72]. Le modèle E [3] a été construit pour planifier une transmission téléphonique de bout en bout, en contexte de conversation.

I.5.2. Modèles utilisant des indicateurs paramétriques

Les modèles d'évaluation de la qualité vocale peuvent utiliser **des indicateurs paramétriques** afin de déterminer la note globale de qualité vocale. Ce type d'indicateur est basé sur des informations issues des statistiques du réseau ou issues d'indicateurs calculés par certains DSP (Digital Signal Processor) et transmis à travers le réseau[9], comme le type de codec, le délai qui correspond au temps de propagation dans le réseau, le niveau du bruit de fond (issu de certaines DSP), l'écho (via les informations remontées par les annuleurs d'écho) ou encore le taux de pertes de paquets. Il existe deux types de modèles utilisant des indicateurs paramétriques : les modèles qualifiés de "bout en bout" prennent en compte la totalité de la transmission (de la bouche du locuteur jusqu'aux oreilles de l'auditeur), alors que les modèles mono-extrémités utilisent seulement les informations disponibles en un point du réseau.

I.5.2.1. Modèle de bout en bout

Le modèle E normalisé sous la référence G.107 [3] utilise des indicateurs paramétriques afin de planifier et de prévoir la qualité vocale de la transmission de bout en bout à l'aide de 18 indicateurs.

Le cœur du modèle est fabriqué à partir de données psychoacoustiques issues des résultats de tests subjectifs. Le résultat brut du modèle appelé *"facteur d'évaluation de l'indice de transmission"* est appelé *R* ou plus communément *la note R*. Elle est obtenue grâce à un modèle additif constitué de cinq facteurs de dégradation définis par la relation suivante :

$$R = R_0 - I_s - I_d - I_e + A \qquad \textbf{Eq. I.12}$$

R_0 représente le rapport signal sur bruit (RSB) et tient aussi compte de l'effet du type de bruit de fond, à savoir si la source de bruit provient de l'environnement du locuteur et de l'auditeur ou bien s'il s'agit d'un bruit généré par le réseau. Le facteur I_s est une combinaison de toutes les dégradations présentes sur le signal vocal (distorsion de la quantification). Le facteur I_d rassemble les dégradations liées au temps de propagation (délai, écho). Le facteur I_e correspond aux dégradations générées par le codec utilisé. Enfin, le facteur d'avantage *A* permet de

[9] Par exemple dans les paquets RTPC XR (Real-Time Control Protocol Extended Reports)

représenter l'indulgence des utilisateurs face à l'évaluation de la qualité vocale dans certaines conditions, telles que le système de communication utilisé (système filaire, mobile, le terminal utilisé, l'emploi du kit main libre ou le combiné).

Chacun de ces cinq facteurs de dégradation est une combinaison de plusieurs indicateurs physiques et/ou psychoacoustiques disponibles dans G.107 [3].

La note *R* est alors exprimée comme un scalaire, pouvant prendre les valeurs de 0 à 100, avec *R = 0* représentant une qualité extrêmement mauvaise, et *R = 100* représentant une qualité extrêmement bonne.

Une relation entre cette note R et la note d'opinion moyenne (note MOS sur une échelle allant de 1 à 5) est définie à partir de résultats de tests subjectifs par l'expression suivante :

$$\begin{cases} R < 0: & MOS = 1 \\ 0 < R < 100: & MOS = 1 + 0{,}035 \cdot R + R \cdot (R - 60) \cdot (100 - R) \cdot 7 \cdot 10^{-6} \\ R > 100: & MOS = 4{,}5 \end{cases} \quad \textbf{Eq. I.13}$$

Le modèle E est utilisé par de nombreux outils car il est libre de droit. Le modèle E a été adapté en 2006 à des transmissions large bande (50 Hz - 7 kHz). L'échelle de la note R a été allongée à une valeur maximale de R = 129 en large bande contre une valeur max de R = 100 en bande étroite (Raake [8]).

I.5.2.2. Modèles mono extrémité

Le modèle CCI (Call Clarity Index ou indice de netteté des logatomes) est un modèle en contexte de conversation (UIT P.562 [72]). Il est équivalent au modèle E mais il est applicable en un point (mono-extrémité). Il utilise les informations issues du dispositif INMD (In-service Non-intrusive Measurement Devices) (UIT P.561 [73]). Ce dispositif consiste à obtenir les informations comme le niveau vocal, le niveau de bruit, l'affaiblissement d'écho, le temps de propagation sur le trajet d'écho (retard d'extrémité) ainsi que la variation du temps de propagation des paquets IP et le taux de perte de paquets IP.

La recommandation P.564 [74] spécifie les critères des modèles objectifs d'évaluation de la qualité vocale ayant les caractéristiques suivantes :

- Application à des flux RTP de voix sur IP,
- Bande étroite (300 Hz – 3,4 kHz) et large bande (50 Hz – 7 kHz),
- Contexte d'écoute,
- Mono extrémité (situé au niveau de l'auditeur),
- Indicateurs paramétriques.

Le modèle "PsyVoIP" développé par Psytechnics et le modèle VQMon développé par Telchemy sont conformes à cette norme. Ces modèles sont principalement utilisés pour le contrôle de la qualité des transmissions. Ils prédisent la note de qualité vocale sur l'échelle MOS (UIT P.800 [11]).

Ces modèles prennent en compte le taux de pertes de paquets, les informations contenues dans l'en-tête des paquets des réseaux RTP, UDP et IP (codec, pourcentage de perte de paquets, burst ou aléatoire, taille des paquets, silences, type de PLC utilisé), et le délai ainsi que la variation du délai (gigue).

Lijing Ding [56] propose un modèle sans référence d'évaluation de la qualité vocale, appliqué à une transmission VoIP, en considérant trois principaux types de dégradations : les pertes de paquets, les clips temporels et le bruit de fond. Ce modèle est basé sur une combi-

naison non-linéaire des trois types de dégradations. Il utilise, entre autres, les indicateurs du modèle E.

Les modèles paramétriques ont l'avantage d'être très rapides du fait qu'ils ne nécessitent que très peu de ressources CPU (Central Processing Unit). Ils peuvent donc être embarqués dans les terminaux ou au cœur du réseau pour donner une note de qualité vocale en temps réel. Cependant, ces indicateurs paramétriques ne sont pas disponibles pour tous les types de transmission téléphonique car les en-têtes des paquets contiennent uniquement les informations des dégradations survenues lors de la transmission à travers le dernier réseau. Il est parfois impossible de savoir quels chemins ont réellement emprunté les paquets et donc, impossible de savoir si le signal correspondant n'a pas subi d'autres dégradations dans des réseaux précédents.

Les modèles utilisant des indicateurs basés sur le signal permettent d'obtenir des résultats plus performants en considérant la totalité de la transmission testée.

I.5.3. Modèles utilisant des indicateurs basés sur le signal

Les indicateurs **basés sur le signal** utilisent les informations numériques du signal vocal pour déduire la note de qualité vocale. Ce type de modèle peut être avec ou sans signal de référence. Dans le cas de l'utilisation d'un signal de référence correspondant au signal original non dégradé situé au niveau de la bouche du locuteur, le modèle est dit intrusif et la note de qualité est obtenue en comparant les deux signaux (le signal de référence et le signal passé à travers le système). Le modèle PESQ [1] utilise ce type d'indicateur. Il est considéré comme le modèle de référence grâce à son efficacité à prédire les notes de qualité vocale.

Dans le cas où le modèle est sans référence (non-intrusif), seul le signal dégradé est utilisé à la prédiction de la note de qualité du signal vocal. Le modèle sans référence normalisé par l'UIT est présenté dans la recommandation P.563 [75].

I.5.3.1. Modèles avec référence

Le modèle PESQ [1] (Perceptual Evaluation of Speech Quality) est l'instrument actuel le plus utilisé et le plus précis pour prédire la note de qualité vocale lors d'une télécommunication soumise à différents types de dégradations comme le codage, le bruit de fond, la présence de pertes de paquets, et la distorsion. Il est utilisé en contexte d'écoute à partir d'indicateurs basés sur le signal.

Le modèle PESQ utilise une représentation perceptive des signaux dégradés et non dégradés. Les caractéristiques des signaux représentées par la fréquence en Hz, et l'intensité en dB du signal sont transformées en caractéristiques perceptives, respectivement la fréquence perçue en barks, et la sonie en sones. Ces deux descripteurs psychoacoustiques sont représentés par la densité de sonie, exprimée en sone /Bark. Un alignement temporel est ensuite réalisé entre les deux signaux avant d'évaluer les différences de représentation perceptive qui déterminent les différences audibles grâce à un modèle cognitif. La note de qualité vocale est alors prédite sur l'échelle MOS [11] à partir de ces différences audibles.

Le modèle PESQ ne prend pas en compte les dégradations liées au délai, au time warping, ni à l'intensité acoustique de la parole qui est considérée à un niveau moyen optimal de 79 dB SPL en écoute monaurale (UIT "Handbook on Telephonometry" [65]).

Des nouveaux modèles d'évaluation de la qualité vocale avec référence en contexte d'écoute utilisant des indicateurs basés sur le signal sont en cours de compétition à l'UIT afin de remplacer le modèle PESQ.

Le modèle PESQM [70] (Perceptual Echo and Sidetone Quality Measure) est basé sur les mêmes techniques que le modèle perceptif PESQ, pour un contexte de locution. Ce modèle de locution est basé sur la comparaison entre les signaux dégradé et non-dégradé. Le signal non-dégradé correspond au signal prononcé par le locuteur qui est capté par le microphone du terminal, tandis que le signal dégradé correspond à celui qui est diffusé dans le haut-parleur de ce même terminal. Les dégradations prises en compte sont l'écho de la parole à travers la transmission et l'effet local[10] généré par le terminal (*sidetone* en anglais). Möller [5] décrit l'effet local comme la somme de trois chemins de propagation: le chemin acoustique, mécanique et électrique. Le chemin acoustique correspond à la transmission aérienne de la parole de la bouche du locuteur vers ses deux oreilles, le chemin mécanique correspond à la propagation de la parole à travers les os de la tête jusqu'aux oreilles (moyennes et internes) du locuteur, et le chemin électrique correspond à la parole enregistrée par le microphone du terminal et rediffusée par le haut-parleur de ce même terminal. L'effet local de type mécanique domine l'effet local électrique pour les fréquences basses comprises entre 600 et 800 Hz. Pour des fréquences supérieures, l'effet local électrique est important pour préserver de bonnes conditions de communication, et notamment en présence de bruit de fond ambiant important.

Le modèle TOSQA (Telecommunications Objective Speech Quality Assessment) développé par J. Berger [76] utilise la comparaison entre le signal dégradé et le signal de référence. Ce modèle présente la particularité de pouvoir être utilisable pour des télécommunications en bande étroite et en bande élargie. De plus, la prédiction de la qualité vocale permet de s'adapter au type de la mesure, à savoir s'il s'agit d'une mesure électrique ou acoustique. Un choix est aussi à faire sur le système de restitution utilisé, à savoir s'il s'agit d'un combiné de haute qualité monaurale, ou d'un combiné de qualité moyenne monaurale, ou encore un casque acoustique diffusant le son en diotique.

Le modèle TOSQA propose plusieurs informations à l'issue des résultats : la valeur d'évaluation de la qualité vocale "TMOS", le facteur de dégradation *Ie* correspondant à la dégradation relative au codage de la parole et le délai de la télécommunication.

Les modèles utilisant un signal de référence nécessitent beaucoup de ressources CPU et une mise en œuvre complexe. De plus, le signal de référence n'est pas connu lors de l'évaluation d'une télécommunication réelle. Ce type de modèle ne peut être appliqué en temps réel, cependant il est très performant pour évaluer la qualité vocale d'une transmission simulée, et très utilisé dans ce domaine.

I.5.3.2. Modèle sans référence

Le modèle P.563 [2] fait l'objet d'une recommandation à l'UIT. Il est proposé par Malfait [75] afin d'évaluer la qualité vocale dans un contexte d'écoute à partir d'indicateurs basés sur le signal, en utilisant seulement le signal dégradé (modèle sans référence). Il prend en compte de nombreux types de dégradations, comme ceux générés par les algorithmes d'annuleurs d'écho ou de réduction de bruits, les pertes de paquets, le type de codec, le transcodage et les bruits de fond présents sur le signal de la parole.

Tout d'abord, un prétraitement est réalisé sur le signal dégradé : un algorithme de détection d'activité vocale (DAV) est utilisé pour repérer les zones actives du signal vocal. Ces zones de parole sont alors ajustées à un niveau de -26 dBov dans le but de s'affranchir de l'effet du niveau de la parole et des problèmes de dynamique du signal vocal. Ensuite, le signal

[10] L'effet local (sidetone) est la diffusion du signal vocal capté par le microphone d'un terminal au haut-parleur de ce même terminal. Il est utilisé afin d'améliorer la qualité de la transmission vocale, afin que le locuteur puisse entendre sa propre voix.

est soumis à une paramétrisation pouvant être décrite par trois blocs fonctionnels indépendants correspondant aux principales classes de distorsion :

L'analyse du naturel de la voix est réalisée à partir de la reconstruction de la parole par les coefficients LPC afin de déterminer le type de voix (masculine, féminine ou fortement robotisée).

L'analyse des bruits additionnels intenses est composée de la mesure de l'intensité sonore du bruit de fond présent sur le signal vocal. Il est exprimé par le rapport signal sur bruit avec le niveau de la parole fixé à -26 dBov.

L'analyse des interruptions, des silences et de l'écrêtage temporel consiste à repérer et à mesurer les chutes de niveau sonore dans le signal.

D'autre part, un ensemble de descripteurs vocaux élémentaires comme le niveau vocal actif, l'activité vocale et les variations de niveau sont utilisés principalement pour l'algorithme de DAV ainsi que pour la phase de prétraitement.

Le modèle de qualité vocale réalise une construction artificielle du signal "original" en améliorant la qualité vocale du signal dégradé à partir d'une approche utilisant les LPC. Ensuite, trois facteurs sont déterminés à partir de la comparaison entre le signal dégradé et le signal amélioré (de la même manière que les modèles avec référence). La note de qualité vocale est alors obtenue comme une régression linéaire de ces trois facteurs.

Ce modèle est cependant très contesté et contestable par rapport à ses performances d'évaluation de la qualité vocale, et de ce fait très peu utilisé en opérationnel.

Le modèle ANIQUE (An Auditory Model for Single-Ended Speech Quality Estimation) est proposé par Kim [77] afin d'évaluer la qualité vocale dans un contexte d'écoute à partir d'indicateurs basés sur le signal en utilisant seulement le signal dégradé (modèle sans référence). Le modèle ANIQUE utilise une représentation perceptive des signaux dégradés et les caractéristiques d'articulation de la voix. Le signal dégradé est représenté par des enveloppes temporelles suivant différentes bandes de fréquences exprimées par l'échelle "Equivalent Rectangular Bandwidth" (ERB). Ces différentes enveloppes temporelles sont alors utilisées pour exprimer le rapport "articulation-to-nonarticulation ratio" (ANR) à la base de l'estimation de la qualité vocale. Ce modèle obtient des performances moins précises que le modèle P.563 et il n'est pas normalisé.

Falk [50] propose aussi un modèle d'évaluation de la qualité vocale sans référence utilisant des indicateurs basés sur le signal. Il utilise un modèle de mélange de gaussiennes afin de reconstruire le signal de référence à partir du seul signal dégradé. La comparaison des deux signaux permet ensuite de prédire la note de qualité vocale.

Grancharov [78] a développé le modèle **LCQA** (Low Complexity Quality Assesment) afin d'évaluer la qualité vocale d'une télécommunication soumise à des dégradations générées par certains codecs. Le modèle LCQA utilise une représentation perceptive du signal dégradé (modèle non-intrusif) à partir de 11 indicateurs basés sur le signal. Ces indicateurs sont ensuite reliés entre eux afin de déterminer la qualité vocale, grâce au modèle de mélange de gaussiennes (GMM "Gaussian Mixture Model").

Raja [79] propose une amélioration du modèle P.563 en utilisant l'outil "Genetics Programming" qui consiste à déterminer la meilleure combinaison des indicateurs en testant toutes sortes d'interactions et différentes fonctions logarithmiques, exponentielles ou polynomiales, dans le but de prédire la qualité vocale.

Les différents modèles exposés dans cette partie obtiennent des performances variables suivant le type de dégradation testé et suivant le domaine d'application. Les modèles utilisant

des indicateurs basés sur le signal sont en général plus performants que les modèles paramétriques. Cependant, les modèles basés sur le signal sont souvent plus lourds à mettre en œuvre et ne peuvent généralement pas être utilisés pour le contrôle en temps réel des télécommunications, contrairement aux modèles paramétriques.

Chapitre II. Positionnement du modèle DESQHI

Le chapitre précédent a permis de faire l'état de l'art des différents types de modèles existants d'évaluation de la qualité vocale, et de montrer leurs limites d'application suivant leurs principales caractéristiques (cf. §.I.5). L'objectif de ce deuxième chapitre est de présenter et de justifier les choix retenus pour la construction du modèle DESQHI "Diagnostic and Evaluation of Speech Quality using Hybrid Indicators". Ils concernent **la structure du cœur du modèle**, **le type d'indicateur**, le type de modèle **intrusif ou non-intrusif,** et le **contexte de la mesure**.

Le diagnostic de la qualité vocale est possible en étudiant les dimensions perceptives mises en cause lors de l'évaluation de la qualité vocale. Pour cela, le modèle proposé est constitué d'un **cœur multidimensionnel** permettant de distinguer les dégradations perceptives de la qualité vocale (cf. Fig. II.1). Cela permet de prédire la qualité vocale globale, mais aussi de déterminer quel(s) attribut(s) perceptif(s) est/sont responsable(s) de la baisse de la qualité vocale.

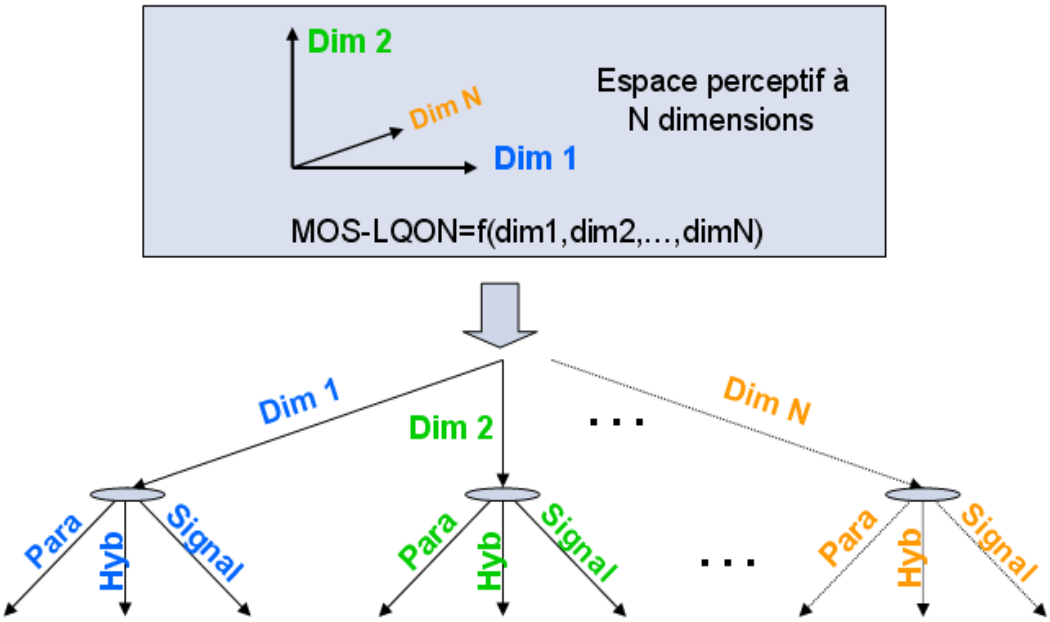


Fig. II.1 Structure globale du modèle DESQHI

Les dimensions perceptives correspondant aux nombreuses dégradations physiques générées par les techniques de télécommunication actuelles (RTC, RNIS, GSM, VoIP) ne sont pas clairement déterminées. La principale raison est que les attributs proposés sont nombreux et que leurs significations ne sont pas toujours assez précises. Certains attributs correspondent parfois à des caractéristiques sensorielles similaires (cf. §.I.4). Nous avons donc choisi de déterminer l'espace perceptif par la méthode sans *a priori*, à partir des mesures de dissimilarité entre les stimuli (cf. §.I.3.2).

Les modèles existants d'évaluation de la qualité vocale utilisent soit des indicateurs basés sur le signal numérique, soit des indicateurs paramétriques basés sur les statistiques du réseau (cf. §.I.5). Le travail présenté dans cette thèse vise, entre autres, à vérifier qu'un modèle hybride utilisant ces deux types d'indicateurs de manière simultanée permet de bénéficier des avantages du faible coût des indicateurs paramétriques pour une application en temps réel, tout en conservant de bonnes performances grâce aux indicateurs basés sur le signal.

Ainsi, en nous basant sur un cœur de modèle multidimensionnel, chacune des dimensions perceptives est représentée par des indicateurs qui sont ou bien paramétriques, ou bien

basés sur le signal ou encore hybrides (cf. Fig. II.1). Le modèle ainsi réalisé est adaptatif en proposant à l'expérimentateur de choisir le type d'indicateur pour chacune des dimensions, suivant les informations disponibles au point de la mesure ou encore suivant le domaine d'application (contrôle différé ou temps réel, planification).

Les modèles intrusifs d'évaluation de la qualité vocale ont fait l'objet de nombreuses études et leurs performances sont de plus en plus efficaces (p. ex. PESQ [1]). Le modèle DIAL de Côté [10] est actuellement le seul qui permet de diagnostiquer et d'évaluer la qualité vocale de manière intrusive. Cependant, ces modèles ne reflètent pas le contexte de l'évaluation subjective absolue d'une transmission téléphonique. Dans la plupart des cas, le signal de référence n'est pas connu *a priori* par l'auditeur qui reconstruit ce signal en fonction de la connaissance préalable du locuteur, du type de service utilisé et de ses habitudes. Le modèle non-intrusif ou mono-extrémité est donc plus représentatif d'une situation dans laquelle un auditeur évalue la qualité vocale d'une transmission téléphonique.

Les modèles non-intrusifs (p. ex. P.563 [2]) ne sont pas encore très performants, ne proposent pas de diagnostic et restent moins nombreux que les modèles intrusifs. Le modèle paramétrique mono-extrémité (P.564 [74]) est quant à lui trop axé sur les dégradations du réseau et ne prend en compte que certains types de dégradations. Il est donc nécessaire de proposer de nouveaux indicateurs afin d'améliorer les performances des modèles non-intrusifs existants et de proposer un diagnostic de la qualité vocale.

Par ailleurs, la prédiction de la qualité vocale en temps réel et à n'importe quel endroit de la transmission nécessite de réaliser un modèle non-intrusif car le signal de référence ne peut être disponible à n'importe quel point de la mesure. C'est pourquoi, nous avons choisi de traiter le cas d'un modèle hybride non-intrusif qui utilise uniquement les informations disponibles au point de la mesure (signal dégradé et/ou informations issues du réseau).

Le développement des techniques de télécommunication permet aujourd'hui de transmettre les informations sur les réseaux avec un débit plus important. Cela permet d'envoyer plus d'information vocale en élargissant la bande passante des fréquences transmises. La bande passante utilisée jusqu'à présent est appelée communément la bande étroite et correspond aux fréquences comprises entre 300 Hz et 3,4 kHz, tandis que la bande élargie correspond aux fréquences comprises entre 50 Hz et 8 kHz. Les modèles existants d'évaluation de la qualité vocale ainsi que les bases de données associées aux jugements subjectifs sont essentiellement disponibles en bande étroite. Afin de pouvoir utiliser ces bases de données et de valider l'intérêt d'un modèle hybride par rapport aux modèles existants, le modèle proposé est construit pour une transmission en bande étroite.

Le cœur multidimensionnel du modèle DESQHI présenté dans le Chapitre III est construit à partir de deux principales expériences réalisées par un certain nombre de sujets. L'une concerne la détermination de l'espace perceptif correspondant aux dissimilarités entre les stimuli, l'autre fournit les mesures d'évaluation de la qualité vocale pour ces mêmes stimuli. La combinaison reliant l'évaluation de la qualité vocale et l'espace perceptif constitue le cœur du modèle DESQHI.

Les Chapitre IV et Chapitre V présentent ensuite la modélisation de chacune des dimensions constituant l'espace perceptif par des indicateurs paramétriques, basés sur le signal et hybrides, afin de proposer un modèle adaptatif au domaine d'application de la mesure.

Enfin, le Chapitre VI dévoile les performances du modèle DESQHI relatives à l'évaluation de la qualité globale et à la précision du diagnostic en employant la base sonore utilisée lors de la construction du cœur du modèle, et sept autres bases sonores inconnues du modèle.

Chapitre III. Cœur du modèle : Analyse multidimensionnelle de la qualité vocale

Le cœur du modèle DESQHI est construit à partir d'une approche multidimensionnelle dans le but d'évaluer la qualité vocale globale. Cette méthode se révèle aussi très utile pour réaliser un diagnostic de la qualité vocale en identifiant les attributs perceptifs responsables de la baisse de la qualité vocale. La Fig. III.1 illustre chacune des étapes de la réalisation du cœur du modèle.

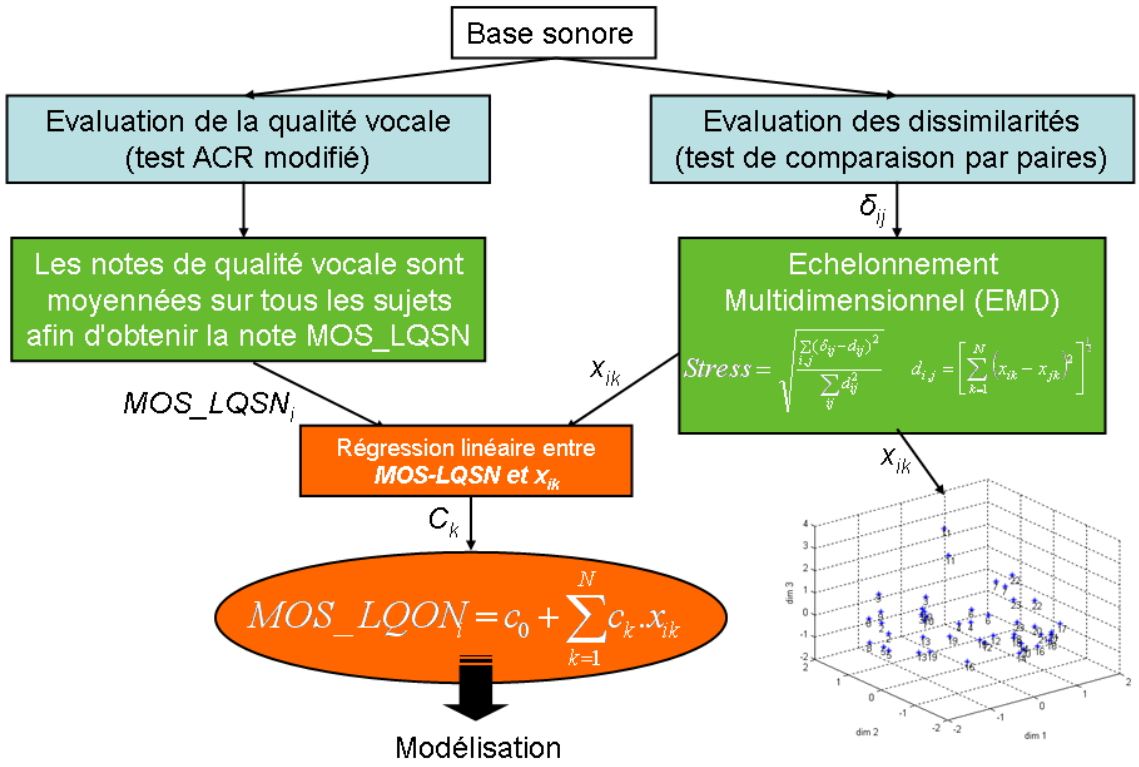


Fig. III.1 Les étapes de la construction du cœur du modèle

La base sonore est constituée de conditions de dégradation représentatives d'une transmission téléphonique réaliste (RTC, RNIS, GSM, VoIP). Le cœur du modèle d'évaluation de la qualité vocale est construit à partir des résultats perceptifs issus de deux expériences. La première consiste à évaluer la qualité vocale des stimuli de manière unidimensionnelle, en demandant directement l'avis des sujets (cf. §.I.2). La deuxième expérience est réalisée à partir de la même base sonore afin de déterminer l'espace perceptif représentatif des différences de sonorité.

Des analyses statistiques sont ensuite réalisées afin de relier les jugements de qualité vocale aux coordonnées de l'espace perceptif pour l'ensemble des stimuli. Les résultats de ces analyses permettent alors de vérifier si l'évaluation de la qualité vocale peut bien être représentée par une approche multidimensionnelle en combinant les différentes dimensions de l'espace perceptif. Dans le cas où cette hypothèse est vérifiée, la combinaison des dimensions perceptives représente alors le cœur du modèle DESQHI.

III.1. Tests subjectifs

Deux expériences ont été réalisées afin de construire le cœur du modèle d'évaluation de la qualité vocale. La première est un test d'évaluation de la qualité vocale par la méthode ACR (Absolute Category Rating) et la deuxième est un test d'évaluation des dissimilarités entre tous les stimuli deux à deux. Ces deux tests ont été réalisés l'un à la suite de l'autre à partir des

mêmes stimuli qui sont diffusés avec le même système de restitution, dans les mêmes salles, suivant le jugement des mêmes sujets.

III.1.1. Stimuli

Chaque stimulus est constitué de deux phrases courtes séparées par environ une seconde de silence. Les signaux de parole présentés sont prononcés soit par une voix d'homme soit par une voix de femme. Ils proviennent d'une base sonore enregistrée par France Telecom.
La double phrase prononcée par la voix de femme est :
"…Nous lui coupons net ses effets … Pascal a un gros problème…".
Cette phrase a une durée totale de 4,6 secondes, avec 3,2 secondes de zones de parole.
Dans le cas de la voix d'homme, les deux phrases courtes utilisées sont :
"…Il s'est endetté pour construire un pavillon… D'ici à ce qu'il lui pousse des ailes…".
La durée totale de ce signal est de 4,8 secondes, avec 3,7 secondes de parole par stimulus.
Le niveau sonore des phrases est fixé à -26 dBov (dB relatif à la surcharge), afin d'obtenir une bonne dynamique de la voix (UIT P.56 [80]).

Les dégradations représentatives des différents systèmes de communication (RTC / RNIS / GSM / IP) ont été simulées et appliquées à ces deux doubles phrases. Elles correspondent aux **codages de la parole**, **aux pertes de paquets**, **aux erreurs de bits** et aux **bruits de fond**. Les conditions de dégradations appliquées sont identiques pour la phrase prononcée par le locuteur femme et la phrase prononcée par le locuteur homme.

La sélection des conditions de dégradations est primordiale car elle va définir le domaine d'application du modèle développé. Le poids appliqué à chacun des quatre types de dégradations physiques doit être du même ordre de grandeur en terme d'impact sur la qualité vocale, sous peine d'écraser l'influence de certaines dégradations par rapport à d'autres. Pour cela, les stimuli présentant des dégradations simples (sans combinaison) ont été construits afin d'obtenir un score de qualité vocale minimale de 3 sur l'échelle MOS, ce qui correspond à une qualité vocale qualifiée de "passable" (cf. Tab. I.1). Dans ce cas, l'estimation de la baisse de qualité vocale a été obtenue par le modèle PESQ [1] pour des raisons pratiques.

Le nombre de stimuli (c.-à-d. le nombre de dégradations possibles) a été limité à cause de la durée de l'expérience. Chaque stimulus a une durée d'environ 5 secondes. Afin d'obtenir un test réalisable, 23 conditions ont été retenues. Cela représente donc 253 paires de stimuli à diffuser aux sujets, équivalant à un test de 2 h en prenant en compte les temps de réponse, l'apprentissage, les temps de pause ainsi que les consignes.

Dans ce cas un plan d'expérience complet n'est pas envisageable surtout si l'on veut tester plusieurs niveaux de chacun des 4 types de dégradations. Il a donc été choisi de tester les dégradations sous forme de combinaisons uniques (par exemple en combinant 1% de perte de paquets avec un codec en G.711 et 2 % avec du G.729 mais pas 1% avec du G.729 et 2 % avec du G.711).

Parmi ces combinaisons uniques, il a été choisi de sélectionner uniquement les combinaisons réalistes. Par exemple, les dégradations générées par une transmission IP sont combinées entre elles avec des combinaisons de codecs (G.711 et G.729) associées à des pertes de paquets, tandis que les codecs utilisés pour les transmissions mobiles (GSM-EFR) sont combinés avec des erreurs de bits. Certaines combinaisons entre les systèmes de transmission ont aussi été représentées, par exemple en combinant des dégradations générées par des transmissions IP vers mobile, afin de simuler des communications entre des interlocuteurs utilisant différents moyens de communication.

Voici la description des quatre types de dégradations physiques qui ont été utilisés pour simuler des transmissions téléphoniques diverses et réalistes.

Les codecs ont été utilisés seuls ou en combinaisons. Les codecs sélectionnés prennent en entrée des signaux avec des bandes passantes audio étroites (4 kHz), avec une fréquence d'échantillonnage de 8 kHz et quantifiés sur 16 bits. Ils ont été sélectionnés pour leurs utilisations courantes. Les quatre codecs qui ont été retenus sont : G.711, G.729, G.726, et GSM-EFR.

- Le codec G.711 [81] est utilisé par le Réseau Téléphonique Commuté (RTC), le RNIS, et la VoIP. Ce codec est basé sur la technique PCM (Pulse Code Modulation) qui consiste à représenter chaque échantillon du signal par sa valeur codée sur 8 bits avec un pas de quantification dépendant du niveau. Ce codec ne dégrade pas beaucoup la qualité vocale car le signal est peu compressé, cependant il nécessite un débit assez élevé (64 kbits/s). La mise en œuvre de ce codec intègre la plupart du temps un algorithme de PLC qui limite la perception des discontinuités dans le signal. Ce codec est testé avec et sans PLC.
- Le codec G.729 [60] est utilisé pour les transmissions IP. Il est basé sur la technique CS-ACELP (Conjugate Structure Algebraic Code Excited Linear Prediction), à un débit de 8 kbits/s. Ce codec ne requiert que très peu de débit, et il est beaucoup utilisé pour la voix sur IP, cependant le signal de parole est assez compressé et génère des dégradations. Ce codec intègre un algorithme de PLC.
- Le codec G.726 [82] est utilisé en Europe pour les terminaux DECT (Digital Enhanced Cordless Telecommunications). Il utilise la technique ADPCM (Adaptive Differential Pulse Code Modulation). Il est employé à un débit de 32 kbit/s). Ce codec a l'inconvénient de générer du bruit sur la parole (cf. §.I.4.2).
- Le codec GSM-EFR (Enhanced Full-Rate) est utilisé pour les communications mobiles. Il utilise la technique ACELP (Algebraic Code Excited Linear Prediction). Il a un débit de 12,2 kbits/s.

Ces codecs ont aussi été présentés en duo afin de simuler le transcodage. Dans ce cas les dégradations causées par chacun des codecs se combinent. Dans le cas du transcodage composé de deux fois le codec G.729, la note MOS-LQON minimale calculée par le modèle PESQ est de 3,5.

Les pertes de paquets ont été simulées grâce au modèle de Gilbert [83]. Ce modèle est une chaîne de Markov à temps discret, d'ordre 1. La discrétisation temporelle se fait par rapport aux trames successives du signal. L'ordre du modèle correspond à l'entropie des différents états. L'entropie d'un système est la mesure de sa complexité. Dans notre application elle correspond aux différents liens existants entre les états. Dans le cas où l'entropie est de 0, la prédiction de l'état consécutif suivant est parfaitement fiable, tandis que plus l'entropie est élevée, plus la prédiction est aléatoire.

Le modèle de Gilbert comprend deux états représentés par un état "0" dans lequel le paquet est transmis, et l'état "1" dans lequel le paquet est perdu (cf. Fig. III.2). *P* est la probabilité de passer de l'état 0 à 1, et *Q* est la probabilité de passer de l'état 1 à 0.

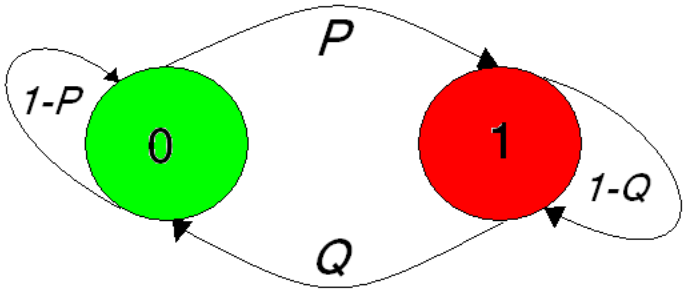


Fig. III.2 Modèle de Markov à l'ordre 1 (modèle de Gilbert 1960)

Ce modèle est alors caractérisé par la matrice de transition *T* défini comme

$$T = \begin{bmatrix} 1-P & P \\ Q & 1-Q \end{bmatrix} \qquad \textbf{Eq. III.1}$$

En résolvant l'équation [Π] = [*T*]·[Π] précédente, les probabilités [Π] = [Π_0 et Π_1] sont définies comme :

$$\Pi_0 = \frac{Q}{P+Q}$$
$$\Pi_1 = \frac{P}{P+Q} \qquad \textbf{Eq. III.2}$$

avec Π_0 le taux moyen d'être dans l'état 0 (pas de perte), et Π_1 le taux moyen d'être dans l'état 1 (pertes de paquets).

Dans la pratique, nous avons renseigné les valeurs du taux de pertes de paquets (Π_1), de la corrélation entre les erreurs de bits et les pertes de trames (valeur par défaut = 0), et de la déviation maximale de la perte de trame fixée à 0. Les pertes obtenues sont aléatoires, avec cependant une ou deux pertes par rafale (burst) suivant le pattern généré.

Le premier pattern "*pl2*" a été construit pour représenter 2% de pertes de paquets présentés de manière aléatoire. Le deuxième pattern "*pl4*" a été généré à partir du premier pattern "*pl2*", en rajoutant 2 % de pertes de paquets, et ainsi de suite. De cette manière, les discontinuités sont situées, au moins, au même endroit dans la phrase, ce qui assure que l'augmentation du pourcentage de pertes de paquets est bien assimilée à une augmentation de la dégradation de la qualité vocale. Cela n'aurait pas toujours été le cas selon que les discontinuités sont situées sur des zones plus ou moins gênantes du signal.

On obtient alors six conditions de pertes de paquets nommées "*pl2*", "*pl4*", "*pl6*", "*pl8*", "*pl10*", "*pl12*", correspondant respectivement à des pourcentages calculés de 2%, 4%, 6%, 8%, 9,6%, 11,4%. Les pourcentages qui ont été définis en entrée du modèle de Gilbert ne sont pas exactement les pourcentages appliqués au signal car le modèle génère les pertes à partir des probabilités d'être dans les différents états.

Certains codages comme le G.711 et le G.729 utilisent des algorithmes de PLC (Packet Loss Concealment) qui estiment le signal manquant afin de limiter la perception des discontinuités. Les conditions de dégradations qui présentent des pertes de paquets sans l'utilisation d'algorithme de PLC sont notées par exemple *"pl2_noplc"* dans le cas de la présence de 2% de pertes de paquets. Le taux d'erreur le plus élevé (11,4%) en utilisant le codage G.711 correspond à un score moyen de qualité vocale d'environ 3 sur l'échelle MOS (estimation obtenue à partir du modèle PESQ).

Les erreurs de bits sont causées par des problèmes de transmissions mobiles. Trois patterns de dégradation ont été générés à partir de la même technique que dans le cas des

pertes de paquets (modèle de Gilbert). Elles sont de 0,2%, 0,41% et 0,65 %. Le taux d'erreurs le plus élevé (0,65%) avec l'utilisation du codage GSM-EFR correspond à un score moyen de qualité vocale d'environ MOS = 3 (estimation obtenue à partir du modèle PESQ).

Le bruit de fond a été généré par un bruit gaussien aléatoire avec une décroissance de 3 dB par octave. Ce bruit est couramment appelé "bruit rose". Il a été choisi pour représenter la dégradation liée aux bruits de fond. Un seul type de bruit de fond a été retenu dans cette étude car le nombre de stimuli est limité par la durée de l'expérience. De plus, une étude approfondie concernant l'influence du type de bruit de fond sur l'évaluation de la qualité vocale a été réalisée et détaillée dans le Chapitre IV de ce manuscrit. Afin de représenter ce type de dégradation dans le modèle DESQHI, sept niveaux de bruit de fond ont été générés par pas de 2 dB, pour des rapports signal sur bruit (RSB) allant de RSB = 36 à 25 dB (36, 34, 32, 30, 28, 26, 25). Les niveaux des bruits de fond sont nommés de 1 à 7, suivant l'ordre croissant des niveaux appliqués. *BDF1* correspond à un RSB appliqué de 36 dB, tandis que *BDF7*, à un RSB de 25 dB.
Les conditions "non bruitées" ont quand même été présentées avec un bruit de fond qualifié de "bruit résiduel", correspondant à un bruit rose à un RSB de 44 dB par rapport au niveau de la parole. Ce bruit de fond est introduit pour plus de réalisme, car en situation téléphonique, il y a toujours un faible bruit généré soit par le codage (bruit de quantification), soit par le bruit ambiant présent autour du locuteur, soit par le combiné utilisé.
L'application des différents codages sur le signal global (parole plus bruit de fond) a modifié les rapports signal sur bruit présentés dans ce paragraphe. Les rapports signal sur bruit diffusés aux sujets ont été calculés et présentés dans le Tab. III.1.

	1	2 BDF1	3	4	5 BDF6	6	7	8 BDF4	9 BDF2	10	11	12
Femme	43,6	35,8	43,6	45,3	28,1	45,3	44,7	30,5	34,2	43,6	43,8	47,5
Homme	43,4	34,2	43,4	45,1	27,5	45,5	44,6	29,2	33,0	43,4	43,5	47,1
	13 BDF3	14	15 BDF 5	16	17	18	19 BDF7	20	21	22	23	
Femme	35,7	48,8	33,4	49,3	49,4	47,8	29,7	47,6	47,3	47,1	47,5	
Homme	34,7	48,7	32,6	49,1	48,8	47,5	29,2	46,8	46,8	46,2	46,9	

Tab. III.1 Rapports signal sur bruit en dB, calculés pour les 23 conditions de dégradations suivant les deux locuteurs (homme et femme)

Les différentes combinaisons possibles entre ces quatre types de dégradations étaient trop nombreuses pour être proposées lors du test subjectif d'évaluation des dissimilarités. Une sélection des conditions de dégradation a été réalisée afin de simuler au mieux des communications réalistes, et limiter le nombre de conditions afin de réduire la durée des expériences. 23 conditions de dégradation ont été finalement retenues :

1- *G.711*	**12- G.729**
2- G.711_BDF1	**13- G.729_BDF3**
3- G.711_G.726	**14- G.729_G.729**
4- G.711_GSMEFR	**15- G.729_G.729_BDF5**
5- G.711_GSMEFR_BDF6	**16- G.729_G.729_pl4**
6- G.711_GSMEFR_pbit02	**17- G.729_G.729_pl8**
7- G.711_GSMEFR_pbit06	**18- G.729_GSMEFR**
8- G.711_ppl10_BDF4	**19- G.729_GSMEFR_BDF7**
9- G.711_pl2_noplc_BDF2	**20- G.729_GSMEFR_pbit02**
10- G.711_pl4	**21- G.729_GSMEFR_pbit04**
11- G.711_pl6_noplc	**22- G.729_pl12**
	23- G.729_pl6

Tab. III.2 Description des 23 conditions de dégradation

Dans un premier temps, le signal vocal prononcé par un des locuteurs (homme ou femme) a été sous-échantillonné à 8 kHz, avec une quantification de 16 bits. Ensuite, un filtre appelé SRI "Système de Référence Intermédiaire" décrit dans la recommandation P.48 [84] a été appliqué à ce signal afin de simuler la réponse en fréquence d'un terminal. Ce signal a alors été égalisé par rapport aux zones actives de la parole au niveau de –26 dBov (ce qui permet ensuite de maîtriser le niveau de restitution en sortie du casque audio pour le test).
On considère que le bruit de fond provient de l'environnement du locuteur, et qu'il est soumis aux mêmes dégradations que le signal de la parole. Par ailleurs, il est probable que le bruit de fond généré par un premier codec est aussi sujet aux dégradations générées par le réseau (p. ex. pertes de paquets, deuxième codec). Le signal de bruit de fond "rose" a donc été soumis au même traitement que le signal vocal (cf. Fig. III.3). Le signal vocal a été mixé avec le signal de bruit de fond pour un des rapports signal sur bruit (RSB = 25 à 44 dB).
En présence de transcodage, le premier codec a été appliqué au signal global (parole et BDF), puis le deuxième codage a alors été appliqué au signal global avant d'appliquer le pattern des pertes de paquets et le pattern d'erreurs de bits. Le deuxième décodage est alors appliqué au signal global.

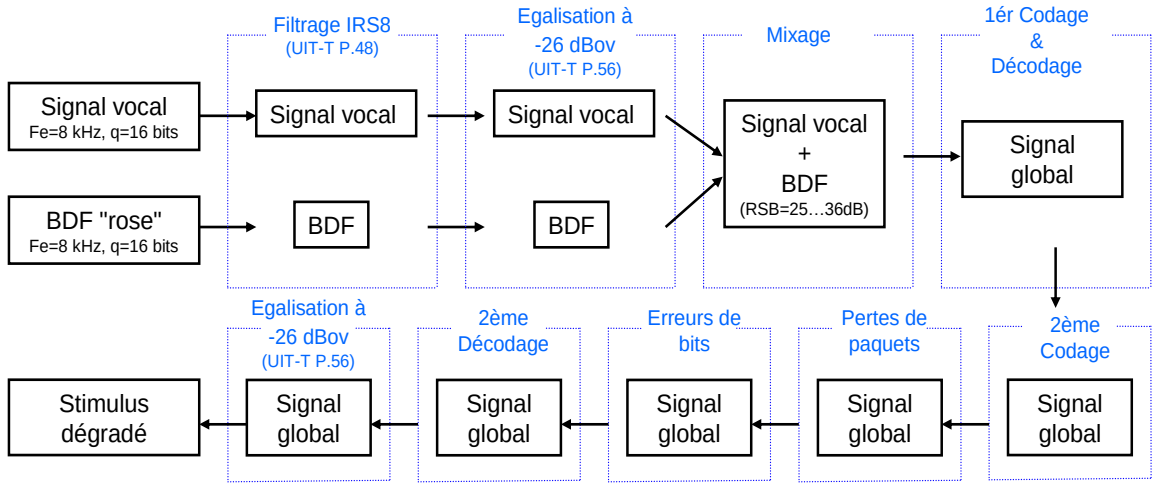


Fig. III.3 Les étapes de la construction des stimuli

Dans le cas de notre étude, le niveau sonore du signal vocal ne doit pas être un facteur influençant les résultats de l'évaluation des dissimilarités car nous considérons que le niveau sonore de la communication peut être ajusté par l'utilisateur dans une télécommunication réelle. Certains types de codec modifient la sonie des stimuli par rapport à d'autres (p. ex.

G.711 et G.729). La dernière étape du traitement a donc consisté à égaliser le niveau sonore des stimuli pour que les zones actives de la parole correspondent à un niveau de -26 dBov en utilisant la norme P.56 [80]. Cela permet de diffuser les stimuli à des niveaux sonores de parole similaires.

La base sonore a tout d'abord été construite à partir de la phrase prononcée par la voix de femme, puis avec la phrase prononcée par la voix d'homme. Les mêmes conditions de dégradations ont été utilisées afin de pouvoir comparer les résultats suivant les deux locuteurs. Deux bases sonores constituées de 23 stimuli chacune ont ainsi été construites.

III.1.2. Restitution sonore

Les deux tests d'évaluation de la qualité vocale et d'évaluation des dissimilarités ont été réalisés dans deux salles juxtaposées de France Télécom prévues à cet effet. L'avantage de ces salles est que huit sujets peuvent réaliser le test simultanément (quatre dans chaque salle), ce qui permet d'avoir un gain de temps.

La restitution des sons s'est faite au travers de casques stéréo "Sennheiser HD 25". Les stimuli ont été diffusés en diotique[11] à un niveau optimal de 69 dB SPL par oreille (cf. §.I.4.8).

Chacun des huit postes comporte deux potentiomètres linéaires ajustables, pouvant fournir des résultats allant de 1 à 100. L'un a été utilisé pour le premier test de similarité, l'autre a été utilisé pour le test d'évaluation de la qualité vocale.

Les stimuli et les paires de stimuli ont été présentés dans un ordre aléatoire dans les deux salles. De plus, dans le cas du test d'évaluation des dissimilarités, les stimuli au sein d'une paire (A-B ou B-A) ont été diffusés dans un ordre aléatoire dans les deux salles.

III.1.3. Sujets

48 sujets, de 18 à 40 ans, ont été sélectionnés pour réaliser ces deux tests. La répartition homme / femme des 48 sujets est à peu près équivalente. La plupart d'entre eux sont naïfs, mais quelques uns ont l'habitude de réaliser ce genre d'expérimentation. 24 d'entre eux évaluent les dissimilarités entre les 23 stimuli prononcés par le locuteur femme. Ce premier test a été réalisé en 2h. Ensuite, le deuxième test concernant l'évaluation de la qualité vocale a été proposé aux mêmes 24 sujets pour une durée de 30 min.

Les 24 autres sujets ont été soumis à ces deux mêmes tests mais en utilisant la base sonore prononcée par le locuteur homme.

III.1.4. Evaluation de la qualité vocale

La méthode d'évaluation de la qualité vocale la plus courante consiste à demander directement l'opinion des sujets sur la qualité vocale. Pour cela, une adaptation du test ACR "Absolute Category Rating" [11] a été choisie (cf. §.I.2.1). L'adaptation de la méthode ACR concerne l'échelle de mesure utilisée. En effet, le test ACR est défini dans la norme comme un test catégoriel à cinq niveaux décrit par les adjectifs présentés dans le Tab. I.1. Nous avons choisi d'utiliser une échelle linéaire renseignée par ces mêmes adjectifs car nous pensons que

[11] Une écoute diotique correspond à la diffusion d'un même signal audio aux deux oreilles de l'auditeur.

certains sujets ont la capacité de déterminer des différences de qualité vocale avec une précision supérieure à celle qui est mise en œuvre dans les cinq catégories du test ACR.

Il a été demandé à des sujets d'écouter chaque échantillon de parole complètement, puis d'indiquer leur opinion sur la qualité globale du signal de la parole en positionnant le curseur à l'emplacement approprié selon l'échelle présentée sur la Fig. III.4. Les sujets ont écouté chaque stimulus une seule fois, puis ils ont disposé de 5 secondes pour donner leurs jugements. Les consignes de ce test sont présentées en Annexe B.

Fig. III.4 Interface du test d'évaluation de la qualité vocale (adapté du test ACR)

Les résultats de chacune des conditions ont été moyennés suivant tous les sujets afin de déterminer la note "MOS-LQSN" (Mean Opinion Score - Listenning Quality Subjective Narrowband) (UIT P.800 [11]).

III.1.5. Evaluation des dissimilarités

L'évaluation des dissimilarités entre les 23 stimuli prononcés par un même locuteur a été réalisée par la méthode de comparaison par paires (cf. §.I.3.2.1) dans le but d'extraire un espace perceptif représentatif des différences de sonorité (cf. §.III.3). Il a été demandé aux deux groupes de 24 sujets d'évaluer la dissimilarité entre deux échantillons vocaux, sur une échelle continue allant d'**identiques** à **très différents**. Cette échelle est renseignée par sept repères afin de guider les sujets dans leur tâche.

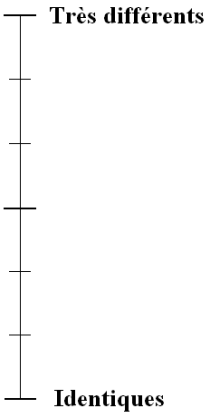


Fig. III.5 Interface du test de dissimilarité

Les consignes de ce test d'évaluation des dissimilarités sont présentées dans l'Annexe C. Toutes les paires de stimuli ont été présentées aux sujets, soit n·(n-1)/2 paires de stimuli, correspondant à 253 stimuli. Les paires ont été diffusées dans un ordre aléatoire selon les 12 groupes de 4 sujets. Chaque paire de stimuli a une durée de 15 secondes (10 secondes de signal sonore et 5 secondes de temps de réponse). L'expérience s'est déroulée en plusieurs sessions d'environ 13 min chacune, séparées par des temps de pause afin de reposer les sujets, pour une durée totale de 2 heures.

III.2. Résultats des tests subjectifs

III.2.1. Résultats du test d'évaluation de la qualité vocale

Les résultats de l'évaluation de la qualité vocale des 23 conditions de dégradation (cf. Tab. III.2) sont présentés sur la Fig. III.6. Les résultats moyennés sur l'ensemble des 48 sujets montrent que les conditions dégradant le plus la qualité vocale sont les conditions 11, 8 et 22. Ces conditions correspondent à la présence de taux élevé de pertes de paquets ainsi qu'à la présence de bruit de fond. Les stimuli jugés de bonne qualité concernent les conditions 1, 3, 4, 10, 12, 14 et 18 qui correspondent pratiquement toutes (à l'exception de la condition 10) à des dégradations liées uniquement au codage de la parole (codage simple et transcodage).

Les résultats sont ensuite présentés en distinguant les réponses obtenues selon les deux locuteurs femme et homme (cf. Fig. III.6).

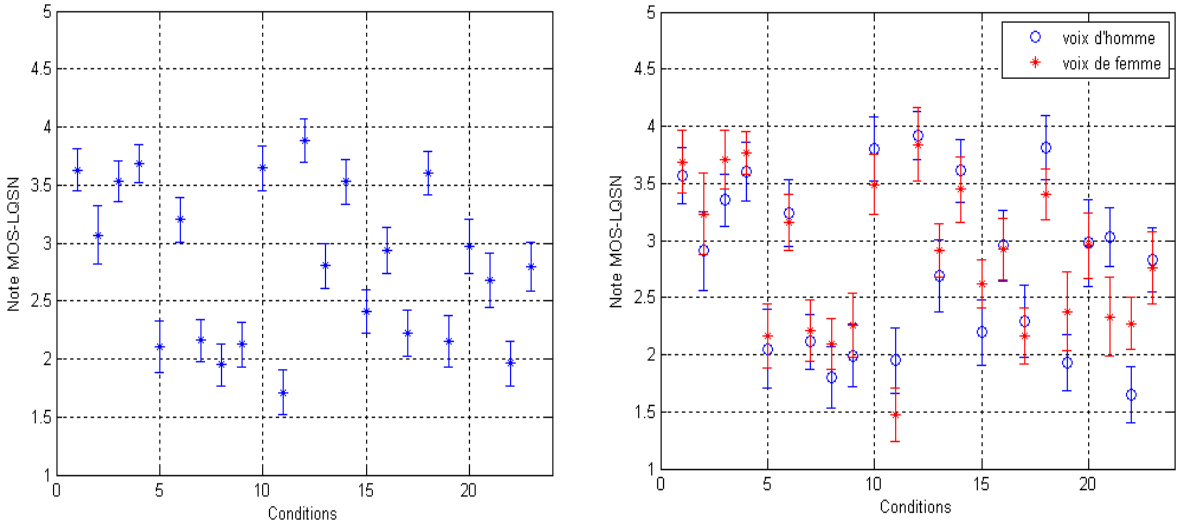


Fig. III.6 Notes MOS-LQSN issues du test d'évaluation de la qualité vocale réalisé par les 48 sujets avec l'intervalle d'incertitude bilatéral à 5% de l'estimation de la moyenne; La figure de gauche représente la moyenne sur tous les sujets et la figure de droite fait la distinction entre les deux locuteurs homme et femme

Les résultats montrent que les résultats sont similaires entre les deux locuteurs, excepté pour les conditions 21 et 22 qui sont significativement différentes entre les deux locuteurs ($\alpha < 0,05$). Nous remarquons aussi une forte différence entre les deux locuteurs pour la condition 11. Ces trois conditions correspondent à des dégradations de pertes de paquets ou à des erreurs de bits.

Les pertes de paquets et les erreurs de bits peuvent engendrer de grandes variations de l'évaluation de la qualité vocale selon l'endroit où se situent les discontinuités. Les patterns des pertes de paquets et d'erreurs de bits sont identiques pour les deux locuteurs, cependant,

comme les phrases sont différentes, il se peut que les pertes ou erreurs soient localisées à des endroits plus ou moins gênants dans la phrase.

Une hypothèse déjà utilisée lors de la construction de nombreux modèles (cf. §.I.4.3 Ding [56] et Kim [58]) est que si la discontinuité se trouve sur les zones actives du signal (parole), l'évaluation de la qualité vocale sera moins bonne que si la discontinuité se trouve dans une zone non active (dans ce cas, la qualité vocale peut ne subir aucune dégradation). Cette hypothèse pourrait être une explication à la différence d'évaluation de la qualité vocale entre les deux locuteurs, pour certaines conditions présentant des discontinuités.

III.2.2. Résultats du test d'évaluation des dissimilarités

Les résultats de l'évaluation des dissimilarités sont composés des 253 jugements de dissimilarité, pour chacun des 48 sujets. Les valeurs obtenues sont comprises entre 1 (identiques) et 100 (très différents).

Le test a été réalisé par 12 groupes de 4 sujets dans deux salles différentes. Les échantillons de parole prononcés par le locuteur femme (sujets numérotés de 1 à 24) ont été diffusés dans la salle 1, tandis que les échantillons prononcés par le locuteur homme (sujets numérotés de 25 à 48) ont été diffusés dans la salle 2. L'ordre de présentation des paires de stimuli était identique pour chacun des 12 groupes, ce qui correspond à des ordres de diffusions identiques pour les sujets 1 à 4, les sujets 5 à 8, les sujets 9 à 12, etc...

Les proximités entre les jugements donnés par les sujets ont été mesurées afin de vérifier ou non l'existence de groupes entre les 48 sujets lors de la tâche demandée (Fig. III.7). Pour cela, la matrice de dimension (253·48) des résultats du jugement des 253 dissimilarités a tout d'abord été centrée et réduite selon les sujets afin de s'affranchir des différentes tactiques d'utilisation de l'échelle. Ensuite, cette matrice a été transformée en des distances euclidiennes pour chaque sujet. Les proximités ont alors été déterminées grâce à la méthode dite "the complete linkage clustering" qui consiste à déterminer les groupes de sujets suivant leurs différences maximales. Ces proximités peuvent alors être représentées par un dendrogramme.

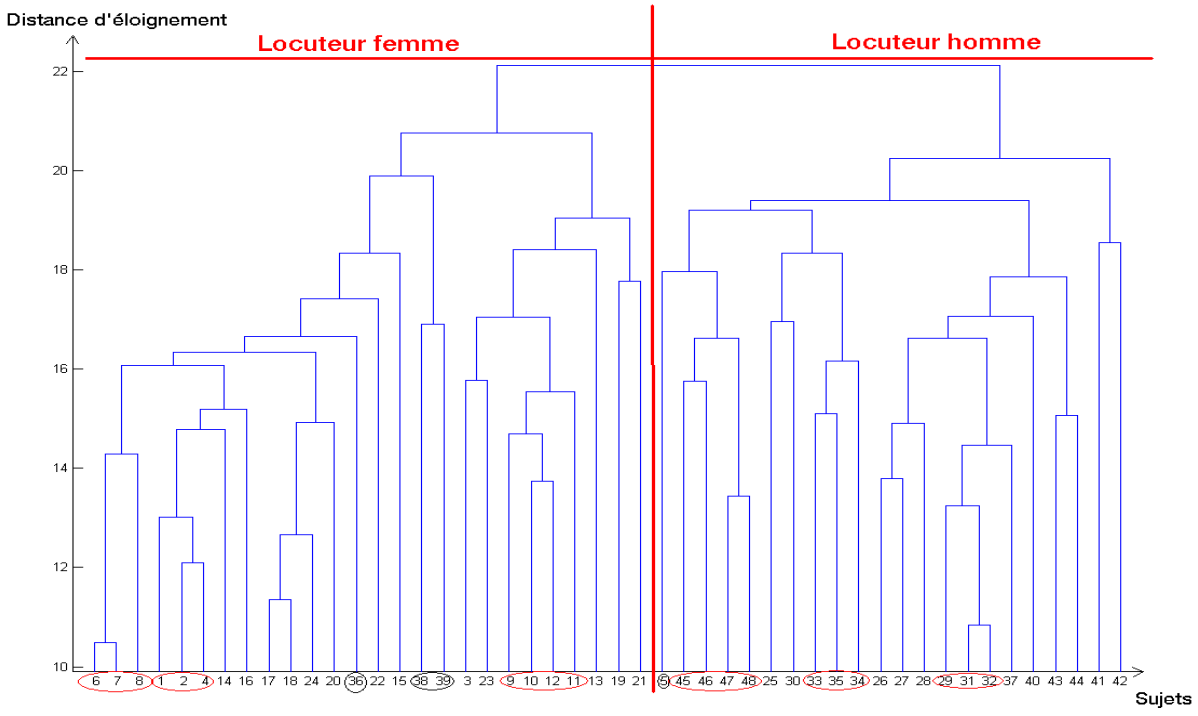


Fig. III.7 Dendrogramme des résultats de l'évaluation des dissimilarités pour les sujets numérotés de 1 à 24 pour ceux évaluant la voix de femme, et les sujets numérotés de 25 à 48 pour ceux évaluant la voix d'homme

Le dendrogramme (Fig. III.7) montre une nette séparation entre deux groupes correspondant au locuteur femme pour les sujets numérotés de 1 à 24 et au locuteur homme pour les sujets numérotés de 25 à 48. Seulement quatre sujets sur les 48 (36, 38, 39, 5) ne sont pas dans le bon groupe de locuteurs. Cela confirme qu'il existe une différence des jugements de dissimilarité entre les deux locuteurs.

D'autre part, les sujets situés dans la même salle ayant été soumis au même ordre de diffusion des paires, sont souvent admis dans les mêmes sous-groupes, comme par exemple les groupes de sujets encerclés en rouge sur la Fig. III.7. Ces groupes correspondent aux sujets qui ont eu des résultats similaires entre eux, lors du jugement des dissimilarités. Ce phénomène indique que l'effet de l'ordre de diffusion des paires de stimuli joue un rôle sur les résultats de ce test. Cet effet n'est cependant pas pris en compte lors des analyses suivantes car il est estompé par le fait que les ordres de présentation étaient différents entre les sous groupes.

Comme le dendrogramme a confirmé l'existence d'une différence entre les deux locuteurs homme et femme, les jugements des dissimilarités ont été moyennés suivant les deux groupes de 24 sujets afin de faire la distinction entre ces deux locuteurs (cf. Fig. III.8).

Le coefficient de corrélation de Pearson r entre les résultats obtenus pour la voix d'homme et les résultats obtenus pour la voix de femme, est de $r = 0{,}80$ ($p < 0{,}01$) Ce coefficient montre que les jugements obtenus sont cohérents suivant les conditions de dégradations, cependant il existe pour certaines d'entres elles, une différence entre les deux locuteurs femme et homme.

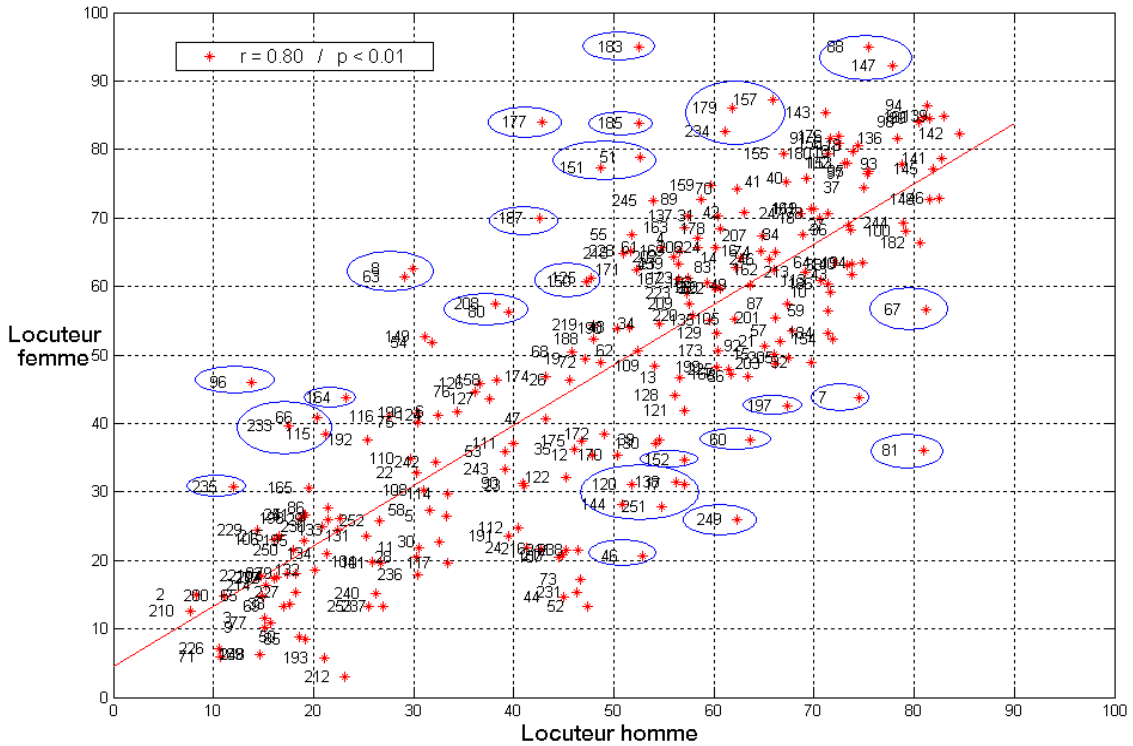


Fig. III.8 Comparaison des résultats de dissimilarité entre les deux locuteurs homme et femme

Il apparaît que les paires de conditions de dégradation les plus différentes entre les jugements issus des locuteurs homme et femme, correspondent la plupart du temps à la présence d'au moins une condition relative à la présence de pertes de paquets ou à des erreurs de bits (cf. Fig. III.8). Plus précisément, plus les erreurs de pertes de paquets ou les erreurs de bits deviennent importantes, plus l'écart entre les deux locuteurs homme et femme devient

important. Certaines de ces paires de stimuli sont encerclées en bleu sur la Fig. III.8. Pour exemple, les paires de stimuli 183, 177, 88 et 51 intègrent au moins la condition présentant 8% de pertes de paquets. Ce phénomène vérifie l'hypothèse posée dans le paragraphe III.2.1 à propos de l'influence de l'emplacement des pertes de paquets et des erreurs de bits dans la phrase. D'autres hypothèses sont proposées afin de justifier cet effet (cf. §.III.3.3.3).

Les résultats des tests d'évaluation de la qualité vocale et de l'évaluation des dissimilarités montrent qu'il existe une différence entre les deux voix utilisées. Lors de l'étape de la détermination de l'espace perceptif, les deux bases sonores ont donc été analysées de manière indépendante.

III.3. Détermination de l'espace perceptif

L'espace perceptif est déterminé par la méthode sans *a priori* (cf. §.I.3.2), pour les raisons exposées au Chapitre II, dans le but d'être relié par la suite à l'évaluation de la qualité vocale. Les résultats du test d'évaluation des dissimilarités sont les données d'entrée de la méthode d'échelonnement multidimensionnel de type INDSCAL métrique. La première partie décrit le choix du nombre de dimensions pertinentes à la représentation de l'espace perceptif. La deuxième partie présente l'espace perceptif, puis chacune des dimensions constituant l'espace est identifiée par des attributs perceptifs.

III.3.1. Choix du nombre de dimensions pertinentes

Le nombre de dimensions optimal pour la représentation de l'espace perceptif est déterminé à partir de la valeur de l'erreur commise lors de la reconstruction des dissimilarités en distances euclidiennes. Cette erreur est déterminée par la valeur du *stress* (cf. Annexe A et §.I.3.2.2) qui est généralement représentée en fonction du nombre de dimensions constituant l'espace sur le graphique appelé "scree plot" (cf. Fig. III.9).

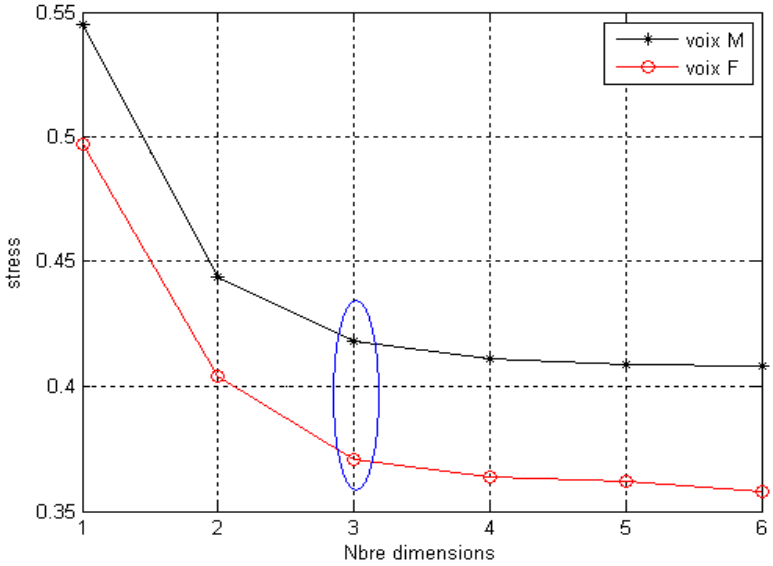


Fig. III.9 Scree plot : Valeur de l'erreur commise lors de la reconstruction des dissimilarités suivant le nombre de dimensions considéré, pour les deux locuteurs (homme et femme)

Les valeurs de stress sont présentées pour les deux locuteurs (femme et homme) sur la Fig. III.9. Ces valeurs montrent que les erreurs commises dans le cas du locuteur homme sont

supérieures à celles de la voix de femme, néanmoins les deux courbes suivent la même décroissance suivant le nombre de dimensions constituant l'espace perceptif.

Dans la pratique, le nombre de dimensions retenu est fixé lorsque la prise en compte d'une dimension supplémentaire n'améliore que très peu la valeur du stress. La Fig. III.9 montre bien ce phénomène en représentant la courbe de l'erreur de reconstruction des dissimilarités, en fonction du nombre de dimensions constituant l'espace perceptif. La courbe forme un coude au niveau du nombre de dimensions optimal correspondant au meilleur compromis entre le nombre minimal de dimensions pour une erreur commise minimale. Le nombre de dimensions optimal est fixé à trois dimensions dans le cas des deux locuteurs, car l'ajout d'une dimension supplémentaire n'apporte que peu d'informations supplémentaires. La valeur de l'erreur commise est alors de 0,37 dans le cas de la voix de femme et de 0,42 dans le cas de la voix d'homme.

Une analyse bootstrap est réalisée en complément pour les espaces constitués de trois et quatre dimensions, afin de tester la fiabilité des espaces obtenus. L'analyse bootstrap consiste à extraire l'espace perceptif un certain nombre de fois (50 fois ici), à partir des données de dissimilarités obtenues pour 24 sujets choisis aléatoirement avec remise à chacun des 50 tirages. Les dispersions des coordonnées des points suivant chacune des dimensions composant l'espace perceptif permet d'observer si la représentation des stimuli est pertinente ou non. Les dispersions sont représentées par les intervalles d'incertitude bilatérale à 5%. (cf. Fig. III.10 et Fig. III.11).

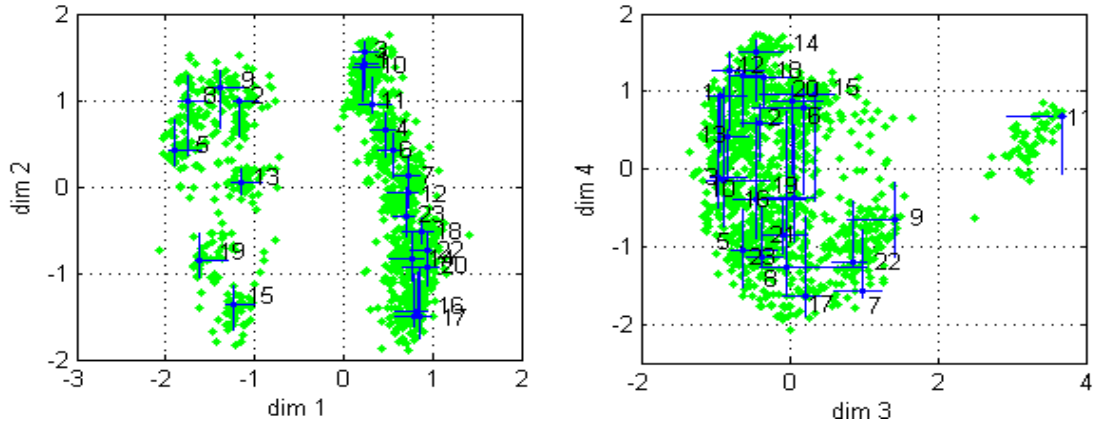


Fig. III.10 Représentation de l'espace perceptif à 4 dimensions pour la voix de femme avec l'analyse bootstrap à 50 tirages, avec les intervalles d'incertitude bilatérale à 5%

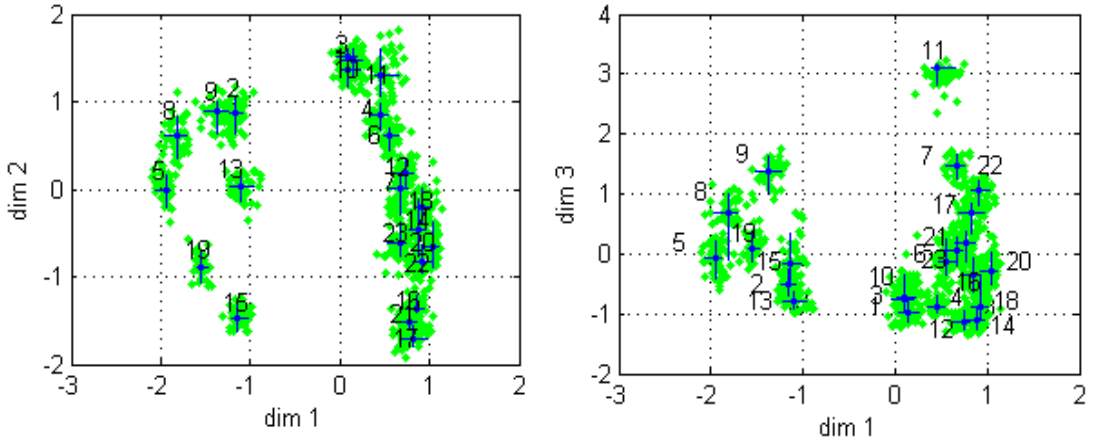


Fig. III.11 Représentation de l'espace perceptif à 3 dimensions pour la voix de femme avec l'analyse bootstrap à 50 tirages, avec les intervalles d'incertitude bilatérale à 5%

La Fig. III.10 montre que la quatrième dimension n'est pas efficace pour distinguer les 23 conditions de dégradation car les intervalles d'incertitude se superposent suivant cet axe. L'espace perceptif à trois dimensions semble être plus fiable car les intervalles d'incertitude ne sont pas trop importants selon les trois axes.

Ce choix du nombre de dimensions est aussi conforté car la combinaison linéaire des trois dimensions permet d'expliquer les notes d'évaluation de la qualité vocale avec une précision définie par le coefficient de détermination ($R^2 = 0{,}86$; cf. §.III.4).

III.3.2. Extraction de l'espace perceptif

Deux espaces perceptifs tridimensionnels sont déterminés, correspondant aux résultats de dissimilarités des stimuli prononcés par la voix de femme (cf. Fig. III.12) et des stimuli prononcés par la voix d'homme (cf. Fig. III.13). Ces deux espaces correspondent aux distances euclidiennes calculées par l'analyse d'échelonnement multidimensionnel. Les coordonnées des points pour chaque dimension sont centrées et réduites. Les conditions de dégradation sont représentées par les numéros de 1 à 23 (cf. Tab. III.2).

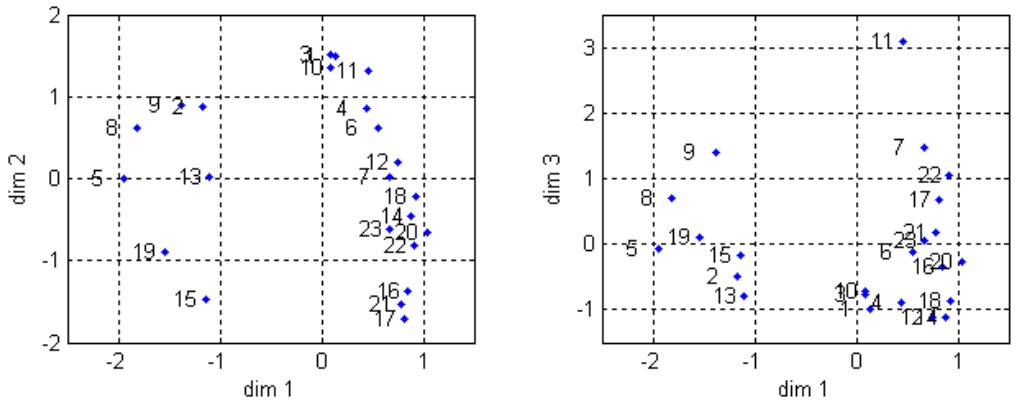


Fig. III.12 Espace perceptif à 3 dimensions de la voix de femme, pour les 23 conditions de dégradation

Au sujet de l'espace perceptif de la voix d'homme, nous avons remarqué des similitudes avec l'espace "femme", et notamment en inversant le sens d'orientation des dimensions 2 et 3. Afin de relier ces deux espaces, une transformation procrustéenne[12] est réalisée sur l'espace "homme", par rapport à l'espace "femme". La transformation procrustéenne appliquée sur l'espace "homme" est définie par l'équation suivante :

$$[Z] = c \cdot [Y] \cdot [R] + [t], \qquad \textbf{Eq. III.3}$$

avec $[Z]$ les coordonnées des 23 conditions de la voix d'homme dans l'espace transformé, $[Y]$ les coordonnées des 23 conditions de la voix d'homme dans l'espace d'origine, c la composante de changement d'échelle, $[R]$ les composantes de rotation de l'espace, et $[t]$ les composantes de translation des 3 dimensions. La composante de changement d'échelle est de 0.919 entre les deux espaces. Les composantes de translation $[t]$ sont négligeables (de l'ordre de 10^{-5}). Les composantes de rotation orthogonale et les composantes de réflexion $[R]$ sont les suivantes :

[12] La transformation procrustéenne est une transformation linéaire incluant les translations, les réflexions, les changements d'échelles et les rotations orthogonales. La transformation procrustéenne des axes n'altère en rien le rapport entre les distances euclidiennes des différentes conditions.

$$[R_{ij}] = \begin{bmatrix} 0{,}99 & -0{,}06 & -0{,}09 \\ -0{,}07 & -1{,}00 & -0{,}07 \\ -0{,}09 & 0{,}07 & -0{,}99 \end{bmatrix} \qquad \textbf{Eq. III.4}$$

R_{ij} représente les transformations de rotation appliquées entre la i^{ème} dimension de l'espace "homme" et de la j^{ème} dimension de l'espace "homme" transformé.
Dans le cas ou $i = j$, les trois différentes composantes indiquent que la première dimension est équivalente à celle d'origine avec $R_{11} = 0{,}99$, tandis que les deuxième et troisième dimensions sont inversées par rapport à l'espace original ($R_{ij} \approx -1$). Lorsque $i \neq j$, il n'y a presque pas d'influence de la transformation ($R_{ij} \approx 0$).

La transformation appliquée à l'espace "homme" par rapport à l'espace "femme" est essentiellement basée sur l'inversion du sens des dimensions 2 et 3. L'espace perceptif de la voix d'homme soumis à cette transformation est représenté sur la Fig. III.13.

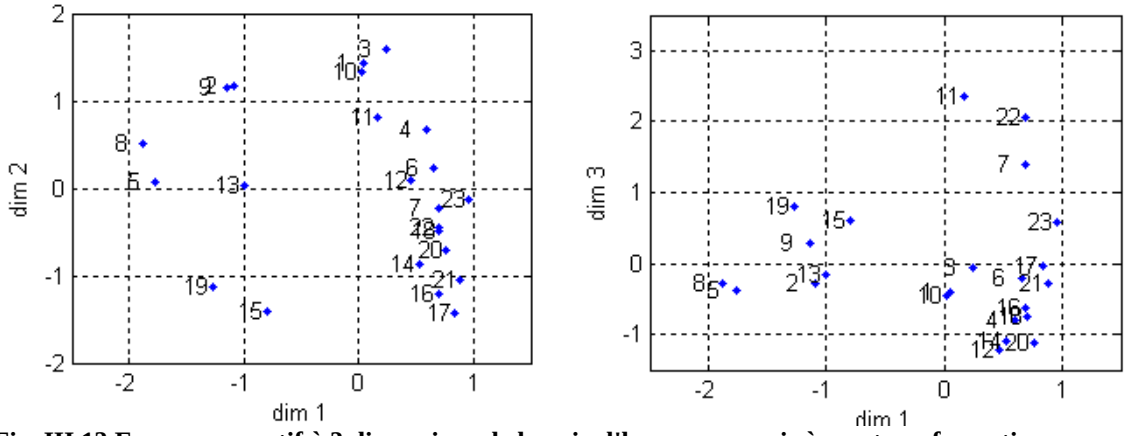
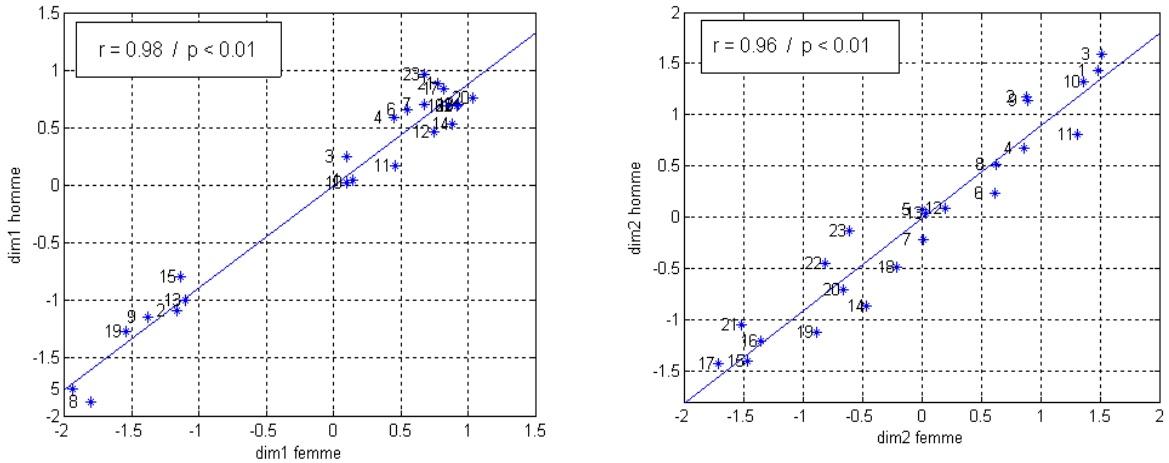


Fig. III.13 Espace perceptif à 3 dimensions de la voix d'homme, soumis à une transformation procrustéenne par rapport à l'espace de la voix de femme, pour les 23 conditions de dégradation

Afin de quantifier les similitudes entre les deux espaces "femme" et "homme", les coefficients de corrélations des positions des conditions de dégradation entre les deux locuteurs femme et homme sont calculés pour chacune des trois dimensions (cf. Fig. III.14).



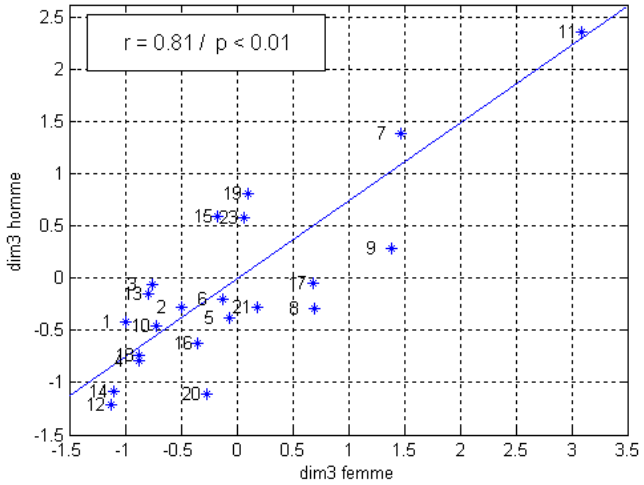


Fig. III.14 Comparaison des positions des conditions de dégradation entre les deux espaces femme et homme, soumis à la transformation procrustéenne, suivant chacune des trois dimensions

Les corrélations entre les voix de femme et d'homme de la première et de la deuxième dimensions ont des valeurs respectivement de *r = 0,98* et *r = 0,96*. Ces deux premières dimensions peuvent donc être considérées comme pratiquement identiques entre les deux locuteurs. Au sujet de la troisième dimension, une corrélation de *r = 0,81* est calculée entre les deux espaces femme et homme. Il existe donc une faible différence entre les deux espaces suivant la troisième dimension qui peut être expliquée par les différences d'emplacement des pertes de paquets suivant les deux locuteurs. Ce phénomène a déjà été mentionné dans les parties III.2.1 et III.2.2, et il est discuté dans le paragraphe III.3.3.3.

Compte tenu de l'importante ressemblance entre les deux espaces prononcés par le locuteur homme et le locuteur femme, les deux espaces sont superposés dans un espace global (cf. Fig. III.15). Cet espace fait l'hypothèse que la phrase et le locuteur sont des facteurs indépendants lors de la détermination de l'espace perceptif. La seule différence relevée concerne l'emplacement des pertes de paquets et les erreurs de bits dans le signal de la parole. Aucune différence n'est remarquée quant à la différence de timbre entre les deux voix (femme et homme).

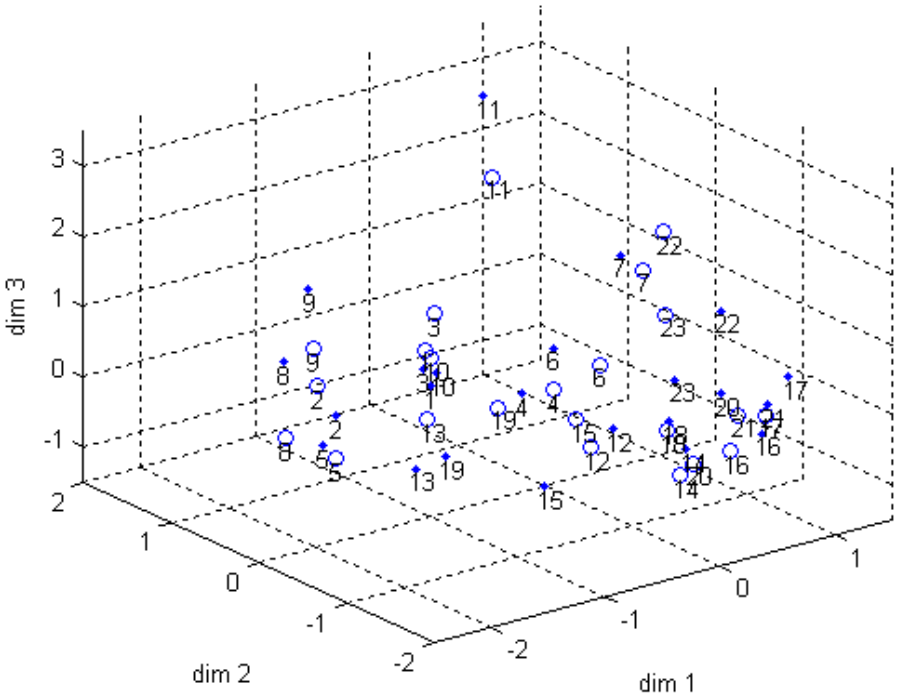


Fig. III.15 Espace perceptif global à 3 dimensions de la voix de femme (♦) et la voix d'homme (O), pour les 23 conditions de dégradation

III.3.3. Identification des dimensions perceptives

Les jugements de dissimilarité entre les stimuli ont permis, grâce à la méthode d'échelonnement multidimensionnel, de distinguer trois principales dimensions perceptives. L'identification des axes n'est pas une étape obligatoire pour relier l'espace perceptif à l'évaluation de la qualité vocale, mais elle constitue une information très utile lors de la modélisation qui est réalisée dans les deux chapitres suivants. Les trois dimensions sont identifiées dans cette partie par des attributs perceptifs (cf. §.I.4), à partir de l'écoute des conditions suivant chaque dimension. Ces trois dimensions correspondent respectivement aux attributs *"Bruyance"*, *"Codage de la parole"*, et *"Continuité"*.

III.3.3.1. Bruyance

La première dimension semble être représentative de la présence de bruit de fond (cf. Fig. III.16 et §.I.4.1).

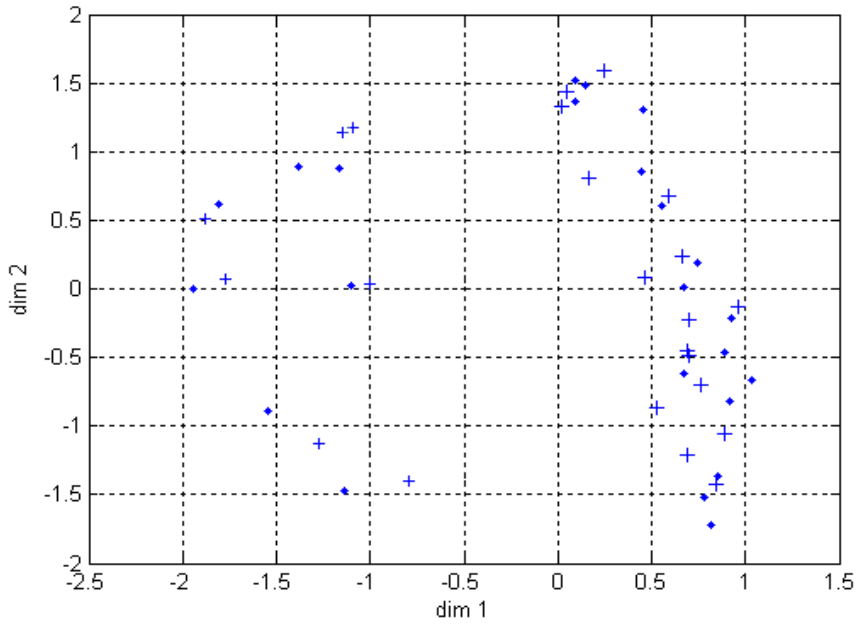


Fig. III.16 Identification de la dimension 1 : Bruyance pour la voix de femme (♦) et la voix d'homme (+)

Deux groupes de stimuli apparaissent clairement, correspondant à la présence ou non de bruit de fond sur le signal de parole. Le groupe de stimuli situé sur la droite (0 < dim 1 < 1 ; cf. Fig. III.16) est constitué des 32 conditions ne comportant pas de bruit ajouté, tandis que le groupe de gauche (-2 < dim 1 < -1,5 ; cf. Fig. III.16) est composé des 14 conditions présentant du bruit de fond ajouté.

Il se pose alors un problème concernant l'échelle catégorielle ou continue de cette première dimension. La détermination du cœur du modèle en est directement affectée.

Cette allure catégorielle est justifiée par la construction des conditions bruitées de la base sonore. Le choix des niveaux sonores appliqués aux conditions bruitées est limité à 7 rapports signal sur bruit qui correspondent à des rapports régulièrement espacés, compris entre 27 dB < RSB < 36 dB (RSB réel calculé sur les signaux (cf. Tab. III.1 et Fig. III.17)). Les conditions non bruitées présentent aussi des niveaux de bruit de fond correspondant à des RSB calculés sur les signaux allant d'environ 43 dB < RSB < 50 dB (cf. Tab. III.1 et Fig. III.17). Ces bruits proviennent du codec et du bruit résiduel et leurs niveaux sonores sont espacés de manière régulière les uns par rapport aux autres.

Il existe un manque de données dans la base sonore pour les conditions bruitées présentant des niveaux sonores compris entre 36 dB < RSB < 43 dB. Ce manque de données peut être à l'origine de l'allure catégorielle de cette première dimension. Nous avons calculé un coefficient de corrélation de $r = 0{,}97$ ($p < 0.01$) entre les coordonnées des points suivant la première dimension et les valeurs de niveau sonore des bruits de fond exprimées par le rapport signal sur bruit (cf. Fig. IV.18). Cependant, cette corrélation doit être considérée avec prudence à cause de la répartition bimodale des conditions suivant le premier axe. Dans ce cas, il est préférable de déterminer les coefficients de corrélation, d'une part suivant les conditions bruitées ($r = 0.77$), et d'autre part suivant les conditions non bruitées ($r = 0.77$) (cf. Fig. III.17).

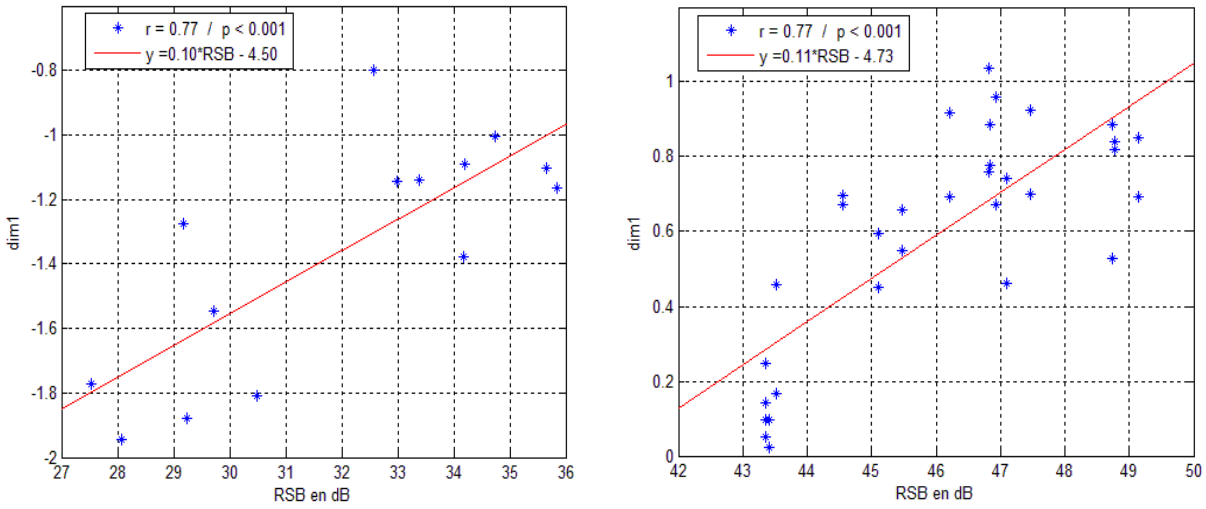


Fig. III.17 Représentation de la première dimension par le rapport signal sur bruit pour les conditions bruitées à gauche et non bruitées à droite

Ces résultats montrent que la dimension bruyance peut être considérée comme continue pour les deux types de conditions (avec et sans bruits ajoutés). Nous remarquons aussi que les équations des droites reliant la dimension et le rapport signal sur bruit sont pratiquement identiques pour les stimuli avec et sans bruit ajoutés (cf. Fig. III.17). Cela permet d'avancer l'hypothèse que des stimuli constitués de rapport signal sur bruit compris entre 36 et 43 dB peuvent être représentés par la même équation pour correspondre à la dimension bruyance.

Ces constatations permettent de vérifier que la dimension relative à la bruyance est bien représentée sur une échelle continue et qu'elle est principalement influencée par le niveau sonore du bruit de fond. Plus les stimuli sont situés vers la partie négative de cette première dimension, plus le niveau sonore du bruit est important.

Le faible niveau de bruit de fond présent sur les stimuli ne contenant pas de bruit ajouté provient du codage de la parole qui génère dans certains cas du bruit de quantification (cf. §.I.4.1). Par exemple, le codage G.711 (stimuli placés dans la partie extrême positive de la dimension 2) génère du bruit de fond de quantification plus important que le codage G.729 (stimuli placés dans la partie extrême négative de la dimension 2) (cf. Fig. III.16 et Fig. III.18).

III.3.3.2. Codage de la parole

L'écoute des stimuli le long de la deuxième dimension a permis d'identifier les attributs relatifs au codage de la parole (cf. Fig. III.18). Les coordonnées des points entre les deux locuteurs sont pratiquement identiques pour cette dimension (cf. Fig. III.14). La Fig. III.18 présente uniquement les conditions prononcées par le locuteur femme pour des raisons de lisibilité.

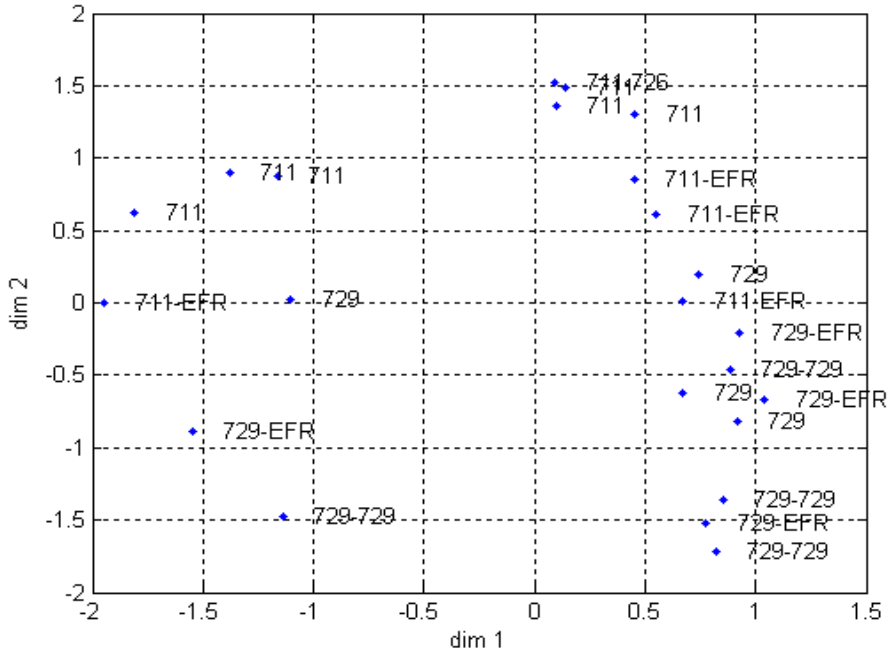


Fig. III.18 Identification de la dimension 2 : Codage de la parole pour la voix de femme (♦)

La répartition des codecs le long de la deuxième dimension peut être représentée de manière simplifiée (cf. Fig. III.19).

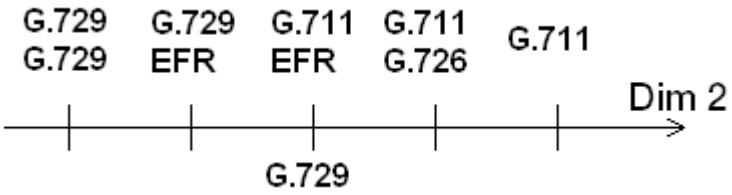


Fig. III.19 Echelle de correspondance entre la dimension 2 et les codages et transcodages

Cette deuxième dimension peut représenter plusieurs attributs. On remarque tout d'abord que cette dimension est étroitement liée à la dégradation de la qualité vocale suivant le codage employé. En effet, il est admis que le codage G.711 obtient le meilleur score de qualité vocale, et le transcodage G.729-G.729 la moins bonne note dans ce panel de codage utilisé.

Le naturel de la voix est étroitement lié à l'évaluation de la qualité vocale, comme le montre Hall [17] (cf. §.I.4.6). Notre deuxième dimension pourrait donc être reliée à cet attribut perceptif.

L'écoute des stimuli selon cette dimension permet aussi d'identifier l'attribut *coloration* de la parole (cf. §.I.4.9). L'utilisation du codage G.729 provoque une sensation sourde de la parole, par rapport au codage G.711 qui ne compresse pas le signal vocal et qui reconstitue le signal de manière plus claire. La partie positive de cette deuxième dimension comprend les voix *claires*, *brillantes* ou encore *colorées*, contrairement à la partie négative où la voix est plus *sourde*, *sombre*. Nous ne remarquons cependant aucune différence entre les deux locuteurs homme et femme suivant cette dimension, comme le montre la Fig. III.14, malgré la différence de timbre entre ces deux voix.

D'autres attributs ont aussi été identifiés tels que *la distorsion* et *le grésillement*, générés par la compression du signal de la parole.

III.3.3.3. Continuité

L'écoute des stimuli suivant la troisième dimension a permis d'identifier les attributs continuité et discontinuité de la parole. Les discontinuités sont principalement générées par les pertes de paquets et par les erreurs de bits. Ces deux dégradations physiques sont représentées par le même axe, malgré la différence de sonorité entre ces deux types de pertes. Les pertes de paquets sont perçues comme des trous dans le signal lorsque la PLC "Packet Loss Concealment" n'est pas utilisée par le codage. Les erreurs de bits, quant à elles, génèrent des dégradations ressemblantes à un effet *bulleux*, avec l'apparition de pics d'intensité. Les pourcentages d'erreurs de bits sont compris entre 0,2 et 0,6 %, tandis que les pourcentages de pertes de paquets correspondent à des valeurs allant de 2 à 12 %.

La Fig. III.20 représente les conditions de dégradations contenant des pertes de paquets ou des erreurs de bits. Plus les conditions de dégradation sont situées dans la partie positive de cette dimension, plus la perception de discontinuité est importante, comme par exemple dans le cas de la condition nommée "PL6no" correspondant à 6 % de pertes de paquets sans l'utilisation de l'outil PLC.

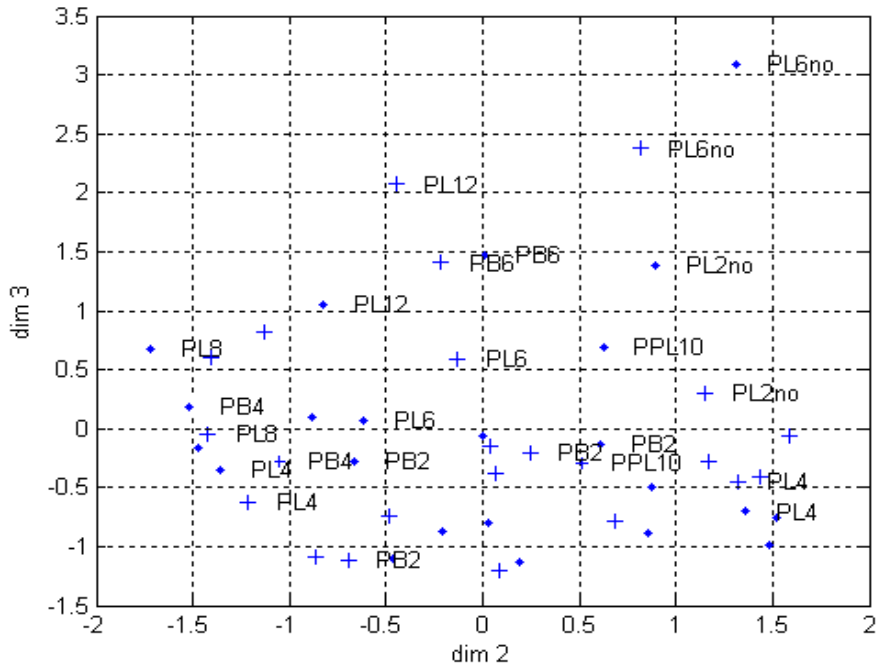


Fig. III.20 Identification de la dimension 3 : Continuité pour la voix de femme (♦) et la voix d'homme (+); "PL" correspond aux pertes de paquets avec PLC, "PLno" correspond aux pertes de paquets sans PLC, tandis que "PB" correspond aux erreurs de bits exprimées en dixième (0,2 / 0,4 / 0,6 %)

La perception de discontinuité est nettement plus prononcée lorsque la PLC n'est pas utilisée.

Les conditions positionnées en dessous de *0* sur l'axe des ordonnées sont constituées d'un mélange entre des conditions soumises ou non à des discontinuités (cf. Fig. III.20). Les stimuli placés dans cette partie correspondent à des signaux perçus continus ou présentant de faibles discontinuités provoquées par l'algorithme de PLC. Il est probable que d'autres dégradations physiques, provenant par exemple du codage de la parole, influencent la continuité du signal de la parole.

Nous observons pour certaines conditions de dégradation une différence entre les deux locuteurs homme et femme. Par exemple la condition constituée de 12% de pertes de paquets (*PL12*) présente une meilleure continuité pour la voix de femme que pour la voix d'homme. Inversement, la condition constituée de 2% de pertes de paquets sans utilisation de l'algo-

rithme de PLC (*PL2no*) présente une meilleure continuité pour la voix d'homme que pour la voix de femme (cf. Fig. III.20). Cela montre l'importance de la localisation de la perte de paquet sur le signal de parole, déjà remarquée dans les parties III.2.1 et III.2.2.

Deux hypothèses sont posées sur l'influence de la localisation des discontinuités.

- La première hypothèse est que les discontinuités présentes sur les zones non-actives de la parole peuvent être entendues s'il y a du bruit de fond, ou ne pas être détectées dans le cas où il n'y a pas de bruit de fond. Dans ce cas la perception de discontinuité du signal serait plus importante dans le cas d'un signal bruité qu'un signal ne contenant pas de bruit de fond.
- La deuxième hypothèse est que les pertes présentes sur les zones actives du signal sont nettement plus dérangeantes que sur les zones inactives. La localisation des pertes sur le signal vocal influencerait donc la perception de discontinuité. Il se pourrait même que la localisation des pertes sur les zones de parole (consonne ou voyelle par exemple) puisse influencer cette dimension, et dans certains cas influencer l'intelligibilité de la parole.

Ces deux hypothèses ont été testées lors de la modélisation de la dimension continuité (cf. §.V.2.2).

III.3.3.4. Comparaison avec les espaces perceptifs existants

Les espaces proposés par Wältermann et Mattila (cf. Tab. I.6) sont réalisés dans un domaine d'application similaire au notre (VoIP, GSM, RTC). Les dimensions bruyance et continuité ont été identifiées dans notre espace tridimensionnel, ainsi que dans les espaces déterminés par Mattila [51] et Wältermann [14]. Par ailleurs, la dimension bruyance est déterminée sur une échelle continue dans le cas de ces deux études, ce qui conforte le choix d'une échelle de bruyance continue et non catégorielle.

Ces auteurs ont aussi déterminé au moins la dimension coloration, mais aussi les dimensions sifflement et naturel de la voix (Mattila cf. Tab. I.6).

Les recherches réalisées par Etame [34] concernent uniquement les dégradations liées au codage de la parole. Dans ce cas, l'analyse révèle un espace à quatre dimensions, correspondant respectivement aux attributs coloration, bruit de fond, bruit sur la parole, et sifflement. La dimension bruit de fond (ou bruyance) est déjà prise en compte dans le cas des études citées ici. Nous remarquerons dans la suite de ce manuscrit que la dimension bruit sur la parole (correspondant aux conditions MNRU) est prise en compte dans notre modèle DESQHI, par la dimension codage de la parole (cf. §.V.1).

Notre deuxième dimension semble faire intervenir un grand nombre d'attributs relatifs au codage de la parole comme coloration, distorsion, distance, naturel de la voix, compression... (cf. §.III.3.3.2). Elle a été identifiée comme le codage de la parole puisqu'elle est bien représentée par le type de codage, cependant il est clair que les différents attributs relevés mettent en évidence un lien avec les dimensions déterminées par les auteurs.

III.4. Prédiction de la qualité vocale par les dimensions perceptives

Le cœur du modèle d'évaluation de la qualité vocale est réalisé à partir d'une relation entre les coordonnées des 46 conditions de dégradation d_{ni} de l'espace perceptif et les 46 notes subjectives d'évaluation de la qualité vocale $MOS\text{-}LQSN_i$.

Nous avons testé différents types de modèle linéaire multiple (avec ou sans interaction). Les modèles prenant en compte les interactions entre les 3 dimensions n'apportent que très peu d'informations supplémentaires à la représentation des notes de qualité vocale, par rapport au modèle linéaire multiple sans interaction.

Le modèle le plus représentatif des notes de qualité vocale $MOS\text{-}LQSN_i$ est le modèle linéaire multiple sans interaction qui est choisi pour constituer le cœur du modèle DESQHI.

Ce modèle est construit à partir d'une régression linéaire multiple entre les notes globales de qualité vocale $MOS\text{-}LQSN_i$ et leurs coordonnées d_{ni} dans l'espace suivant les N dimensions ($N = 3$) :

$$MOS-LQSN_i = c_0 + \sum_{n=1}^{N} c_n . d_{ni} + \varepsilon \qquad \textbf{Eq. III.5}$$

Cette régression linéaire multiple permet de déterminer les coefficients c_n qui relient les coordonnées des conditions de l'espace tridimensionnel à la note d'évaluation de la qualité vocale, à une erreur ε près. Les coefficients c_n sont alors utilisés afin de déterminer les notes dites objectives $MOS\text{-}LQON_i$ (cf. Eq. III.6). Ce modèle fait l'hypothèse que les trois dimensions sont indépendantes entre elles en formant un espace orthogonal. Les corrélations de Pearson et le facteur de significativité entre les positions de chaque point ont été déterminés entre les trois dimensions (cf. Tab. III.3).

	Dim1 Vs Dim2	Dim1 Vs Dim3	Dim2 Vs Dim3
Coefficient de corrélation *r*	*r = -0,22*	*r = -0,05*	*r = -0,01*
Facteur de significativité (*p-value*)	*p = 0,14*	*p = 0,74*	*p = 0,96*

Tab. III.3 Corrélation de Pearson et la valeur de p des conditions de dégradation entre chaque dimension

Ces résultats montrent qu'il n'existe pas de lien significatif entre chaque dimension.

Le modèle linéaire multiple est déterminé à partir des 46 stimuli. La performance du modèle est mesurée en comparant les notes MOS-LQSN, avec les notes MOS-LQON, à l'aide du coefficient de détermination R^2 (cf. Fig. III.21).

$$MOS-LQON_i = 2{,}82 + 0{,}30 \cdot \dim_1 + 0{,}25 \cdot \dim_2 - 0{,}55 \cdot \dim_3 \qquad \textbf{Eq. III.6}$$

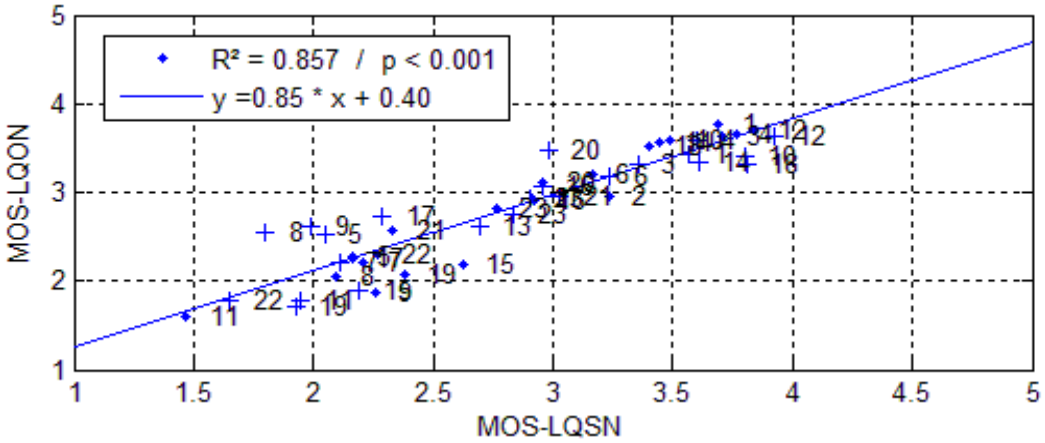


Fig. III.21 Estimation de la qualité vocale par l'analyse tridimensionnelle et performance par rapport aux notes MOS-LQS pour l'espace global (femme (♦) et homme (+))

L'hypothèse de départ concernant le caractère multidimensionnel de l'évaluation de la qualité vocale est donc vérifiée, grâce à la bonne performance obtenue par le coefficient de détermination entre les notes MOS-LQSN issues du test subjectif et les notes MOS-LQON issues de l'analyse multidimensionnelle ($R^2 = 0,86$).

Les coordonnées des conditions ont été centrées et réduites pour chacune des trois dimensions. Le poids de l'influence de chacune des dimensions sur l'évaluation de la qualité vocale est donné par les coefficients multiplicateurs des trois dimensions. La continuité est la dimension altérant le plus l'évaluation de la qualité vocale avec un coefficient $\mathbf{c_3 = -0,55}$. Ensuite, la bruyance et le codage de la parole obtiennent des poids similaires sur l'évaluation de la qualité vocale, respectivement avec les valeurs $\mathbf{c_1 = 0,30}$ et $\mathbf{c_2 = 0,25}$.

Ces trois coefficients de pondération sont étroitement liés à la base sonore utilisée (cf. §.III.1.1). Dans le cas des conditions de dégradation comprenant uniquement du codage de la parole, les notes de qualité vocale sont supérieures à *MOS-LQON = 3,5* (d'après le modèle PESQ [1]). Ce type de dégradation influence donc moins la qualité vocale par rapport aux conditions constituées de pertes de paquets, d'erreurs de bits ou encore de bruit de fond.

Nous avons aussi testé un modèle de catégorie en considérant que la première dimension est représentée par une échelle catégorielle (cf. §.III.3.3.1). Dans ce cas, deux régressions linéaires multiples entre les notes globales de qualité et les deux dernières dimensions ont été déterminées suivant que les stimuli sont bruités ou non. Les performances sont données par le coefficient de corrélation entre les *MOS-LQSN* et *MOS-LQON*.

- Conditions non-bruitées (32) :
  $MOS - LQON = 2,96 + 0,28 \cdot \dim 2 - 0,54 \cdot \dim 3$
  La performance de ce modèle est de $R^2 = 0,89$ (coefficient de détermination)

- Conditions bruitées (14) :
  $MOS - LQON = 2,40 + 0,05 \cdot \dim 2 - 0,37 \cdot \dim 3$
  La performance est de $R^2 = 0,28$ (coefficient de détermination).

- Conditions mixtes bruitées et non-bruitées (46)
  $$MOS - LQON = \begin{cases} 2,40 + 0,05 \cdot \dim 2 - 0,37 \cdot \dim 3 & \textit{pas de bruit de fond} \\ 2,96 + 0,28 \cdot \dim 2 - 0,54 \cdot \dim 3 & \textit{bruit de fond} \end{cases}$$

Ce modèle catégoriel ne représente pas bien les conditions bruitées ($R^2 = 0,28$), contrairement aux conditions non bruitées ($R^2 = 0,89$).

Les expériences spécifiques à la bruyance réalisées par la suite montrent que le niveau sonore des bruits de fond joue un rôle prépondérant lors de l'évaluation de la qualité vocale (cf. Fig. IV.4), et qu'il est bien défini suivant une échelle continue. Ce modèle catégoriel n'est donc pas retenu.

III.5. Structure globale du modèle DESQHI

Ce chapitre a permis de déterminer un espace perceptif tridimensionnel, représentatif de l'évaluation de la qualité vocale des communications téléphoniques actuelles (RTC / RNIS / GSM / IP). La combinaison linéaire des trois dimensions permet de représenter les notes de qualité vocale issues du test perceptif avec une excellente performance ($R^2 = 0,86$ entre les

notes MOS-LQSN et LQON cf. §.III.4). Cette combinaison linéaire représente le cœur du modèle DESQHI.

Afin d'obtenir un modèle automatique, les coordonnées des stimuli sur chacune de ces trois dimensions perceptives doivent maintenant être estimées en utilisant différents types d'indicateurs. Le modèle DESQHI propose d'utiliser des indicateurs paramétriques, des indicateurs basés sur le signal, ou encore des indicateurs hybrides suivant les informations disponibles au point de mesure (cf. Fig. III.22).

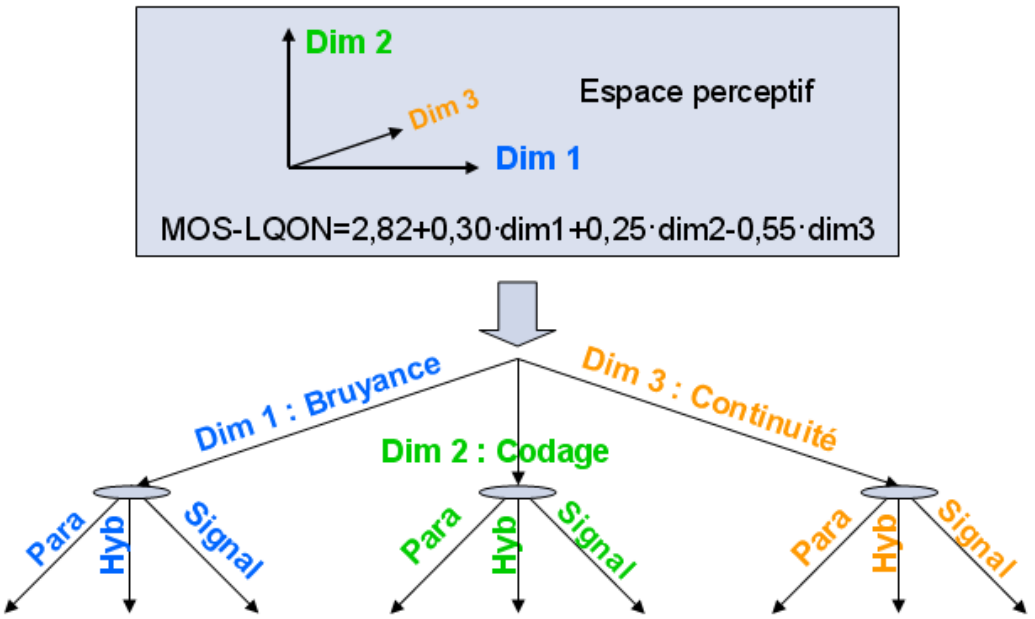


Fig. III.22 Structure globale du modèle DESQHI

La structure globale du modèle DESQHI permet de diagnostiquer la qualité vocale en proposant quatre notes moyennes d'opinion (MOS-LQON). L'une représente la qualité globale du signal vocal déterminée par le modèle linéaire multiple. Les trois autres notes correspondent aux notes relatives à chacune des trois dimensions. Elles sont déterminées en fixant les valeurs des deux autres dimensions non utilisées à leurs valeurs par défaut (cf. §.VI.1.2).

Les deux chapitres suivants présentent la modélisation de chacune des trois dimensions. Tout d'abord le Chapitre IV présente une étude réalisée sur la bruyance, puis le Chapitre V présente les différents indicateurs utilisés pour modéliser le codage de la parole ainsi que la continuité.

Chapitre IV. Modélisation de la bruyance

La dimension bruyance est généralement représentée par le niveau sonore du bruit de fond, comme le font les modèles existants d'évaluation de la qualité vocale.

Les études réalisées sur la bruyance (cf. §.I.4.1) ont permis de remarquer qu'il existe une influence du contenu informationnel du bruit de fond lors de l'estimation de l'intelligibilité. En situation de télécommunication, l'intelligibilité de la parole est souvent bonne car les bruits de fond n'ont généralement pas un niveau sonore suffisant pour engendrer des phénomènes de masquage énergétique sur le signal vocal.

Certains auteurs ont remarqué en plus de l'influence du niveau sonore, une influence du type de bruit de fond lors de différentes analyses psychoacoustiques (Ellermeier [46, 48], Bernex et Barriac [29], Scholz [36], cf. §.I.4.1). Cependant, l'influence du type de bruit de fond est souvent abordée grossièrement ou alors elle concerne d'autres applications que l'évaluation de la qualité vocale d'une communication téléphonique.

Ce chapitre présente l'analyse de la dimension bruyance, en posant notamment l'hypothèse d'une influence de l'origine du bruit de fond (issu du réseau ou provenant de l'environnement du locuteur) lors de l'évaluation de la qualité vocale.

Une première partie présente les tests subjectifs réalisés afin de quantifier l'effet de la bruyance sur l'évaluation de la qualité vocale. Les résultats sont analysés dans la deuxième partie, afin de proposer lors d'une troisième partie, la modélisation de la bruyance qui prend en compte le niveau sonore et le type de bruit de fond, lors de l'évaluation de la qualité vocale. La dernière partie analyse principalement l'influence du type de bruit de fond lors d'une transmission en bande élargie.

IV.1. Tests subjectifs d'évaluation de la qualité vocale bruitée

Des tests subjectifs sont réalisés à partir d'une base sonore construite à cet effet, pour vérifier et évaluer l'influence du type de bruit de fond lors de l'évaluation de la qualité du signal vocal.

IV.1.1. Choix de la méthode

Les modèles existants d'évaluation de la qualité vocale ont généralement été construits à partir de notes MOS issues de tests ACR (UIT P.800 [11]). Le test ACR "Absolute Category Rating" (cf. §.I.2.1) a donc été utilisé pour évaluer la qualité vocale des échantillons de parole bruités afin de pouvoir comparer nos résultats (subjectifs et objectifs) aux modèles existants.

IV.1.2. Stimuli

Les stimuli ont été restitués en écoute monaurale pour plus de réalisme, en utilisant le casque "Sennheiser HD 25" dont l'une des deux oreillettes est rétractable. Il a aussi été envisagé d'utiliser un combiné, mais le casque a été retenu pour deux principales raisons.

La première est que le sujet peut être gêné si on lui demande de tenir le combiné pendant une heure, contrairement au casque. Il sera moins fatigué et sera moins gêné lors de la période de réponse.

La deuxième raison concerne le positionnement du combiné qui diffuse le signal différemment suivant la prise en main de l'appareil (p. ex. la position du haut-parleur, la pression exercée sur l'oreille). L'utilisation du combiné provoquerait un biais dans les résultats.

IV.1.2.1. Description des échantillons de parole

Les échantillons de parole sont issus d'une base sonore phonétiquement équilibrée, enregistrée à France Telecom. Huit doubles-phrases prononcées par quatre locuteurs (deux hommes et deux femmes) ont été retenues. Les deux phrases sont espacées d'un silence de deux secondes, afin d'entendre le bruit de fond. Chaque stimulus a une durée de huit secondes.

Les échantillons de parole ont été rééchantillonnés à 8 kHz, avec une quantification de 16 bits. Le filtre SRI (UIT P.48 [84]) a ensuite été appliqué aux signaux afin de simuler la réponse en fréquence d'un terminal.

Ces signaux ont enfin été égalisés à un niveau de -26 dBov (UIT P.56 [80]). Ces huit échantillons de parole sont alors dégradés par différents signaux de bruits de fond présentés par la suite.

IV.1.2.2. Description des bruits de fond

A. Bruits issus du réseau

La transmission du signal vocal par le réseau RTC peut provoquer l'apparition de bruits de circuit (cf. §.I.1.2.1). Il est aussi possible que des bruits électriques soient générés par les phénomènes d'interférences entre le signal électrique contenant le signal de la parole et la porteuse électrique. Ce bruit est généralement constitué d'un signal harmonique de fréquence fondamentale de 50 Hz.

Les codecs génèrent des bruits de quantification. Ce type de bruit apparaît lorsque le signal analogique est converti en signal numérique, et réciproquement. Ils peuvent être soit stationnaires sur la totalité du signal transmis (zones actives et non-actives), soit modulés en amplitude, c'est-à-dire uniquement présents sur les zones actives du signal de parole. Ces derniers sont généralement appelés bruits sur la parole et sont souvent représentés par les conditions MNRU (Modulated Noise Reference Unit) (cf. §.I.4.2).

B. Bruits issus de sources acoustiques environnant les utilisateurs

Avec l'apparition de la télécommunication mobile, les utilisateurs peuvent être contraints de téléphoner dans n'importe quel environnement bruyant (gare, voiture, ville, restaurant…).

Les bruits d'environnement peuvent être brefs (événements isolés tels qu'un klaxon, aboiement, cri...). On ne tiendra pas compte de ce type de bruit pour simplifier l'étude réalisée. De plus, on suppose qu'ils sont moins gênants lors de la communication comparativement à des bruits présents sur l'ensemble du signal vocal.

La plupart des bruits d'environnement ne sont pas stationnaires. On peut cependant les distinguer en deux classes suivant leurs niveaux de fluctuation temporelle :

- les bruits faiblement fluctuants tels que le bruit intérieur d'une automobile, le bruit intérieur d'un train, le bruit de vent, le bruit de ville, le brouhaha d'un restaurant, le bruit de nature, le bruit de cocktail party...
- les bruits fortement fluctuants tels que la musique, la parole...

Le bruit présent dans l'environnement de l'auditeur n'est pas pris en compte lors de l'évaluation de la qualité vocale par ce même auditeur pour les raisons expliquées au paragraphe I.4.1. Le bruit de fond provenant du côté du locuteur va, quant à lui, être capté par le microphone du terminal d'entrée, et sera donc soumis aux mêmes dégradations que le signal vocal (codecs, pertes de paquets).

C. Choix des bruits de fond

Les trois signaux choisis pour simuler **les bruits de fond issus du réseau** sont décrits ci-dessous :

- **Le bruit rose** présente toutes les fréquences de la bande téléphonique avec une diminution de 3 dB/oct.
- **Le Bruit de Parole Stationnaire** appelé "BPS" est construit de manière à représenter le spectre moyen de la parole, mais avec une enveloppe temporelle constante. Ce bruit ressemble au "bruit marron" (bruit aléatoire avec une diminution de 6 dB/oct).
- **Le bruit électrique** est simulé par un signal harmonique de forme rectangulaire et de fréquence fondamentale 50 Hz.

Les trois signaux retenus pour représenter **les bruits d'environnement** sont :

- **Le bruit de restaurant** qui comporte un mélange de conversation incompréhensible appelé aussi "cocktail party", des bruits de vaisselle et de chaise.
- **Le bruit de ville** qui comporte des bruits d'accélération de voitures et des bruits de klaxon.
- **Le bruit de parole** qui comporte une voix intelligible d'homme enregistrée à partir d'une émission de télévision.

Ces six bruits ont une durée égale à huit secondes et sont filtrés par le système de référence intermédiaire (SRI) [84], pour simuler la réponse en fréquence d'un terminal émetteur en bande étroite (300 Hz – 3,4 kHz). Les spectres de ces six bruits sont représentés sur la Fig. IV.1.

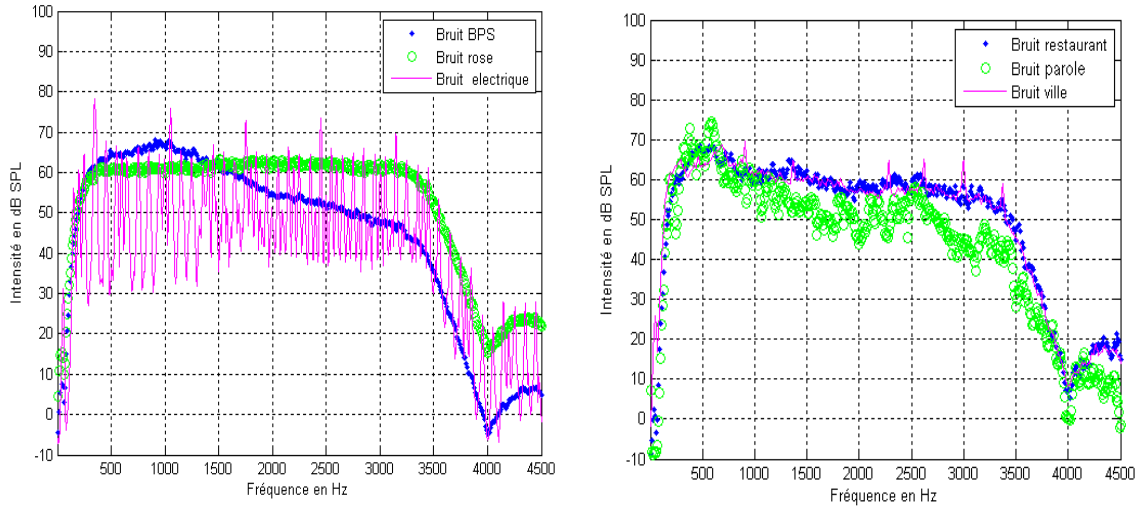


Fig. IV.1 Spectres fréquentiels des 3 bruits issus du réseau à gauche et des 3 bruits d'environnement à droite, utilisés pour les tests subjectifs

D. Choix des niveaux sonores des bruits de fond

Les niveaux sonores du bruit de fond représentatifs des communications actuelles correspondent à des rapports signal sur bruit (RSB) allant de 15 à 40 dB. Sachant que le niveau de la parole est défini par la norme à 79 dB SPL (UIT "Handbook on Telephonometry" [65]), cela suppose d'avoir des niveaux de bruit entre 64 et 39 dB SPL. Ces niveaux sonores sont faibles par rapport à ceux de la parole et n'influencent pas l'intelligibilité.

Notre étude privilégie la diversité des bruits de fond plutôt que les niveaux sonores de restitution, en sélectionnant trois niveaux de bruit de fond :

- RSB = 32 dB → N1 = 47 dB SPL
- RSB = 24 dB → N2 = 55 dB SPL
- RSB = 16 dB → N3 = 63 dB SPL

Le bruit de fond choisi comme référence est le bruit rose stationnaire soumis au filtrage du système de référence intermédiaire (UIT P.48 [84]) et égalisé au niveau de -26 dBov afin de correspondre au niveau sonore des zones actives du signal de la parole. Les trois rapports signal sur bruit ont alors été appliqués à ce bruit rose. Les six bruits de fond sélectionnés ont des caractéristiques spectro-temporelles bien différentes, comme le montre la Fig. IV.1. Une égalisation des niveaux physiques de ces six bruits ne fournirait pas des niveaux sonores équivalents du point de vue subjectif. Les six bruits de fond doivent donc être égalisés en niveau d'isosonie à l'aide d'un test perceptif que nous décrivons dans la partie suivante.

IV.1.2.3. Test préliminaire d'égalisation de la sonie des bruits de fond

Les modèles d'estimation de la sonie proposés par Zwicker [38] ou Moore [49] sont efficaces dans le cas de sons stationnaires. C'est le cas de nos trois bruits issus du réseau (rose, BPS, et électrique), mais pas des trois bruits d'environnement qui sont non stationnaires. D'après les conclusions de Boullet [85], *"en ce qui concerne les sons non stationnaires et impulsionnels, […] les modèles ne permettent pas d'estimer correctement la sonie globale"*.
Il est donc difficile d'égaliser en sonie nos six bruits de fond avec de tels modèles. Le test subjectif reste le moyen le plus efficace afin de réaliser cette égalisation.

Les recherches effectuées par Boullet [85] traitent entre autres de la comparaison de plusieurs méthodologies de tests pour trouver le meilleur moyen d'estimer la sonie de sons non-stationnaires : *"L'ensemble des résultats a permis de conclure que la méthode d'ajustement présente ce meilleur compromis avec une précision de 4,7 phones et une fiabilité de l'ordre de 2 phones. Notons aussi qu'elle dure quatre fois moins longtemps (20 minutes pour 10 sons) que les autres méthodes ayant des écarts- types équivalents."*

La méthode d'ajustement a donc été choisie pour le test préliminaire d'égalisation de la sonie. Elle consiste à présenter en alternance un son de comparaison et le son dont on cherche à mesurer la sonie. Il est demandé aux auditeurs d'ajuster le niveau du son de comparaison, à l'aide d'un curseur.

Pour chacun des trois rapports signal sur bruit, le bruit rose est utilisé comme la référence lors de l'égalisation de la sonie des cinq autres bruits.

Les résultats de ce test préliminaire, détaillés dans l'Annexe G, ont permis de fixer pour chacun des trois niveaux, les cinq gains à appliquer aux cinq bruits pour que les six bruits soient restitués à un niveau d'isosonie équivalent lors du test d'évaluation de la qualité vocale (cf. Fig. IV.2).

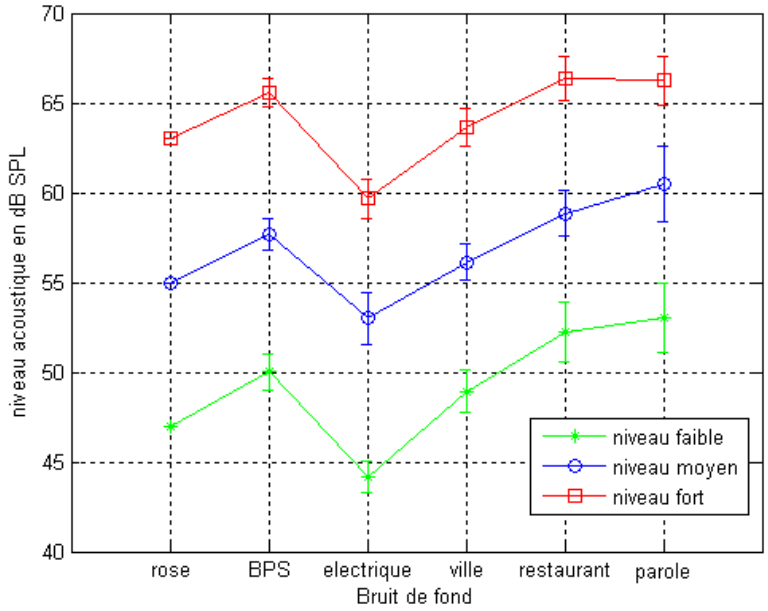


Fig. IV.2 Résultats du test d'égalisation des 3 niveaux d'isosonie et les intervalles de confiance à 95% des 6 bruits de fond selon 20 sujets experts

Les valeurs des trois niveaux d'isosonie sont calculées à partir du bruit rose grâce au modèle de Zwicker [38] qui est performant pour ce type de son. Les trois niveaux d'isosonie sont de 62 phone, 70,5 phone et 78 phone, ou encore de 4,6 sone, 8,2 sone et 14 sone.

Remarque :

Il a été constaté par Ellermeier and al. [46] que l'identification de la source d'un son pouvait dans certains cas influencer l'égalisation de la sonie (cas du son d'alarme d'une horloge, d'un bruit de grésillement, et d'une cloche). Dans le cas du son de cloche, il apparaît que la version dépourvue de signification est perçue plus forte que le son original. L'effet inverse est relevé pour les deux autres. Une étude similaire a aussi été menée par Hellbrück *et al.* [86]. Dans le cas de notre expérience, les sujets nous ont souvent rapporté que le bruit électrique est particulièrement gênant. Nous pouvons faire l'hypothèse que ce bruit a été égalisé à des niveaux inférieurs aux autres à cause de ses caractéristiques.

IV.1.2.4. Construction de la base sonore

La base sonore a été construite à partir des signaux de parole composés de huit phrases, prononcées par quatre locuteurs (2 hommes et 2 femmes) (cf. §.IV.1.2.1), et des six bruits de fond introduits dans le paragraphe IV.1.2.2. La Fig. IV.3 résume les différentes étapes de la construction de la base sonore.

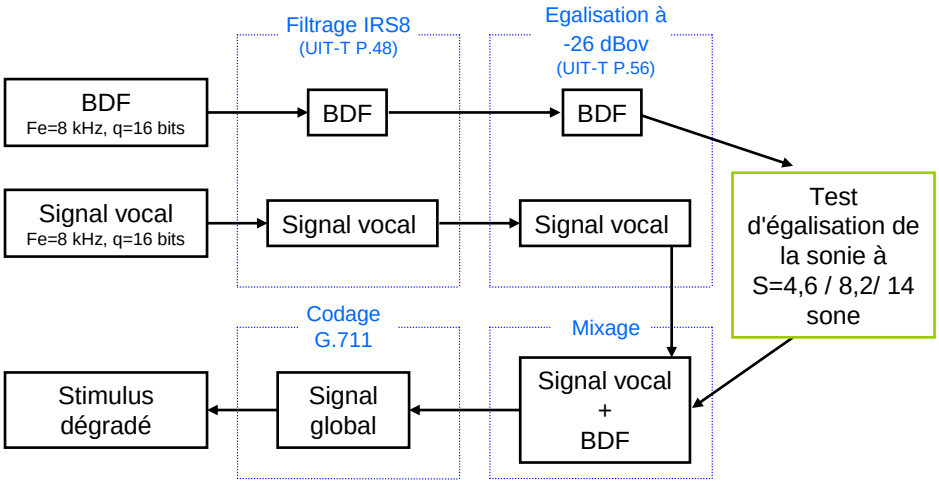


Fig. IV.3 Schéma de construction de la base sonore

Lors de la construction des stimuli, on ne fait pas la distinction entre les bruits issus du réseau et les bruits de l'environnement. On considère qu'ils proviennent tous de l'environnement du locuteur, et qu'ils sont soumis aux mêmes dégradations que le signal de parole. Les signaux de bruit de fond ont donc été soumis au même traitement que le signal vocal (cf. Fig. IV.3). Les signaux de bruit de fond et de parole ont été soumis au filtrage simulant l'utilisation d'un terminal émetteur (UIT P.48 [84]), puis ils ont été égalisés à -26 dBov (UIT P.56 [80]). Le test préliminaire permet d'égaliser les bruits de fond aux trois niveaux d'isosonie. Les signaux globaux ont été obtenus en mixant les six bruits de fond aux huit échantillons de parole pour les trois niveaux d'isosonie. Les signaux globaux ont ensuite été codés et décodés en G.711. Il a été vérifié que le rapport signal à bruit des stimuli n'a pas été influencé lors de cette dernière étape, et que les trois niveaux d'isosonie des bruits de fond correspondent bien à 4,6 sone, 8,2 sone et 14 sone.

La base sonore a aussi été complétée de huit stimuli sans bruit de fond. Pour que ces derniers soient réalistes d'une télécommunication, il a été ajouté un faible niveau de bruit résiduel simulé par un bruit rose de rapport signal sur bruit *RSB = 44 dB*.

IV.1.3. Plan d'expérience

Le test d'évaluation de la qualité vocale par la méthode ACR (cf. §.I.2.1) a été réalisé par 24 sujets naïfs âgés de 19 à 54 ans, avec une répartition équivalente entre les hommes et les femmes. Ce test a été réalisé par groupes de huit sujets avec un ordre aléatoire de présentation des stimuli, et en utilisant les mêmes casques monauraux "Sennheiser HD 25" utilisés lors du test d'égalisation de la sonie.

Trois variables ont été testées lors du test d'évaluation de la qualité vocale (cf. Tab. IV.1) :

- les six bruits de fond utilisés,
- les trois niveaux de diffusion des bruits de fond,
- les huit phrases prononcées par quatre locuteurs.

Les conditions non bruitées (contenant uniquement du bruit résiduel) ont aussi été incluses dans le test d'évaluation de la qualité vocale. En tout, 152 stimuli ont été présentés aux sujets. Chaque stimulus a une durée de huit secondes. Le temps de réponse est fixé à cinq secondes. La durée d'écoute de chaque stimulus est donc de 13 secondes, ce qui représente 33 minutes d'écoute. En intégrant le temps mis pour spécifier les consignes aux sujets, la session

d'apprentissage ainsi que les pauses réalisées toutes les 13 minutes, le test a été réalisé en une heure.

BDF	Niveaux d'isosonie	Phrase & Locuteur	Nb de stimuli	Durée des stimuli	Durée réelle du test
6	3	8	144	31,2 min	1 h
1	1	8	8	1,73 min	

Tab. IV.1 Plan d'expérience du test ACR

IV.2. Résultats du test d'évaluation de la qualité vocale

IV.2.1. Analyse statistique (ANOVA)

Une première analyse statistique a été réalisée sur les résultats du test subjectif, afin de déterminer les variables qui ont un effet significatif sur l'évaluation de la qualité vocale en présence de différents types de bruit de fond. L'analyse de la variance, ANOVA (**AN**alysis **O**f **Va**riance), peut être réalisée seulement si les données ont une distribution normale. Nous avons donc vérifié que la distribution des résultats de qualité vocale donnés par les 24 sujets suit la loi normale grâce au test de Lilliefors. Nous avons vérifié ce point pour les conditions de bruits de fond et de niveaux, en moyennant les résultats sur tous les sujets.

La méthode "**ANOVA répétée à trois facteurs**" a été utilisée, car les résultats obtenus sont **répétés** suivant les 24 sujets, et les conditions de dégradation sont constituées de **trois variables** correspondant aux six types de bruit, aux trois niveaux de diffusion et aux huit phrases. Les résultats de cette analyse sont représentés dans le Tab. IV.2.

Source	SS	Df	MS	F	p
Niveau de bruit	551,9	2	275,9	72,1	0,000***
Type de bruit	73,0	5	14,6	4,1	0,002**
Phrase	115,1	7	16,4	11,1	0,000***
Niveau & Type de bruit	60,4	10	6,0	12,3	0,000***
Niveau & Phrase	10,8	14	0,8	1,8	0,042*
Type de bruit & phrase	28,6	35	0,8	1,7	0,009**
Bruit & Niveau & Phrase	42,5	70	0,6	1,3	0,034*

Tab. IV.2 ANOVA à mesure répétée à trois facteurs sur les résultats du test d'évaluation de la qualité vocale

La valeur de *p* indique le niveau de significativité de la variable testée. Plus la valeur de *p* est petite, plus l'influence de la variable testée joue un rôle significatif sur l'évaluation de la qualité vocale.

Toutes ces variables sont significatives car $p < 0,05$.

* → significatif à $p < 0,05$

** → significatif à $p < 0,01$

*** → significatif à $p < 0,001$

SS : somme des carrés des écarts

Df : degrés de liberté

MS : moyenne au carré

F : valeur de la statistique

p : probabilité de dépassement

D' après les valeurs de *F* et *p*, nous avons relevé les variables influençant le plus l'évaluation de qualité vocale (cf. Tab. IV.2) :

- Le Niveau de bruit (*F = 72,10 / p < 0,001*),
- L'interaction Type de bruit & Niveau (*F = 12,3 / p < 0,001*)
- Les phrases prononcées par différents locuteurs (*F = 11,1 / p < 0,001*)
- Le type de bruit (*F = 4,1 / p < 0,01*)

La prédiction automatique de la qualité vocale des conditions bruitées doit être réalisée en s'affranchissant de l'influence des différents locuteurs et des différentes phrases pour des raisons de faisabilité du modèle. Les analyses présentées par la suite sont donc systématiquement réalisées à partir des notes de qualité vocale *MOS-LQSN* moyennées suivant les 24 sujets, et les 8 phrases. Nous nous intéressons tout d'abord à l'influence du niveau sonore du bruit de fond, puis à l'influence de l'interaction entre le type et le niveau sonore du bruit de fond, lors de l'évaluation de la qualité vocale.

IV.2.2. Influence du niveau sonore du bruit de fond

L'influence du niveau sonore du bruit de fond lors de l'évaluation de la qualité vocale est évaluée en représentant les notes MOS-LQSN moyennées suivant les six bruits en fonction des quatre niveaux sonores exprimés en phone (Fig. IV.4).

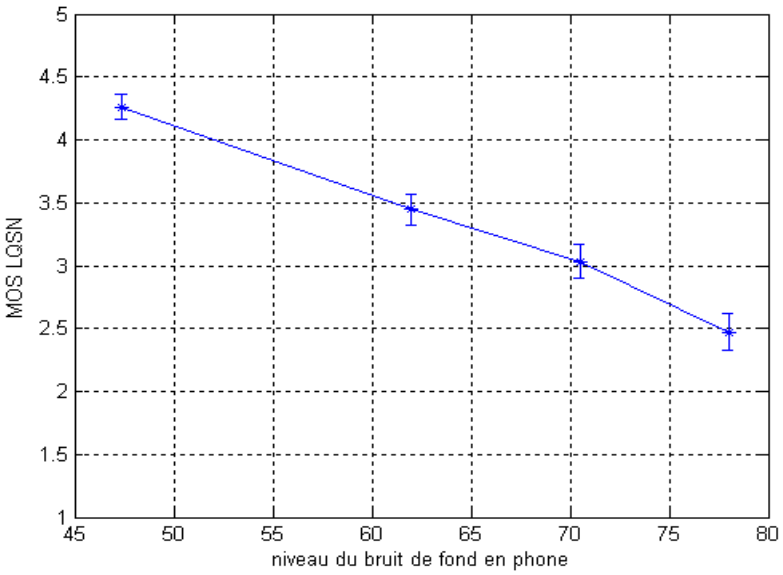


Fig. IV.4 Notes MOS-LQSN moyennées sur les phrases et les 6 BDF, présentées en fonction du niveau sonore des BDF, avec l'intervalle de confiance à 95 %

Les résultats étaient attendus, à savoir que les notes de qualité vocale décroissent quasi-linéairement avec l'augmentation du niveau sonore des bruits de fond.

IV.2.3. Influence de l'interaction entre le type et le niveau sonore du bruit de fond

L'influence de l'interaction entre le type et le niveau sonore du bruit de fond est analysée en représentant les notes MOS-LQSN moyennées suivant les phrases et les sujets pour les quatre niveaux d'isosonie, en distinguant les six bruits de fond (Fig. IV.5).

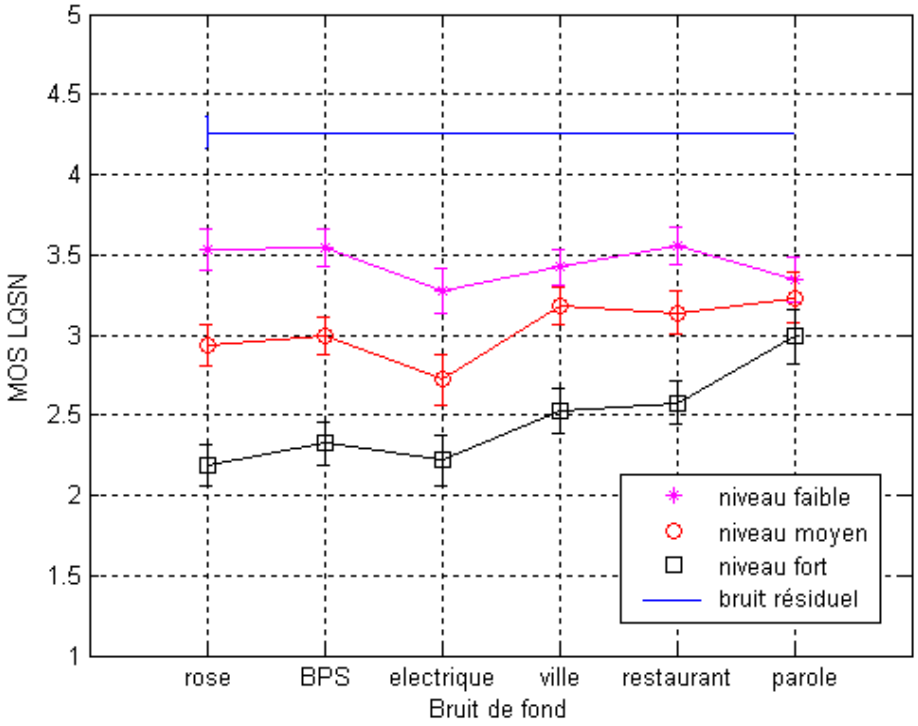


Fig. IV.5 Notes MOS-LQSN moyennées sur les phrases pour les trois niveaux d'isosonie (62 phone, 70,5 phone, et 78 phone) et le niveau du bruit résiduel (47,4 phone) en fonction des 6 BDF, avec l'intervalle de confiance à 95 %

On remarque tout de suite l'effet prépondérant du niveau sonore des bruits de fond lors de l'estimation de la qualité vocale, comme montré au paragraphe IV.2.2. La note maximale obtenue sur l'ensemble des dégradations testées correspond aux conditions comprenant uniquement le bruit résiduel diffusé au niveau de 47,4 phone (*MOS-LQSN = 4,26*), tandis que les notes minimales sont principalement obtenues pour les bruits présentés au niveau fort (78 phone).

Avec l'augmentation du niveau sonore du bruit de fond (niveaux moyen et fort), les différences d'évaluation de la qualité vocale entre les conditions bruitées sont de plus en plus prononcées. Cet effet est analysé à l'aide du test de Student bilatéral suivant chaque paire de conditions bruitées (cf. Annexe E). Ces tests révèlent que pour le niveau faible, il y a peu de différences significatives entre les conditions bruitées (seulement 4 paires de conditions bruitées sont significativement différentes sur 15). Ce phénomène se traduit par l'allure de la courbe qui a tendance à se rapprocher de la droite horizontale (cf. Fig. IV.5). Lorsque le niveau sonore est moyen, nous observons alors 9 différences significatives sur 15, tandis que dans le cas du niveau fort, 12 paires de bruits de fond sur 15 sont significativement différentes lors de l'évaluation de la qualité vocale.

On remarque que les stimuli bruités par un signal de parole intelligible ont des notes de qualité vocale équivalentes selon les trois niveaux d'isosonie (cf. Fig. IV.5). Deux hypothèses pourraient expliquer ce phénomène. La première concerne le masquage informationnel pro-

voqué par le contenu phonétique et lexical du signal de bruit de fond qui pourrait générer une indulgence lors de l'évaluation de la qualité vocale. La deuxième hypothèse concerne la fluctuation temporelle importante du bruit de parole. L'alternance des zones intenses et calmes de ce bruit pourrait faciliter la distinction entre le message cible et le signal de bruit de fond.

Lorsque le niveau sonore du bruit de fond augmente, il apparaît progressivement des différences significatives lors de l'estimation de la qualité vocale entre les bruits provenant de l'environnement du locuteur (ville, restaurant, parole) et les bruits issus du réseau (rose, BPS, électrique) (cf. Fig. IV.5). Les bruits de fond sont analysés suivant ces deux catégories (cf. Fig. IV.6).

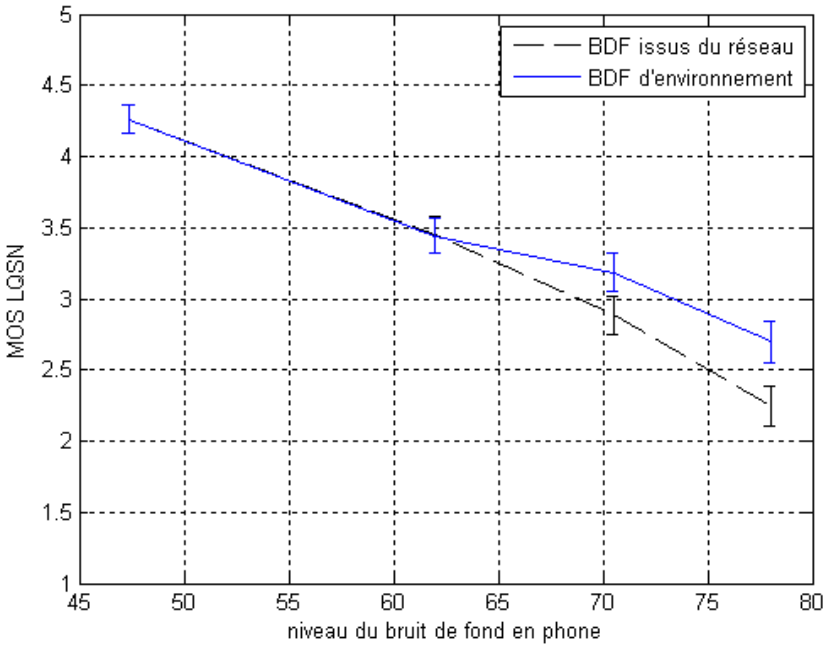


Fig. IV.6 Notes MOS-LQSN moyennées sur les phrases et le type de BDF, en distinguant les 3 bruits d'environnement (ville, restaurant, parole) et les 3 bruits issus du réseau (rose, BPS, électrique), en fonction du niveau sonore des BDF, avec les intervalles de confiance à 95 %

A partir du niveau moyen (70,5 phone), il existe une différence significative lors de l'évaluation de la qualité vocale, entre les bruits issus du réseau et les bruits de l'environnement (Fig. IV.6).

Une deuxième expérience similaire à celle décrite précédemment a été réalisée avec d'autres bruits de fond (bruit rose, bruit de sèche-cheveux, bruit de parole, bruit de musique), pour seulement un niveau d'isosonie (70,5 phone). Les résultats de l'évaluation de la qualité vocale par la méthode ACR ont été moyennés suivant les 24 sujets et les 8 phrases, puis reportés sur la Fig. IV.7.

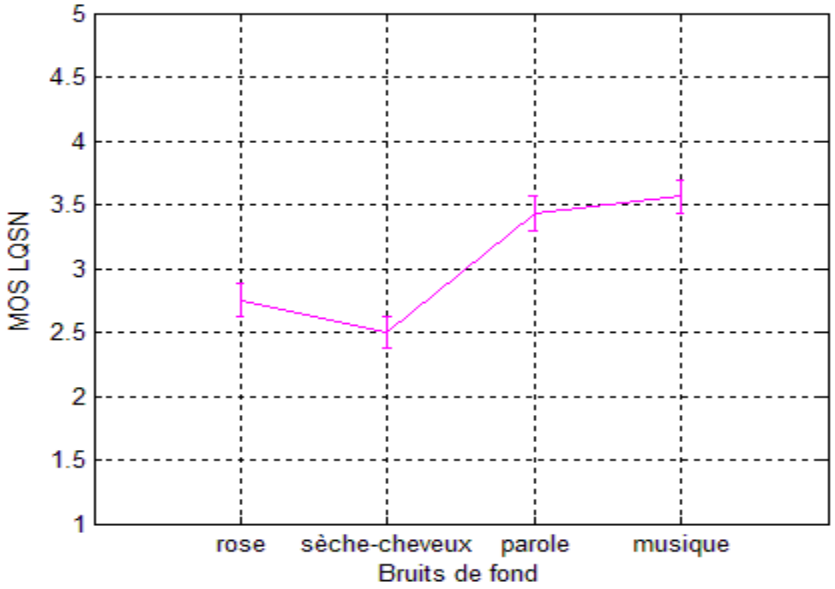


Fig. IV.7 Notes MOS-LQSN moyennées sur les phrases pour le niveau d'isosonie de 70,5 phone en fonction des 6 BDF, avec l'intervalle de confiance à 95 %

Les résultats de ce test subjectif montrent une meilleure évaluation de la qualité vocale pour les stimuli comportant du bruit de musique et de parole, comparativement aux bruits rose et de sèche-cheveux. Ce dernier est considéré de la même manière que le bruit rose lors de l'évaluation de la qualité vocale. A la fin du test d'évaluation de la qualité vocale, il a été demandé de manière informelle aux sujets s'ils avaient reconnu les différents bruits de fond ou non. Ce bruit a été reconnu seulement par quelques sujets (environ 1 personne sur 4).

Les résultats de ces deux expériences montrent que les bruits de fond peuvent être séparés en deux classes suivant qu'ils proviennent de l'environnement du locuteur ou qu'ils sont générés par le réseau de télécommunication (cf. Fig. IV.6 et Fig. IV.7). Nous discutons ci-dessous des deux causes les plus probables de la distinction entre ces deux grandes classes de bruit.

- La principale caractéristique physique qui distingue ces deux types de bruits est le niveau de fluctuation temporelle des signaux. En effet, les bruits de l'environnement sont non-stationnaires, tandis que les bruits issus du réseau sont stationnaires. Cette caractéristique est vérifiée pour la plupart des bruits présents lors d'une télécommunication. L'alternance de zones plus ou moins intenses du bruit de fond pourrait améliorer les résultats de l'évaluation de la qualité vocale, par rapport à un bruit stationnaire.
- Une autre caractéristique qui peut distinguer ces deux types de bruits concerne leurs contenus informationnels. Si le bruit est reconnu par l'auditeur comme provenant d'une source acoustique présente dans l'environnement du locuteur, il n'est pas vraiment assimilé comme une dégradation de la télécommunication mais comme un bruit naturel. Dans le cas contraire, lorsque le bruit n'est pas identifié, il est généralement considéré comme étant une dégradation gênante à la télécommunication. Autrement dit, un auditeur pourrait être indulgent lors de l'évaluation de la qualité vocale lorsque le bruit de fond est reconnu comme étant naturel à la communication. Cette dernière hypothèse peut être comparée à l'étude proposée par Ellermeier *et al.* [48] qui évaluent le jugement de gêne entre des sons originaux et leurs correspondants respectifs dépourvus d'information. Ces auteurs ont vérifié pour certains types de son que les signaux originaux sont moins gênants que leurs correspondants dépourvus d'informations. C'est le cas des sons de machine à café, de porte, de pièce de monnaie, de chasse d'eau, de robinet, de cloche et de verre.

Ces deux hypothèses semblent être reliées entre elles car il est très rare que les bruits issus de sources acoustiques soient stationnaires et que les bruits générés par le réseau soient non-stationnaires. Lorsqu'un bruit d'environnement est stationnaire, il n'est en général pas reconnu par les utilisateurs. Ce bruit est alors assimilé à un bruit issu du réseau comme le montrent les résultats de l'évaluation de la qualité vocale pour les bruits roses et de sèche-cheveux (cf. Fig. IV.7).

Les résultats de ces deux expériences et nos connaissances en matière d'influence du bruit de fond lors de l'évaluation de la qualité vocale permettent de considérer deux sous-classes pour chacune des deux classes déterminées auparavant.
Les quatre classes sont définies suivant l'influence du type de bruit de fond lors de l'évaluation de la qualité vocale.

- Classe 1 → Les bruits *"intelligibles"* provoquent une forte indulgence lors de l'évaluation de la qualité vocale. Ils sont en général constitués d'une seule source ou bien de plusieurs sources corrélées entre elles. Ces bruits sont fortement fluctuants comme par exemple de la musique et de la parole.
- Classe 2 → Les bruits *"d'environnement"* provoquent une indulgence lors de l'évaluation de la qualité vocale. Ils sont en général constitués d'un mélange de plusieurs sources décorrélées entre elles. Ces bruits sont faiblement fluctuants comme des bruits de ville, de restaurant, de nature… Ces bruits peuvent fournir une information supplémentaire à l'auditeur comme par exemple la localisation du locuteur.
- Classe 3 → Les bruits de *"souffle"* provoquent une dégradation de l'évaluation de la qualité vocale. Ces bruits sont stationnaires et proviennent souvent du réseau comme par exemple le bruit de quantification. Dans certains cas, les bruits de souffle peuvent provenir de l'environnement du locuteur comme par exemple du bruit de vent, du bruit d'intérieur de voiture, de sèche-cheveux et en général tous les bruits d'environnement stationnaires qui ne sont pas reconnus par l'auditeur.
- Classe 4 → Les bruits de *"grésillement"* provoquent une forte dégradation de l'évaluation de la qualité vocale. Ces bruits sont stationnaires et proviennent principalement du réseau, comme par exemple le bruit électrique. Les auditeurs leur attribuent souvent les adjectifs *gênant*, *rugueux*, *agressif*.

La Fig. IV.8 suivante présente les résultats du test subjectif d'évaluation de la qualité vocale en faisant la distinction entre ces quatre classes de bruit de fond.

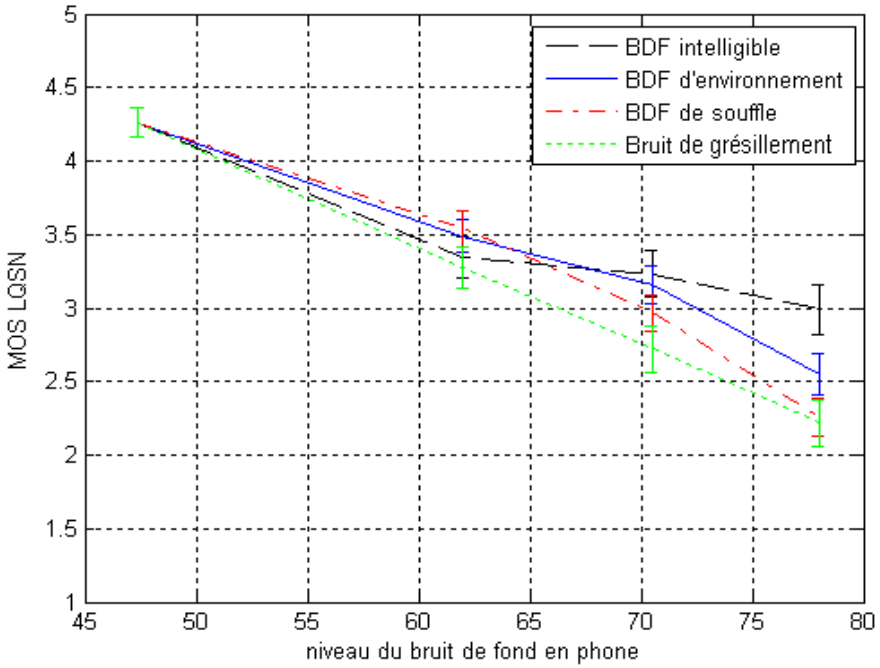


Fig. IV.8 Notes MOS-LQSN moyennées sur les phrases, et les BDF en distinguant les 4 classes de BDF (grésillement, souffle, environnement et intelligible), avec les intervalles de confiance à 95%

L'évaluation de la qualité vocale entre les classes *"intelligible"* et *"d'environnement"* est significativement différente pour le niveau fort (78 phone). Cependant, ce n'est pas le cas pour des niveaux inférieurs de bruit de fond. Dans le cas des classes *"souffle"* et de *"grésillement"*, il apparaît une différence significative pour le niveau moyen (70,5 phone), mais pas pour le niveau fort (78 phone).

Cette catégorisation est tout de même retenue afin de distinguer les différents types de bruit de fond. Elle permettra lors de l'étape de la modélisation de la bruyance, d'améliorer la prédiction de la qualité vocale et de proposer un diagnostic avancé du type de bruit de fond présent lors de la télécommunication (cf. §.VI.1.2).

IV.2.4. Comparaison des résultats avec les modèles objectifs existants

Les résultats de l'évaluation de la qualité vocale sont comparés aux outils de mesure actuels pour estimer l'efficacité des techniques existantes de prédiction de la qualité vocale pour des conditions bruitées. L'indice d'articulation est tout d'abord appliqué aux stimuli de notre base sonore, puis les modèles PESQ, TOSQA, P.563, et le modèle E sont testés.

IV.2.4.1. Indice d'Articulation

L'indice d'articulation (IA) permet d'estimer l'intelligibilité de la parole bruitée à partir des rapports signal sur bruit déterminés pour chacune des bandes de fréquence par tiers d'octave (cf. §.I.4.7).

L'IA est calculé sur un des échantillons de parole (voix de femme à 79 dB SPL) masqué par les six bruits de fond aux trois niveaux d'isosonie définis par le test préliminaire d'égalisation des sonies (cf. Annexe G). La condition comportant le niveau minimal de bruit (47,4 phone) correspond à un IA de 74,5 % compte tenu de la limitation de la largeur de bande (200 Hz – 3,4 kHz au lieu de 200 Hz à 8 kHz) (cf. Fig. IV.9).

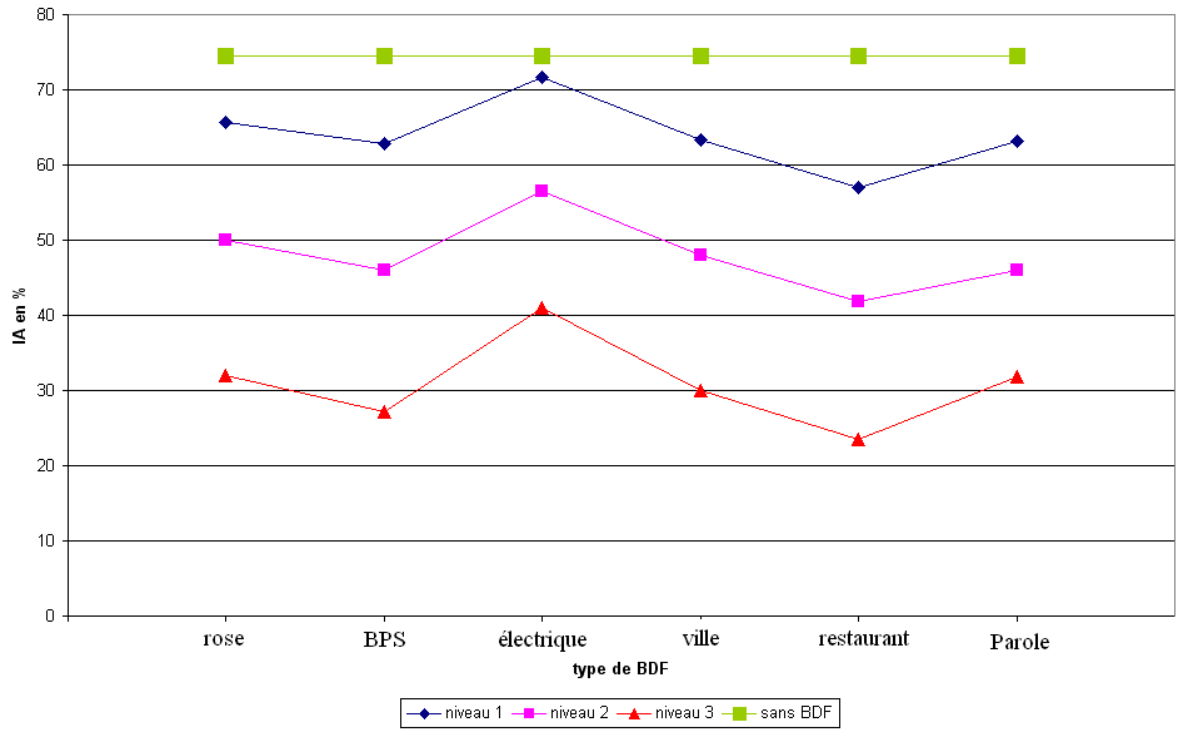


Fig. IV.9 Calcul de l'Indice d'Articulation (IA) en % pour les 6 BDF diffusés aux 3 niveaux d'isosonie et au niveau du bruit résiduel (sans BDF)

Le lien entre l'indice d'articulation et la qualité vocale est présenté dans le Tab. I.5. Nous remarquons que la condition sans bruit de fond correspond à une qualité de communication "*excellente*", le niveau faible de bruit (62 phone) correspond à une qualité "*bonne*", le niveau moyen (70,5 phone) correspond à une qualité "*satisfaisante*", et le niveau fort (78 phone) correspond à une qualité "*mauvaise*". Cela correspond bien à la baisse de la qualité vocale lorsque le niveau sonore du bruit de fond augmente.

En ce qui concerne les différences entre les bruits de fond selon chaque niveau d'isosonie, les résultats de l'IA (cf. Fig. IV.9) ne correspondent pas aux notes MOS-LQSN déterminées par le premier test subjectif (cf. Fig. IV.5). Par contre, ces IA sont cohérents avec les résultats du test préliminaire d'égalisation des niveaux présentés sur la Fig. IV.2. L'IA augmente lorsque le niveau du bruit diminue et inversement. Ce phénomène est vérifié pour tous les bruits de fond aux trois niveaux d'isosonie, excepté dans le cas du bruit de parole.

L'IA est déterminé à partir des rapports signal sur bruit pour chaque bande de tiers d'octave. Cet indicateur ne suffit pas à l'évaluation de la qualité vocale en présence de différents types de bruit de fond.

IV.2.4.2. Comparaison avec les modèles existants d'évaluation de la qualité vocale

Nous avons analysé les différentes corrélations entre les résultats des tests subjectifs et les résultats provenant directement de certains modèles existants présentés dans le paragraphe I.5 (PESQ, TOSQA, G.107 et P.563). Voici un récapitulatif des différentes corrélations entre les notes MOS-LQSN issues du test subjectif et les notes MOS-LQON calculées par les différents modèles.

	r	p
PESQ	*0,91*	$4{,}2 \cdot 10^{-10}$
TOSQA	*0,87*	$2{,}6 \cdot 10^{-8}$
Modèle E	*0,88*	$1{,}2 \cdot 10^{-8}$
P.563	*0,65*	$5{,}52 \cdot 10^{-4}$

Tab. IV.3 Récapitulatif des corrélations de Pearson *r* entre les notes MOS-LQSN issues du test subjectif, et les notes MOS-LQON calculées à partir de 4 modèles existants d'évaluation de la qualité vocale, et les facteurs de significativité *p*

De manière générale, l'outil PESQ semble être efficace pour l'estimation de la qualité vocale de conditions bruitées ($r_{pesq} = 0{,}91$).

Il est remarqué pour tous les modèles testés dans cette partie que plus le niveau sonore des bruits de fond augmente, plus l'estimation de la qualité vocale est approximative. On a observé, par ailleurs, que l'influence du type de bruit de fond est de plus en plus prononcée avec l'augmentation du niveau sonore des bruits. Ces modèles sont limités puisqu'ils ne prennent pas en compte l'influence du type de bruit de fond lors de l'évaluation de la qualité vocale, et principalement lorsque les bruits sont présentés à des niveaux sonores élevés.

La modélisation de la bruyance proposée dans la partie suivante a pour but de prendre en compte le niveau sonore du bruit et l'influence du type de bruit de fond lors de l'évaluation de la qualité vocale, afin de proposer un modèle de bruyance complet, et performant.

IV.3. Construction du modèle de bruyance

Les informations issues des statistiques du réseau ne permettent pas, pour l'instant, de savoir si le signal de parole est bruité ou non. Il n'est donc pas possible de déterminer le niveau sonore ou encore la nature du bruit de fond par des indicateurs paramétriques. La dimension bruyance est donc uniquement représentée par des indicateurs basés sur le signal.

L

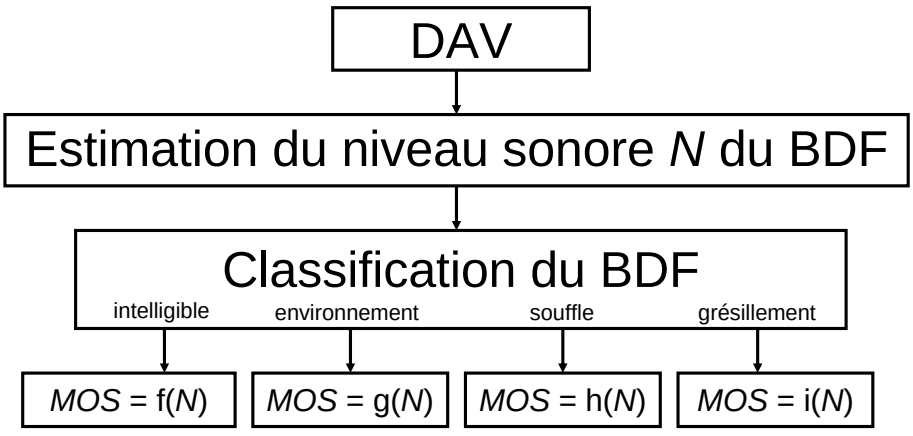


Fig. IV.10 Structure globale du modèle de bruyance

Le premier module est constitué de l'algorithme de détection d'activité vocale (DAV) décrit dans l'annexe B de la recommandation G.729 de l'UIT [60]. Cet algorithme est utilisé afin d'identifier de manière automatique les zones contenant uniquement du bruit de fond.

Le deuxième module correspond à l'estimation du niveau sonore du bruit de fond présent sur le signal de la parole. Plusieurs indicateurs sont disponibles comme le rapport signal sur bruit ou la sonie de Zwicker [38].

Ces deux premiers modules sont utilisés seulement lorsque le modèle global de bruyance est appliqué. Lors de la construction des deux prochains modules, la DAV est réalisée manuellement afin d'être sûr de la bonne détection des zones inactives de la parole, et les niveaux sonores des bruits de fond sont renseignés par leurs valeurs exactes.

Nous avons vérifié que le type de bruit de fond joue un rôle sur l'évaluation de la qualité vocale. Les résultats montrent notamment que les bruits peuvent être séparés suivant quatre classes correspondant aux bruits de *grésillement*, de *souffle*, d'*environnement* et *intelligibles* (cf. §.IV.2.3). Le troisième module présenté dans la partie IV.3.1 suivante consiste à réaliser une classification automatique du type de bruit de fond afin de déterminer la classe correspondante.

Enfin, le quatrième module détaillé dans la partie IV.3.2 présente les quatre fonctions utilisées pour prédire la qualité vocale. Elles sont déterminées grâce à des régressions entre le niveau sonore du bruit de fond et la note d'évaluation de la qualité vocale issue du test subjectif, suivant chacune des quatre classes de bruit de fond.

IV.3.1. Classification automatique du bruit de fond

La classification de bruit de fond dans des signaux audio a déjà fait l'objet de travaux connus. Par exemple, Ma [87] décrit une méthode de classification de bruit reposant sur un modèle de Markov caché (HMM). Selon la méthode décrite, dix bruits d'environnement (bar, plage, rue, bureau, …) peuvent être classifiés en utilisant des coefficients MFCC et leurs dérivées temporelles. En tout, une trentaine d'indicateurs est utilisée pour classifier les bruits avec un pourcentage de bonne classification de 90%. Cependant, cette méthode est très coûteuse en temps de traitement compte tenu du nombre élevé d'indicateurs utilisés.

El-Maleh [88] décrit une technique de classification de bruit de fond dans le contexte de la téléphonie mobile, en particulier quatre types de bruit de fond sont classifiés : rue, cocktail party, usine, bus. Les indicateurs utilisés pour la classification sont les fréquences de raies spectrales (LSF : *line spectral frequencies*). Différents types de classificateurs utilisant ces caractéristiques sont alors comparés, en particulier un arbre de décision (DTC : *decision tree classifier*) et un classificateur quadratique gaussien (QGC : *quadratic gaussian classifier*). Cependant, cette dernière technique utilise encore des indicateurs (LSF) coûteux à calculer.

Ainsi, les techniques de classification de bruits susmentionnées sont complexes et imposent des temps CPU utilisés assez importants, en raison notamment du type et du nombre élevé d'indicateurs requis pour mettre en œuvre la classification.

La partie suivante vise à proposer une technique de classification du bruit de fond présent dans un signal audio, qui soit relativement simple à mettre en œuvre et qui soit peu coûteuse en ressources de traitement comparativement aux méthodes connues. Cette classification pourra alors être appliquée lors de la mesure de la qualité vocale, ou encore lors d'applications de traitement du signal de parole telles que la réduction de bruit ou la reconnaissance vocale. La classification automatique des bruits permet aussi de compléter le diagnostic de la télécommunication en proposant un diagnostic avancé du type de bruit de fond présent lors de la télécommunication (cf. §.VI.1.2).

IV.3.1.1. Base sonore utilisée

Pour obtenir une classification efficace, un nombre important de bruits de fond doit être considéré. Il faut aussi prendre en compte les différentes dégradations survenant lors d'une télécommunication, comme le niveau sonore du bruit, le codage du signal, les pertes de paquets ou la largeur de bande passante.

La base sonore est constituée des six bruits de fond diffusés aux trois niveaux d'isosonie (cf. §.IV.1.2.2.C). Afin que la classification reste pertinente pour d'autres dégradations présentes lors d'une télécommunication, ces six bruits sont aussi présentés avec des dégradations correspondant à 3 % de pertes de paquets et avec les codages G.729 et G.711. Nous obtenons une base sonore constituée de 43 bruits de fond, cependant elle ne contient que six bruits différents (rose, BPS, électrique, ville, restaurant, parole). Cette base sonore a donc été complétée de 48 nouveaux bruits de fond (cf. Annexe F), soumis à six conditions de dégradations. Les dégradations appliquées à ces 48 sons correspondent à des transmissions en bande étroite (300 Hz – 3,4 kHz), mais aussi à des transmissions en bande élargie (50 Hz – 8 kHz) afin que la classification soit applicable à différents types de transmission. Un total de 288 + 43 = 331 stimuli est utilisé pour réaliser la classification.

Chaque stimulus a été échantillonné à 8 kHz, puis filtré suivant la norme du système de référence intermédiaire (UIT SRI [84]), codé et décodé en G.711 ainsi qu'en G.729 dans le cas de la bande étroite, puis chaque son a été échantillonné à 16 kHz, puis filtré (UIT P.341 [89]), puis codé et décodé en G.722 (large bande). Ces trois conditions dégradées ont ensuite été restituées à deux niveaux de restitution (*RSB = 16 et 32*). Chaque bruit dure quatre secondes.

IV.3.1.2. Calcul des indicateurs basés sur le signal

Didiot [62] présente et calcule plusieurs indicateurs à partir du signal pour la classification automatique de la musique et de la parole. En tout, quatre types d'indicateurs sont référencés : *les indicateurs temporels* basés sur la représentation temps/puissance des sons, les *indicateurs fréquentiels* basés sur la représentation fréquence/puissance des sons, *les indicateurs mixtes* basés sur les deux descriptions précédentes ainsi que *les indicateurs cepstraux*. Didiot conclut que les indicateurs les plus efficaces à la segmentation de la parole/musique sont les coefficients MFCC (Mel Frequency Cepstrum Coefficients).

Istrate [63] et Istrate et Vacher [90] effectuent des recherches sur la détection et la reconnaissance des sons pour la télésurveillance médicale. Ils se servent de différents types de bruits comme la parole ou des bruits de la vie courante (cris / applaudissement / ronflements / éternuements / perforeuse / vaisselle / sèche-cheveux /…). Dans le cas de la classification parole / autres sons, les auteurs utilisent 16 coefficients MFCC et l'énergie normalisée pour obtenir des performances de classification assez bonnes (4,5 % de taux d'erreurs de classification environ).

Dans le cas de la classification des sons de la vie courante (déterminée par sept classes), les auteurs utilisent plusieurs sortes de combinaisons d'indicateurs, cependant la plus efficace semble être la combinaison des 16 coefficients MFCC couplés avec le nombre de passages par zéro (ZRC), le point spectral de coupure ainsi que le centre de gravité spectrale. Avec ces indicateurs, les auteurs obtiennent des taux d'erreurs de classification d'environ 10%.

Dans le cas de la parole, les coefficients MFCC sont utilisés avec le modèle de mélange de distributions de Gauss (ou "Gaussian Mixture Model" GMM) pour la classification et reconnaissance du locuteur. Mais lorsque la classification doit s'étendre à une base sonore comportant différents types de bruit, il apparaît que l'ajout d'indicateurs acoustiques semble améliorer l'efficacité de la classification (Istrate [63]).

Dans notre étude, il est intéressant d'explorer certains de ces indicateurs, en considérant une grande diversité de bruits de fond représentatifs de la réalité (bruit de parole, cocktail party, musique, environnement, usine, nature, circuit…). Ces indicateurs devront être choisis en considérant que les signaux traités sont issus d'une communication téléphonique, impliquant diverses dégradations (largeur de bande limitée, codecs, pertes de paquets, niveau sonore faible par rapport au niveau de la parole).

Les indicateurs les plus pertinents pour la classification de bruit de diverses natures sont décrits et calculés pour les 331 bruits de la base sonore. Ils correspondent aux indicateurs temporels, aux indicateurs fréquentiels, et aux indicateurs cepstraux.

A. Les indicateurs temporels

La corrélation du signal est un indicateur utilisant le coefficient de corrélation de Bravais-Pearson entre le signal entier, et ce même signal, mais décalé d'un échantillon. Il est défini comme

$$Corr\_signal = \frac{\sum_{i=1}^{N-1}(x_i - \langle x \rangle).(x_{i+1} - \langle x \rangle)}{\sqrt{\sum_{i=1}^{N-1}(x_i - \langle x \rangle)^2}.\sqrt{\sum_{i=1}^{N-1}(x_{i+1} - \langle x \rangle)^2}}, \qquad \textbf{Eq. IV.1}$$

avec:
N : le nombre d'échantillons du signal x
x_i : les valeurs du signal (de *1* à *N-1*),
x_{i+1} : les valeurs du même signal décalées d'un échantillon (de *2* à *N*)
$<x>$: les moyennes respectives des deux vecteurs

Ce coefficient donne une estimation de la continuité du signal temporel : lorsqu'un son est généré de manière aléatoire (bruit blanc ou bruit rose), les échantillons *n* et *n+1* peuvent être éloignés l'un de l'autre car le signal présente des sauts d'amplitude complètement décorrélés entre le temps *n* et *n+1*. Dans ce cas, la corrélation entre ces deux signaux (les mêmes décalés de 1 échantillon) nous donne un indicateur de corrélation faible (proche de zéro).

Dans le cas d'un son harmonique de fréquence pure, les échantillons *n* et *n+1* sont assez proches. La corrélation entre les deux signaux (les mêmes décalés de 1 échantillon) donnera alors une valeur se rapprochant de *1*.

Cet indice de corrélation donne une idée de la nature harmonique ou non du son. Par exemple, un son de parole aura une valeur de corrélation nettement plus élevée ($Corr\_signal \rightarrow 1$) qu'un son généré aléatoirement comme un bruit rose ou un bruit blanc ($Corr\_signal \rightarrow 0$).

Le taux de passage par zéro (aussi appelé ZCR *« Zero Crossing Rate* ») donne principalement une idée sur le voisement du signal de la parole. Le voisement est une propriété de certains sons de la parole. Un son est dit voisé si sa production s'accompagne d'une vibration des cordes vocales, et non voisé dans le cas contraire. Le taux de passage par zéros représente le nombre de fois où le signal, dans sa représentation temporelle, passe par la valeur centrale de l'amplitude (généralement zéro) sur une fenêtre temporelle définie par la longueur des trames. Chaque trame est composée de 512 échantillons, avec un recouvrement suivant les trames successives de 256 échantillons. Pour une fréquence d'échantillonnage de 8 kHz, chaque trame a une durée de 64 ms, avec un recouvrement de 32 ms.

Voici l'équation permettant de comptabiliser ce taux de passage par zéro :

$$ZCR(tr) = \sum_{i=1}^{N} \left(\frac{|\mathrm{sgn}(x_i(tr)) - \mathrm{sgn}(x_{i-1}(tr))|}{2} \right), \quad \textbf{Eq. IV.2}$$

avec $x_i(tr)$ le i^{ieme} échantillon de la trame *tr* et *N* le nombre total d'échantillons de la trame *tr*.

Le ZRC aura une valeur élevée dans les zones non-voisées et faible dans les zones voisées. Dans le cas de la parole qui est constituée d'une alternance entre sons voisés et non voisés, le ZRC aura une variation importante. **La variation du taux de passage par zéro** est défini par l'écart type des valeurs du taux de passage par zéro suivant la totalité des trames composant le bruit de fond. Dans le cas de la musique, cette variation entre sons voisés et non voisés sera moins importante, et lorsque les sons sont stationnaires, cette différence de ZRC au cours du temps sera minime.

La variation du niveau sonore *Vn* représente l'indicateur de non-stationnarité du son. Il est défini par l'écart-type des valeurs d'énergie *P(tr)* de toutes les trames *tr* du signal.

$$P(tr) = \frac{1}{L_{tr}} \sum_{i=1}^{L_{tr}} x_i^2 \quad \textbf{Eq. IV.3}$$

Plus le son est non-stationnaire, plus l'indicateur sera élevé.

B. Les indicateurs fréquentiels

Les indicateurs fréquentiels sont calculés à partir de la DSP[13] du signal. Ces indicateurs caractérisent l'enveloppe spectrale du signal. Ils permettent ainsi de capter le contenu fréquentiel du signal à un moment donné, comme par exemple les formants, les harmoniques, etc… Les indicateurs décrits dans cette partie sont déterminés par trames de 256 échantillons, correspondant à une durée de 32 ms pour une fréquence d'échantillonnage de 8 kHz. Il n'y a pas de recouvrement des trames.

Le centre de gravité spectrale est défini dans le paragraphe I.4.9. Le centre de gravité spectrale donne une idée de la répartition fréquentielle du signal considéré.

Le flux spectral (SF) aussi appelé la variation de l'amplitude du spectre est une mesure permettant d'estimer la vitesse de changement du spectre de puissance d'une trame *tr*. Cette mesure est calculée à partir de la corrélation croisée normalisée entre deux amplitudes successives du spectre $a_k(tr\text{-}1)$ et $a_k(tr)$.

$$SF_{tr} = 1 - \frac{\sum_k a_k(tr-1).a_k(tr)}{\sqrt{\sum_k a_k(tr-1)^2}\sqrt{\sum_k a_k(tr)^2}}, \quad \textbf{Eq. IV.4}$$

avec : *k* représentant les différentes composantes fréquentielles, et *tr* les trames successives sans recouvrement composées de 256 échantillons chacune.
En d'autres termes, la valeur SF_{tr} correspond à la différence d'amplitude du vecteur spectral entre deux trames successives. Elle est proche de *0* si les spectres successifs sont similaires, et

[13] La Densité Spectrale de Puissance (DSP) d'un signal est issue de la transformée de Fourier (module du carré de la transformée d'un signal, sur le temps d'intégration).

de *1* pour des spectres successifs très différents. Cette valeur est élevée pour la musique, car le contenu varie fortement d'une trame à l'autre. Pour la parole, avec l'alternance de périodes de stabilité (voyelle) et de transitions (consonne-voyelle), cette mesure prend des valeurs très différentes et varie fortement au cours d'une phrase.

Les valeurs du flux spectral obtenues pour toutes les trames du signal sont moyennées, afin d'avoir une valeur de flux spectral moyenne sur l'ensemble du signal à analyser.

Le point spectral de coupure à 95% ou "Spectral Rolloff Point" est présenté dans le paragraphe I.4.9. Cet indicateur permet de caractériser l'alternance parole voisée, parole non voisée. L'énergie est concentrée dans les basses fréquences pour les voyelles, d'où une petite valeur du point spectral de coupure, alors qu'elle se situe davantage dans les hautes fréquences pour les fricatives et nous avons par conséquent une valeur plus élevée du point spectral de coupure. Pour la musique dont l'énergie est plus uniformément répartie sur toutes les bandes de fréquences, cette mesure ne varie que très faiblement.

Dans notre application, le point spectral de coupure est défini pour chaque trame. L'indicateur utilisé est ensuite défini comme étant la variation des composantes obtenues sur toutes les trames du bruit de fond.

C. Les indicateurs cepstraux

Ce type d'indicateur est utilisé pour représenter des signaux constitués de parole. Ils sont brièvement présentés dans le paragraphe I.4.6. Pour plus d'information, on peut se reporter aux références suivantes : D'Alessandro et Demars [61], Didiot [62], et Istrate [63]. Les MFCC (Mel Frequency Cepstrum Coefficients) sont généralement utilisés pour classifier les sons composés de parole.

IV.3.1.3. Apprentissage

Le module d'apprentissage de la classification utilise les indicateurs détaillés précédemment afin de déterminer ceux qui sont les plus pertinents pour la classification automatique des bruits dans l'une des quatre classes définies dans le paragraphe IV.2.3. Les indicateurs utilisés sont les suivants :

- La corrélation du signal
- La variation du nombre de passages par zéro
- La variation du niveau sonore
- Le centre de gravité spectrale
- Le flux spectral
- Le point spectral de coupure
- Dix Coefficients MFCC

Ces 16 indicateurs sont calculés pour les 331 bruits de fond. Parallèlement, ces bruits sont renseignés de manière empirique, par l'une des quatre classes de bruit définies dans le paragraphe IV.2.3 (bruits *intelligibles*, bruits *d'environnement*, bruits *de souffle*, et bruits *de grésillement*).

Certains des bruits, parfaitement identifiables à l'origine, perdent toutes leurs significations avec les dégradations introduites. Nous remarquons ainsi une meilleure identification du bruit lorsqu'il est présenté en large bande par rapport à une transmission en bande étroite. Dans le cas de bruit de vent ou de nature, la distinction des classes "*environnement*" et "*souffle*" peut être confuse.

La classification est réalisée à partir d'un algorithme proposé par Breiman [91]. Cet outil permet de déterminer un arbre de décision qui classifie les données avec le moins d'erreurs

possible en utilisant un nombre d'indicateurs fixé. Les paramètres d'entrée de cet algorithme sont constitués des 16 x 331 indicateurs et des classes associées aux 331 bruits de fond. Cet algorithme propose un arbre de décision afin de représenter ces quatre classes à l'aide des indicateurs les plus pertinents. L'arbre de décision optimal est choisi grâce à un compromis entre le nombre d'indicateurs utilisés (pour une application en temps réel), et le pourcentage de bonnes classifications des bruits de fond.

Lors de l'apprentissage, l'arbre de décision est obtenu à partir de 250 bruits de fond choisis aléatoirement sur les 331. Les 81 bruits restants sont utilisés ensuite dans une phase de validation de la classification. L'arbre de décision retenu est représenté sur la Fig. IV.11.

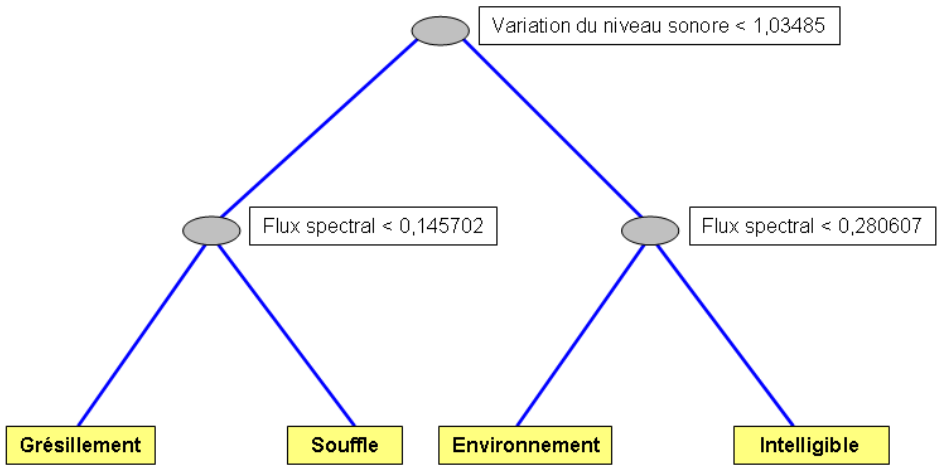


Fig. IV.11 Arbre de classification des BDF

La classification proposée utilise seulement deux indicateurs pour classer les 331 bruits de fond de l'apprentissage dans les quatre classes correspondantes, avec une proportion correctement classifiée de 86,2%.

Le premier **indicateur est temporel** et correspond à **la variation du niveau sonore**. Si la valeur de cet indicateur est inférieure à *1,035*, le bruit est considéré comme stationnaire et il a de fortes probabilités d'être généré par le réseau de télécommunication. Si la valeur de l'indicateur temporel est supérieure à *1,035*, le bruit est considéré comme non-stationnaire. Il a alors de grandes probabilités de provenir de l'environnement du locuteur.

Les deux sous-classes sont ensuite distinguées grâce à un **indicateur fréquentiel : le flux spectral**. Dans le cas des bruits stationnaires, si la valeur de l'indicateur fréquentiel est inférieure à *0,146*, le bruit appartient à la classe *grésillement*. Dans le cas contraire, le bruit est considéré comme du bruit de *souffle*.

Dans le cas des bruits non-stationnaires, lorsque la valeur de l'indicateur fréquentiel est inférieure à *0,281*, le bruit appartient à la classe *environnement*, sinon, le bruit est considéré comme *intelligible*.

IV.3.1.4. Validation

Les performances de l'arbre de décision sont testées sur les 80 stimuli non utilisés lors de l'étape d'apprentissage.

Le Tab. IV.4 récapitule les proportions correctement classifiées pour quatre combinaisons différentes de la base sonore :

Base sonore	250 BDF apprentissage	80 BDF inconnue	330 BDF totale
proportions correctement classifiées (en %)	86,2	91,5	87,3

Tab. IV.4 Proportions correctement classifiées en % pour 3 combinaisons différentes de la base sonore

L'arbre de décision présente une proportion correctement classifiée de 87,3 % sur la totalité de la base sonore. Dans le cas des 80 bruits non utilisés lors de l'apprentissage, la proportion est de 91,5 %.

Plus précisément, voici les pourcentages de bonne classification pour chaque classe, dans le cas de l'emploi de toute la base sonore :

- 100% pour la classe "grésillement",
- 96,4% pour la classe "souffle",
- 79,2% pour la classe "environnement",
- 95,9% pour la classe "intelligible".

Il apparaît que la classe "environnement" obtient des proportions correctement classifiées plus faibles que les autres classes. Ce problème est dû à la différenciation de certains bruits de "souffle" ou "d'environnement", comme les bruits de vent ou de sèche-cheveux par exemple.

IV.3.1.5. Comparaison avec la classification d'Istrate [63]

Voici les résultats des proportions correctement classifiées, présentés par Istrate [63], pour différentes combinaisons d'indicateurs :

Type des paramètres	Nombre	proportions correctement classifiées (en %)
16MFCC+Energie+ZCR+RF+Centroïde+Δ+ΔΔ	60	92.9
16MFCC+Energie+ Δ+ΔΔ	51	89.3
16MFCC+Energie+ZCR+RF+Centroïde+Δ	40	90
16MFCC+Energie+ Δ	34	87.3
16MFCC+Energie+ZCR+RF+Centroïde	**20**	**89.9**
16MFCC+Energie	17	86

Tab. IV.5 Proportions correctement classifiées en % selon les résultats de la thèse d'Istrate [63]

La combinaison retenue par Istrate [63] est celle en gras. Elle est composée de 16 coefficients MFCC, l'énergie, le taux de passage par zéro, le point spectral de coupure, et le centroïde (brillance). En tout, 20 coefficients sont nécessaires à la classification.

La comparaison de ces résultats avec les nôtres sont à prendre avec précaution, car les protocoles sont différents :

- La base sonore n'est pas la même
- Le type de protocole utilisé pour choisir la base sonore de l'apprentissage est différent (type de "leave-one-out" spécifique au corpus de taille réduite). Ce protocole stipule qu'à tout instant la base d'apprentissage comprend tous les fichiers sons sauf un, qui est le fichier test Istrate [63].
- La classification est réalisée avec des GMM (Gaussian Mixture Model), contrairement à l'utilisation de l'arbre de décision dans notre cas.

Avec seulement deux descripteurs simples à déterminer, l'arbre de décision proposé permet d'obtenir une proportion correctement classifiée de 87.3% sur la base sonore complète, contre vingt descripteurs pour un score de 89.9 % suivant les résultats d'Istrate. La différence de précision est assez faible pour un temps CPU utilisé jouant en notre faveur.

IV.3.2. Prédiction de la qualité vocale en fonction du niveau sonore et de la classe du bruit de fond

Les résultats du test subjectif présentés dans la partie IV.2.3 ont montré que l'évaluation de la qualité vocale est influencée par le niveau sonore et le type de bruit de fond.

Les bruits de fond de chaque stimulus sont identifiés par la classe correspondante (*intelligible*, *environnement*, *souffle* ou *grésillement*). Pour chacune des quatre classes, une analyse de régression est menée entre les résultats de l'évaluation de la qualité vocale et les niveaux sonores des bruits de fond. Le modèle de bruyance est construit selon deux représentations du niveau sonore : la sonie et le rapport signal sur bruit (RSB).

Les quatre niveaux d'isosonie déterminés lors du test préliminaire d'égalisation de la sonie des bruits de fond (1,67 sone, 4,6 sone, 8,2 sone et 14 sone cf. §.IV.1.2.3 et Annexe G) sont utilisés pour prédire la qualité vocale. Dans ce cas, les régressions obtenues sont logarithmiques (cf. Fig. IV.12). La même analyse est effectuée à partir des quatre niveaux sonores des bruits de fond exprimés par les rapports signal sur bruit issus du test préliminaire d'égalisation de la sonie (Annexe G). Cette fois les quatre régressions obtenues sont linéaires (cf. Fig. IV.13).

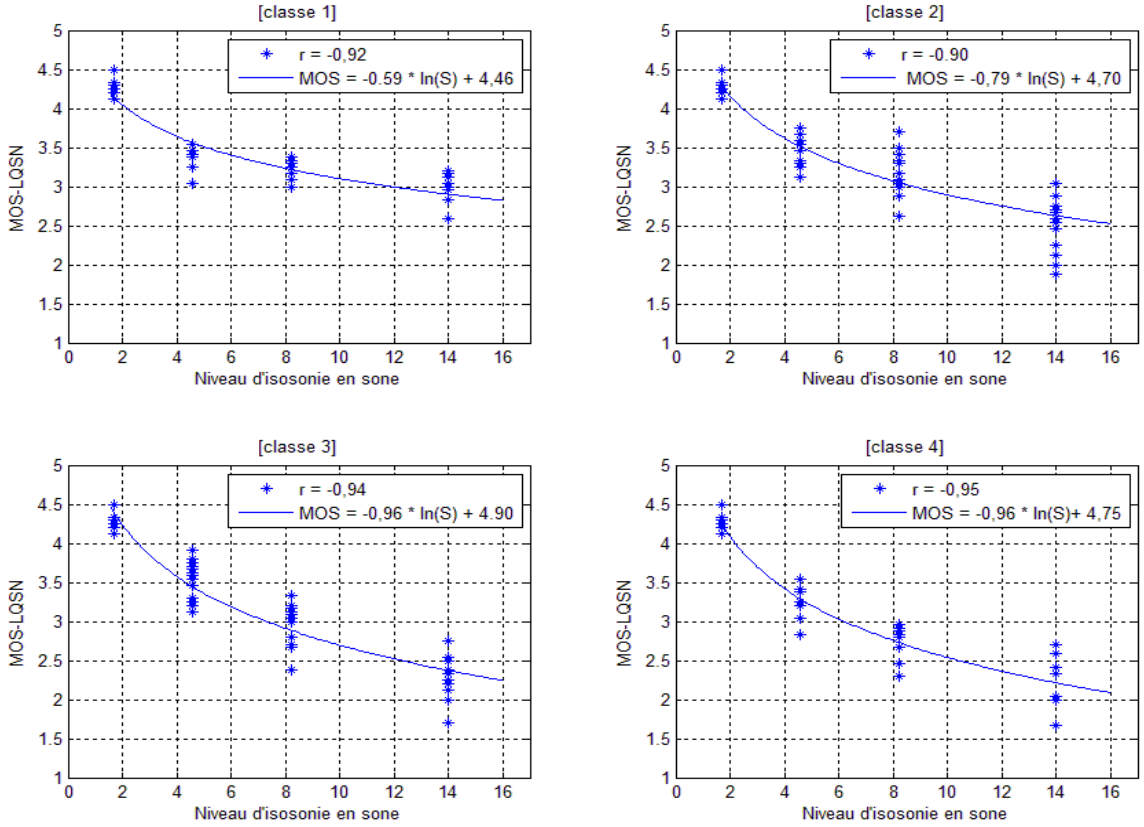


Fig. IV.12 Régressions logarithmiques entre les notes d'évaluation de la qualité vocale issues du test subjectif (MOS-LQSN) et les niveaux d'isosonie en sone pour les 4 classes de BDF (respectivement *intelligible, environnement, souffle* et *grésillement*)

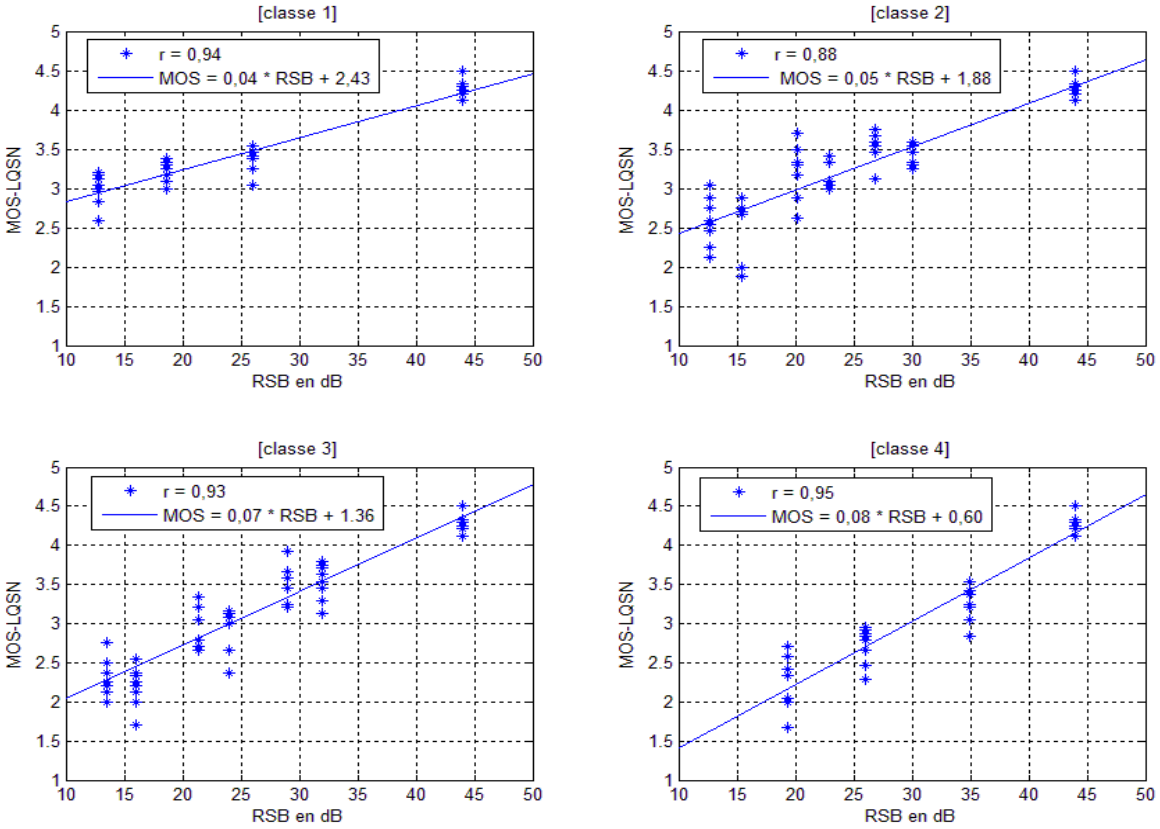


Fig. IV.13 Régressions linéaires entre les notes d'évaluation de la qualité vocale issues du test subjectif (MOS-LQSN) et les RSB en dB pour les 4 classes de BDF (respectivement *intelligible, environnement, souffle* et *grésillement*)

Remarques :

- Les quatre classes ne comportent pas le même nombre de stimuli. En effet, il n'y a qu'un seul bruit de fond dans la classe "intelligible" (parole) ainsi qu'un bruit de fond dans la classe "grésillement" (électrique), contre deux dans chacune des classes "souffle" (rose et BPS) et "environnement" (ville et restaurant). Cependant, l'équilibre est rétabli entre les bruits de fond issus de l'environnement du locuteur (classes 1 et 2) et les bruits issus du réseau (classes 3 et 4).
- Dans le cas des régressions avec la sonie, si les niveaux sonores des bruits de fond avaient été exprimés en phone, les régressions auraient été linéaires.

Les performances de la prédiction de la qualité vocale sont similaires pour les deux indicateurs (RSB et sonie) comme le montrent les coefficients de corrélation *r* (cf. Fig. IV.12 et Fig. IV.13). Lors de l'application du modèle de bruyance, l'indicateur RSB est nettement plus simple et plus rapide à calculer, ce qui constitue un avantage certain par rapport à l'indicateur de la sonie comme par exemple celui proposé par Zwicker [38].

Le Tab. IV.6 récapitule les relations de la prédiction de la qualité vocale suivant la classe du bruit de fond et suivant le niveau sonore (niveau d'isosonie *S* ou rapport signal sur bruit *RSB*).

Classe 1 **Bruit intelligible**	**Classe 2** **Bruit d'environnement**	**Classe 3** **Bruit de souffle**	**Classe 4** **Bruit de grésillement**
$MOS = 4{,}46 - 0{,}59 \cdot \ln(S)$	$MOS = 4{,}70 - 0{,}79 \cdot \ln(S)$	$MOS = 4{,}90 - 0{,}96 \cdot \ln(S)$	$MOS = 4{,}75 - 0{,}96 \cdot \ln(S)$
$MOS = 2{,}43 + 0{,}04 \cdot RSB$	$MOS = 1{,}88 + 0{,}05 \cdot RSB$	$MOS = 1{,}36 + 0{,}07 \cdot RSB$	$MOS = 0{,}60 + 0{,}08 \cdot RSB$

Tab. IV.6 Prédiction de la qualité vocale *MOS* en fonction du niveau d'isosonie *S* exprimé en sone et en fonction du rapport signal sur bruit *RSB* en dB, pour chacune des 4 classes de bruit de fond

IV.4. Performance et validation du modèle de bruyance

Le modèle de bruyance est appliqué à la base sonore utilisée lors de la construction du modèle. Cette base sonore comprend 152 stimuli composés de 19 conditions de dégradations répétées suivant 8 phrases prononcées par 4 locuteurs (cf. §.IV.1.2).

Une deuxième base sonore inconnue du modèle est ensuite appliquée au modèle de bruyance. Elle est constituée de différents types de bruits de fond, et de dégradations relatives à l'utilisation d'algorithmes de débruitage (Gautier-turbin et Gros [92])

IV.4.1. Application à la base sonore connue du modèle

Le modèle de bruyance utilise les quatre modules présentés sur la Fig. IV.10 :

- La DAV (Détection d'Activité Vocale),
- La sonie calculée par le modèle de Zwicker [38] ou le rapport signal sur bruit déjà calculé lors du module de DAV
- La classification automatique du type de bruit de fond (cf. §.IV.3.1)
- La prédiction de la qualité vocale par les régressions reliant le niveau sonore des bruits de fond aux notes de qualité vocale (cf. §.IV.3.2)

Les performances de ces modules ont été évaluées séparément et en combinaison, afin de mesurer la fiabilité du modèle de bruyance, à partir de la base sonore utilisée pour construire le modèle (cf. §.IV.1.2).

Tout d'abord, le modèle est utilisé en fixant les trois premiers modules manuellement afin de tester uniquement la performance des quatre régressions (régressions logarithmiques pour la sonie et régressions linéaires pour le RSB). La performance du modèle est représentée par la corrélation de Pearson entre les notes MOS-LQSN issues du test subjectif et les notes MOS-LQON calculées. Ces deux types de notes sont moyennés suivant les phrases. Pour les deux types d'indicateurs du niveau sonore, la performance mesurée est de $r = 0,98$, $p < 0,001$ (cf. Fig. IV.14). Ce résultat montre que les quatre régressions logarithmiques suivant le niveau d'isosonie et les quatre régressions linéaires suivant le rapport signal sur bruit sont bien représentatives des résultats du test d'évaluation de la qualité vocale.

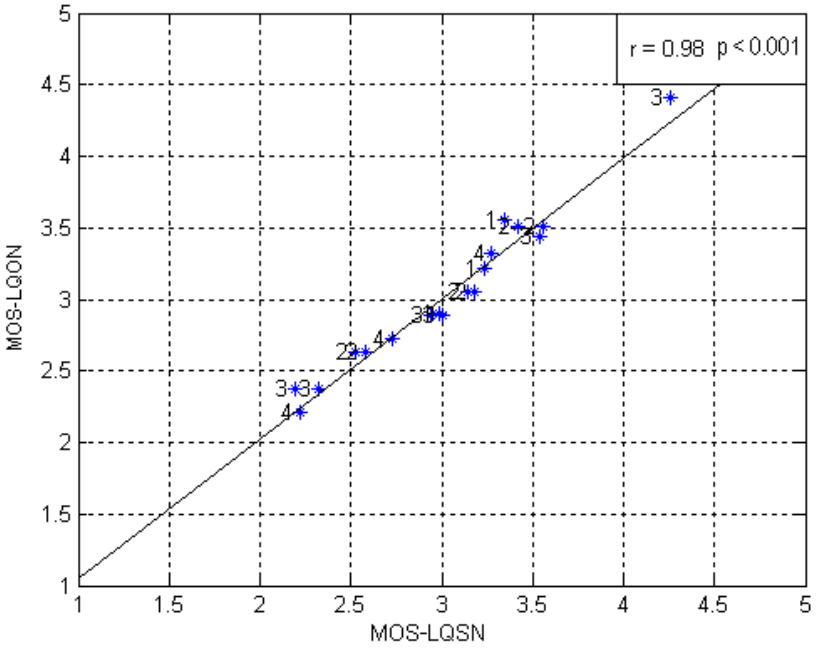


Fig. IV.14 Performance des 4 régressions logarithmiques (DAV manuelle + classification manuelle + niveau d'isosonie issus du test préliminaire), suivant les résultats moyennés selon les phrases du test d'évaluation de la qualité vocale ; les numéros correspondent à la classe de BDF (1→ Intelligible / 2→ environnement / 3→ souffle / 4→ grésillement)

Dans le cas où on ne tient pas compte des différentes classes de bruit de fond (en considérant que tous les bruits appartiennent à la classe souffle), la corrélation est alors de $r = 0,93$, $p < 0,001$ au lieu de $r = 0,98$, $p < 0,001$. Cela montre une augmentation des performances de l'évaluation de la qualité vocale lorsque le type de bruit de fond est pris en compte grâce aux quatre classes proposées.

Lorsque le modèle de bruyance est appliqué avec les modules de régression et de classification automatique du bruit de fond, nous calculons une corrélation de $r = 0,98$, $p < 0,001$. Ce résultat montre que le module de classification est performant sur cette base sonore d'apprentissage.

En appliquant le modèle avec les modules de régression, de classification automatique et du calcul du niveau sonore du bruit de fond par le RSB ou bien par la sonie de Zwicker, nous calculons une performance globale de $r = 0,93$, $p < 0,001$. L'estimation du niveau so-

nore par les deux méthodes (RSB et sonie de Zwicker) cause une légère diminution de la performance du modèle, qui reste cependant acceptable. Cette analyse ne permet pas de fixer l'indicateur de niveau sonore le plus pertinent. Néanmoins, le RSB est moins coûteux à calculer que la sonie.

Dans les deux cas de figure précédents, la détection active de la parole (DAV) a été réalisée manuellement. Le modèle de la bruyance a aussi été testé dans sa globalité avec les trois modules (DAV, classification, indicateur de niveau sonore). La corrélation est alors de $r = 0,92$, $p < 0,001$, quel que soit l'indicateur de niveau sonore du bruit de fond.

Nos résultats ont été comparés à deux modèles basés sur le signal. Le modèle PESQ est intrusif et obtient une corrélation de Pearson de $r = 0,91$, $p < 0,001$. Ce résultat est similaire à la performance du modèle non-intrusif proposé dans ce chapitre mais PESQ nécessite d'avoir accès au signal de référence et ne peut être appliqué en temps réel. Le modèle P.563 est non intrusif et obtient une corrélation de $r = 0,65$, $p < 0,001$. Les modèles existants ne prennent pas en compte l'influence de l'interaction entre le niveau sonore et le type de bruit de fond. Ils sont donc moins représentatifs du jugement des utilisateurs lorsque les stimuli testés présentent différents types de bruits de fond.

IV.4.2. Validation du modèle de bruyance sur une base sonore inconnue

La performance du modèle de bruyance est évaluée sur une nouvelle base sonore issue d'une étude réalisée par Gautier-Turbin et Gros [92] qui traite de l'impact de l'emploi de différentes échelles d'évaluation de la gêne due aux bruits de fond (P.835 et une échelle de dégradations). La base sonore est constituée de stimuli bruités, soumis à deux algorithmes de débruitage.

Les conditions de dégradation sont composées de trois types de bruits de fond qui sont restitués à deux niveaux sonores différents :

- Bruit de rue (comprenant des bruits d'oiseaux / de pas / de souffle)
- Bruit de bureau (bruit rose)
- Bruit de foule (bruit de cocktail party)

Quatre doubles-phrases prononcées par deux locuteurs sont dégradées par chacune de ces six conditions. Au total, 24 stimuli sont disponibles pour chacun des deux algorithmes de débruitage.

Le débruitage est un outil utilisé en télécommunication pour limiter la présence de bruit de fond sur le signal vocal afin d'augmenter l'intelligibilité ou/et la qualité de la parole. Deux techniques de débruitage sont testées à deux degrés de débruitage :

- Un débruitage "léger" va déformer légèrement le signal global (voix et bruit de fond), mais en conservant un niveau sonore de bruit de fond élevé.
- Un débruitage "agressif" va déformer le signal global (voix et bruit de fond), mais en réduisant fortement le niveau sonore du bruit de fond.

Le modèle de bruyance ne prend pas en compte les dégradations de la parole. Nous utilisons donc uniquement les conditions de dégradation traitées par les débruiteurs "légers" car le signal de parole n'est que faiblement dégradé et que le bruit de fond est encore présent, contrairement aux stimuli traités par les débruiteurs "agressifs". Les deux algorithmes de débruitage sont traités séparément pour limiter l'effet des différences de dégradation de la parole entre les deux techniques de débruitage utilisées.

Lors de la même étude, les notes MOS-LQSN associées aux conditions de dégradation ont été obtenues par un test subjectif selon la méthode normalisée de l'ITU-T P.835 [93]. Cette technique consiste à demander aux sujets d'évaluer la qualité du signal sonore selon trois critères :

- Evaluer uniquement la qualité du signal de parole
- Evaluer uniquement la qualité du signal du bruit de fond
- Evaluer la qualité globale du signal (parole et bruit de fond)

Afin de pouvoir comparer les résultats de ce test subjectif à notre modèle, nous considérons que les notes globales MOS-LQSN issues du test P.835 sont équivalentes à celles obtenues par le test ACR (UIT-T P.800 [11]). Les performances du modèle de bruyance sont alors déterminées en comparant les notes MOS-LQON issues de notre modèle, aux notes MOS-LQSN globales issues du test subjectif (UIT P.835 [93]).

Le modèle de la bruyance est appliqué aux deux bases sonores en utilisant les quatre modules (DAV, niveau sonore (sonie et RSB), classification et régression). Les deux indicateurs de niveau sonore sont testés et obtiennent des résultats similaires, avec toutefois un temps CPU utilisé nettement moins important dans le cas de l'utilisation du RSB.

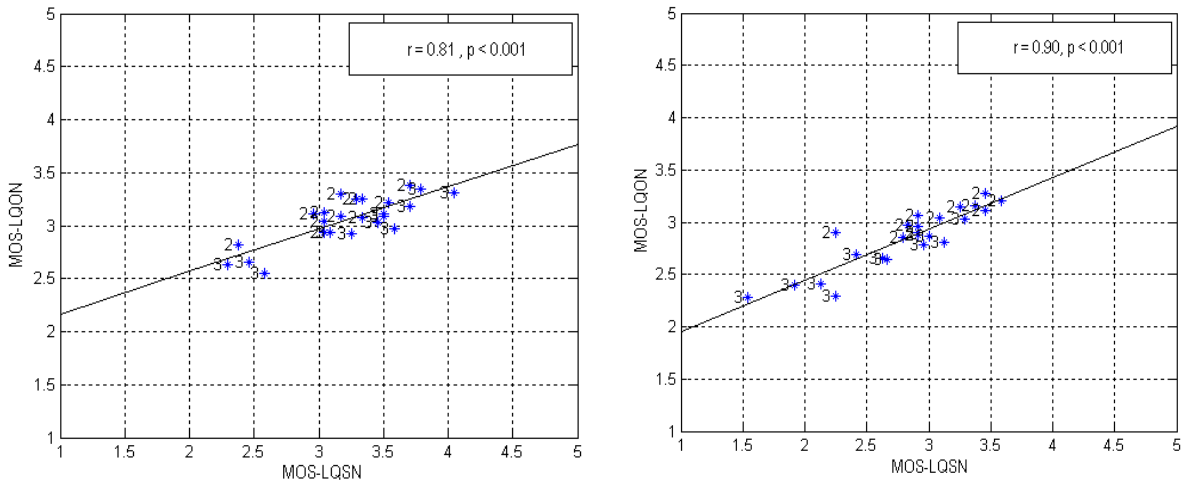


Fig. IV.15 Performance du modèle de bruyance (avec la sonie) à partir d'une base sonore inconnue au modèle ; à gauche à l'aide des résultats du 1er algorithme de débruitage et à droite à partir du 2ème algorithme de débruitage ; les numéros correspondent à la classe de BDF (1→ Intelligible / 2→ environnement / 3→ souffle / 4→ grésillement)

Le modèle de bruyance est efficace pour prédire la qualité vocale avec une corrélation de *r = 0,81* obtenue dans le cas des résultats issus du 1er algorithme de débruitage, et *r = 0,90* dans le cas du 2^{éme} algorithme de débruitage (cf. Fig. IV.15).

Le modèle permet en plus d'identifier le type de bruit de fond présent sur le signal avec une classification sans aucune erreur pour les bruits de foule (classe 2) et de bureau (classe 3). Cependant, le bruit de rue est classé 5 fois sur 8 en tant que bruit de souffle (classe 3), et 3 fois sur 8 en tant que bruit d'environnement (classe 2). Ce type de bruit fait partie des sons difficilement classifiables. Cela dépend de l'utilisateur interrogé, du niveau sonore de restitution du bruit, des dégradations du signal ou encore du contexte de communication.

IV.5. Etude de la bruyance en bande élargie

L'étude détaillée dans ce chapitre a permis de développer le modèle de bruyance à partir de stimuli transmis en bande étroite ($f < 4\ kHz$). De nouvelles technologies de transmission de la parole ont été développées en bande élargie ($f < 8\ kHz$), notamment sur les réseaux utilisant la VoIP, et aussi sur les transmissions mobiles. La télécommunication en bande élargie permet de transmettre pratiquement tout le contenu spectral utilisé par la production de la parole ($50\ Hz < f < 8\ kHz$). Cela permet de communiquer avec une voix plus naturelle et plus réaliste, et ainsi d'améliorer la qualité des services proposés.

Le bruit de fond d'environnement présent sur le signal de parole est aussi transmis avec plus de réalisme et plus de naturel en condition de bande élargie.

Nous pouvons donc faire l'hypothèse que dans le cas d'une transmission en bande élargie, il existe une influence du bruit de fond sur l'évaluation de la qualité vocale dont l'effet est encore plus important que dans le cas d'une transmission en bande étroite.

Les expériences réalisées précédemment ont donc été refaites afin d'étudier l'influence du niveau sonore et surtout l'influence du type de bruit de fond, pour des conditions de dégradation présentées en bande élargie.

IV.5.1. Tests subjectifs

Les mêmes tests subjectifs que ceux présentés dans la partie IV.1 (test préliminaire d'égalisation de la sonie par la méthode d'ajustement et test d'évaluation de la qualité vocale par la méthode ACR) ont été réalisés une fois en condition de bande étroite par 24 sujets, et une fois en condition de bande élargie par 24 autres sujets.

Il y a trois différences entre le test réalisé précédemment et les tests présentés dans cette partie :

- La largeur de bande passante des stimuli est de 4 kHz dans le cas de la bande étroite, et de 8 kHz dans le cas de la bande élargie. Le filtrage SRI P.48 [84], et le filtrage P.341 [89] ont été appliqués respectivement pour les conditions en bande étroite et élargie, afin de simuler le terminal émetteur.
- La base sonore est constituée de conditions bruitées par huit bruits de fond différents. Un bruit de fond est de classe intelligible (musique), quatre bruits de fond appartiennent à la classe d'environnement (cantine, sport, ville, piscine), deux bruits sont de souffle (rose, mer) et un bruit est de la classe grésillement (bruit électrique). Ces huit bruits de fond sont chacun mixés à huit double-phrases prononcées par quatre locuteurs pour trois niveaux d'isosonie. Les huit phrases sont aussi diffusées aux sujets sans bruit de fond. Au total, 200 stimuli en bande étroite et 200 stimuli en bande élargie sont diffusés aux sujets lors des deux tests d'évaluation de la qualité vocale.
- Le moyen de restitution est diotique[14].

IV.5.2. Résultats des tests subjectifs

Les résultats de l'évaluation de la qualité vocale des deux tests réalisés (bande étroite et élargie) sont moyennés suivant les 24 sujets et les 8 phrases en faisant la distinction entre les

[14] Une écoute diotique correspond à la diffusion d'un même signal audio aux deux oreilles de l'auditeur.

quatre classes déterminées au §.IV.2.3 (bruits intelligibles, d'environnement, de souffle et de grésillement) (cf. Fig. IV.16).

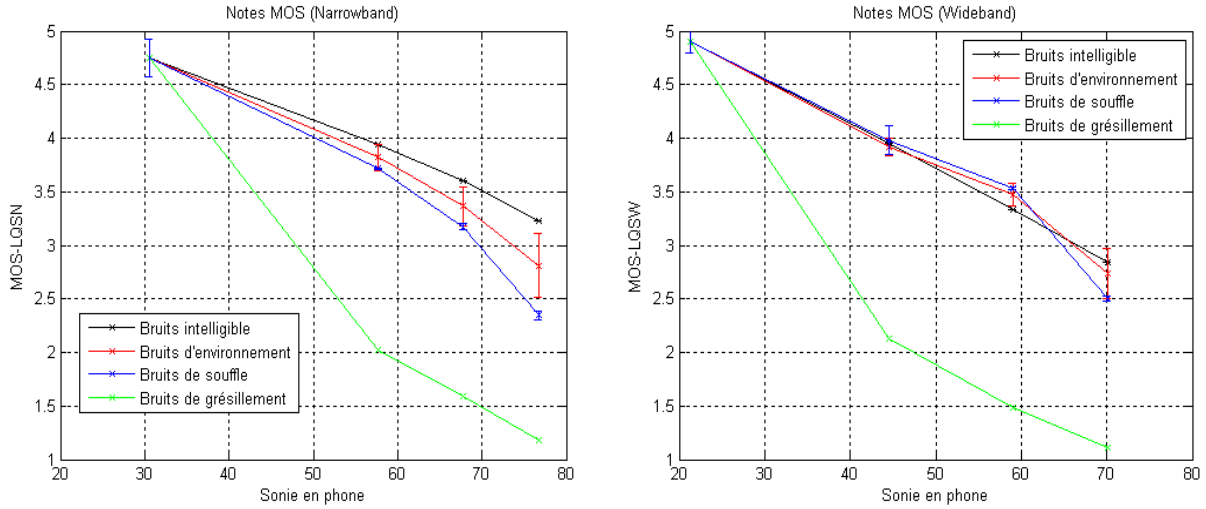


Fig. IV.16 Notes MOS-LQSN moyennées sur les phrases suivant le niveau sonore des BDF en faisant la distinction entre les 4 classes de bruits, avec l'intervalle de confiance à 95%

La classification des bruits de fond est bien appropriée aux résultats du test en bande étroite (cf. Fig. IV.16 de gauche).

Dans le cas de la bande élargie, ce résultat n'est pas aussi net. Les classes de bruits intelligibles, d'environnement et de souffle commencent à se distinguer à partir du niveau d'isosonie de 70 phone. Les niveaux sonores des bruits diffusés en bande élargie sont trop faibles par rapport aux niveaux des bruits diffusés en bande étroite. Nous faisons l'hypothèse que les bruits en bande élargie n'ont pas été diffusés à des niveaux assez forts, afin de quantifier entièrement l'influence du type de bruit de fond.

A partir d'un niveau de bruit de fond de 70 phone, les conditions bruitées appartenant à la classe intelligible sont moins gênantes que celles appartenant à la classe d'environnement, qui sont elles mêmes moins gênantes que les conditions de la classe de souffle. Le bruit de fond électrique appartenant à la classe grésillement provoque une gêne importante sur la qualité vocale. Le bruit électrique est très mal perçu à cause de ses caractéristiques rugueuses. Il est encore plus dérangeant que le bruit électrique généré lors du premier test (cf. §.IV.2.3).

Les résultats des deux tests d'évaluation de la qualité vocale permettent de conclure que le type de bruit de fond joue un rôle important lors de l'évaluation de la qualité vocale. Par ailleurs, l'influence du type de bruit de fond n'est pas plus élevée dans le cas d'une transmission en bande élargie que lors d'une transmission en bande étroite. Elle semble similaire, bien que de nouvelles analyses doivent être réalisées pour des niveaux sonores de bruit de fond plus élevés. Les quatre classes de bruits déterminées précédemment sont adaptées aux transmissions en bande étroite et en bande élargie. Le modèle de bruyance est donc appliqué aux deux bases sonores afin de vérifier ses performances dans le cas des transmissions en bande étroite et en bande élargie.

IV.5.3. Application du modèle de bruyance aux deux bases sonores (bandes étroite et élargie)

Le modèle de bruyance développé pour un contexte de transmission en bande étroite (cf. §.IV.3) a été appliqué aux deux bases sonores contenant 200 stimuli chacun (cf. §.IV.5.1).

Quelques adaptations ont été réalisées aux modèles dans le cas de la base sonore en bande élargie :

- Afin de calculer les indicateurs pour les signaux présentés en bande élargie, la durée des trames analysées doit être équivalente à celles analysées en bande étroite (64 ms). Les fréquences d'échantillonnage sont de 16 kHz en bande élargie contre 8 kHz en bande étroite. Le nombre d'échantillons analysés par trame est donc doublé (1024 échantillons par trame en bande élargie au lieu de 512 en bande étroite). Pour les mêmes raisons, la longueur de recouvrement des trames successives est doublée (512 échantillons au lieu de 256).
- Certains indicateurs utilisés par le modèle deviennent incohérents lorsque la puissance acoustique est trop élevée dans les fréquences inférieures à 100 Hz. Un filtrage passe-haut de fréquence *f = 100 Hz* a donc été appliqué à tous les stimuli de la base sonore de la bande élargie.

Pour ces deux bases sonores, nous avons comparé les performances de la prédiction de la qualité vocale lorsque la classification automatique des bruits de fond est appliquée ou non (cf. Fig. IV.17 dans le cas de la base sonore en bande étroite).

La classification automatique du bruit de fond est performante pour les conditions présentées dans les deux bandes passantes. Nous remarquons cependant quelques erreurs lorsque la détection d'activité vocale n'est pas bien déterminée car il subsiste parfois un résidu de signal de parole dans le bruit.

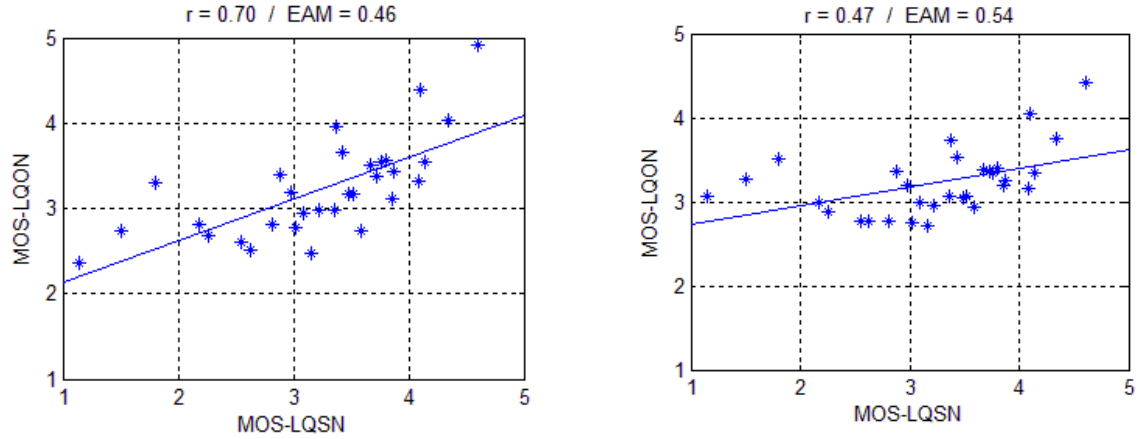


Fig. IV.17 Application du modèle de bruyance à la base sonore du test en bande étroite, avec la classification automatique à gauche et sans classification à droite (BDF de souffle uniquement). Les notes MOS sont moyennées suivant les 8 phrases.

On relève des corrélations de *r = 0,70* avec la classification et *r = 0,47* lorsque tous les bruits sont placés dans la classe des bruits de souffle, ce qui montre que la classification du type de bruit de fond améliore la prédiction de la qualité vocale. Les trois conditions présentant des MOS-LQSN comprises entre 1 et 2 correspondent aux stimuli avec du bruit électrique, diffusé pour les trois niveaux sonores. Ces trois conditions sont surestimées par le modèle DESQHI. Quel que soit le niveau sonore de ce bruit, l'évaluation de la qualité vocale correspondante est mauvaise.

Dans le cas de la base sonore en bande élargie, nous observons le même effet, avec une corrélation de *r = 0,61* lors de l'utilisation de la classification automatique des bruits de fond et *r = 0,43* lorsque les bruits sont placés dans la classe souffle. La différence de corrélation est moins flagrante que dans le cas des stimuli en bande étroite, mais cela montre que le phénomène est similaire pour les transmissions en bande étroite et en bande élargie. Cependant, le modèle de bruyance pourrait être amélioré pour les conditions transmises en bande élargie.

Le modèle de bruyance a été construit à partir de résultats subjectifs obtenus lors d'une restitution monaurale du signal de parole. Les tests subjectifs décrits dans cette partie ont été réalisés en diotique. Il a été remarqué qu'il existe une influence du système de restitution lors de l'évaluation de la qualité vocale (Nagle *et al.* [94]). D'autres expériences devraient être réalisées pour analyser l'influence du système de restitution, lors de l'évaluation de la qualité vocale de conditions bruitées.

Le modèle P.563 [2] a été testé sur la base sonore en bande étroite, mais il n'est pas adapté aux conditions bruitées de notre base sonore ($r = 0{,}02$). Le modèle paramétrique G.107 [95] ne dispose pas des informations relatives aux bruits de fond à partir des statistiques du réseau. Il ne permet pas de prédire la qualité vocale de ces conditions bruitées.

IV.6. Conclusion

L'étude de la bruyance proposée dans ce chapitre a permis de relever et de vérifier l'influence du type de bruit de fond, en plus de l'effet prédominant du niveau sonore des bruits, lors de l'évaluation de la qualité vocale. Cette étude a été menée pour des transmissions téléphoniques en bande étroite. Des éléments de réponse sont aussi présentés pour des transmissions en bande élargie.

L'influence du type de bruit de fond lors de l'évaluation de la qualité vocale peut être modélisée en classifiant les bruits dans quatre classes (intelligible, environnement, souffle et grésillement).

Les résultats des tests subjectifs ont permis de construire un modèle de bruyance non-intrusif et basé sur le signal, qui permet d'évaluer la qualité vocale de conditions bruitées, en prenant en compte leurs types et leurs niveaux sonores. Ce modèle comprend quatre modules : le module de détection de l'activité vocale, le module d'estimation du niveau sonore, le module de classification des bruits et le module de prédiction de la qualité vocale.

Les performances de ce modèle montrent de bons résultats comparativement aux modèles existants qui ne prennent pas en compte l'influence du type de bruit de fond.

Le modèle semble aussi être applicable à des conditions bruitées diffusées en bande élargie, en appliquant préalablement un filtrage passe-haut sur les signaux de bruit de fond, et en considérant des durées de trames équivalentes à 64 ms.

Par ailleurs, le modèle de bruyance permet de diagnostiquer la transmission en donnant une information très utile sur l'origine du bruit de fond. Ce diagnostic avancé permet dans notre cas d'améliorer la prédiction de la qualité vocale, mais il peut aussi être utile dans d'autres domaines comme la décision d'utiliser un algorithme de débruitage ou encore la reconnaissance de l'environnement du locuteur.

Le modèle de bruyance est appliqué à la base sonore utilisée lors de la détermination de l'espace perceptif afin de représenter la première dimension bruyance (cf. §.III.3.3.1). Pour cela, une régression linéaire est déterminée entre les notes MOS-LQON prédites par le modèle de bruyance et les coordonnées des 46 stimuli projetés suivant la première dimension :

$$DIM_1 = -8{,}433 + 2{,}005 \cdot MOS - LQON \qquad \textbf{Eq. IV.5}$$

Cette base sonore est uniquement composée de bruit rose. La modélisation utilise donc uniquement la régression déterminée pour la classe de bruit de souffle, pour les deux indica-

teurs de niveau sonore (sonie de Zwicker dans le cas de la Fig. IV.18 de gauche et par le rapport signal sur bruit à droite).

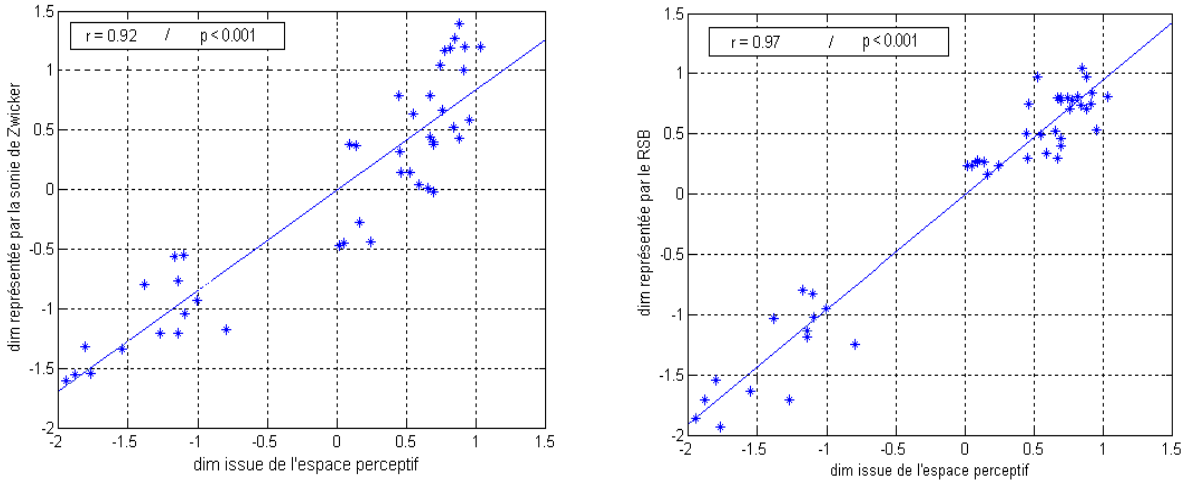


Fig. IV.18 Représentation de la première dimension de l'espace perceptif (cf. §.III.3.3.1) par le modèle de bruyance utilisant l'indicateur de sonie à gauche et le RSB à droite

Dans le cas du modèle utilisant la sonie de Zwicker, nous remarquons une légère surestimation de la bruyance pour les stimuli prononcés par la voix de femme par rapport à ceux prononcés par la voix d'homme.

Les coefficients de corrélation obtenus entre la dimension issue de l'espace perceptif et les dimensions calculées par la sonie et le rapport signal sur bruit doivent être considérés avec prudence car la répartition des conditions suivant cette première dimension n'est pas uniforme (cf. §.III.3.3.1). Néanmoins, nous remarquons que le modèle de bruyance utilisant l'indicateur RSB est légèrement plus performant que celui utilisant l'indicateur de sonie (respectivement $r = 0{,}97$, $p < 0.001$ et $r = 0{,}92$, $p < 0.001$).

De manière générale, nous avons remarqué que les performances de la représentation de la bruyance avec l'utilisation du rapport signal sur bruit sont similaires ou meilleures que l'indicateur de sonie utilisant le modèle de Zwicker. Le temps CPU utilisé pour calculer le rapport signal sur bruit est aussi nettement inférieur à celui de la sonie de Zwicker. Ces différentes raisons nous permettent de retenir l'indicateur du rapport signal sur bruit pour représenter le niveau sonore du bruit de fond lors de la modélisation de la bruyance.

Chapitre V. Modélisation du codage de la parole et de la continuité du signal

Le modèle *Diagnostic and Evaluation of Speech Quality using Hybrid Indicators* (DESQHI) est basé sur la représentation des trois dimensions perceptives correspondant à la bruyance, le codage de la parole, et à la continuité (cf. §.III.3.3). La bruyance a fait l'objet du chapitre précédent : un modèle de bruyance utilisant des indicateurs basés sur le signal a été proposé. Il mesure uniquement les dégradations causées sur les zones inactives du signal. Ce chapitre présente la modélisation des deux dimensions codage de la parole et continuité en proposant différents types d'indicateurs (paramétriques, basés sur le signal ou hybrides), afin de proposer un modèle adaptatif selon les informations disponibles au point de la mesure ou encore selon le domaine d'application (contrôle différé ou temps réel, planification).

Tous les indicateurs proposés dans ce chapitre ont été déterminés à partir des 46 conditions de dégradations et de leurs coordonnées respectives suivant chacune des deux dimensions perceptives.

V.1. Modélisation du codage de la parole

Le codage de la parole est nécessaire afin de pouvoir transmettre l'information vocale via des réseaux numériques avec un minimum de débit, mais il génère des dégradations de la qualité vocale.

L'identification de la deuxième dimension a permis d'associer différents attributs issus de la littérature et relatifs au codage de la parole (*naturel de la voix*, *coloration*, *distorsion*, *sifflement*, *grésillement*). Cette dimension est aussi étroitement liée à la *dégradation de la qualité vocale* (cf. §.III.3.3.2).

V.1.1. Indicateur paramétrique

Les dégradations de la qualité vocale provoquée par le codage employé peuvent être représentées par les facteurs de dégradation *Ie* (UIT G.113 [55] cf. Tab. V.1). Ils ont été déterminés à partir de tests subjectifs d'évaluation de la qualité vocale, pour des conditions soumises à divers codages de la parole. Chaque facteur a été défini comme la différence d'évaluation de la qualité vocale obtenue entre la condition soumise au codec testé et la condition non-codée correspondante. Le facteur de dégradation est exprimé sur l'échelle de la note R (cf. §.I.5.2.1). Le modèle E [3] utilise ce facteur *Ie* afin de prédire la qualité vocale de stimuli dégradés par le codage. Le transcodage[15] n'est pas pris en compte par la recommandation de l'UIT G.113 [55]. Le modèle E [3], et Möller [5] suggèrent d'ajouter les coefficients *Ie* entre eux lors de l'utilisation de transcodage. Cette technique semble efficace pour la plupart des codecs. Il existe des exceptions de combinaison de codages, où l'ordre d'apparition du codage a une influence sur les valeurs de *Ie*, notamment dans le cas du codec G.728 (Möller [5]). En pratique, les statistiques du réseau permettent d'obtenir l'information sur le codage utilisé lors d'une télécommunication (à minima concernant le réseau IP dans lequel est situé le point de mesure).

[15] Le transcodage correspond à l'utilisation de plusieurs codages de la parole successifs. C'est le cas lorsque plusieurs réseaux sont mis à contribution lors d'une télécommunication (GSM et VoIP par exemple) ou lors de l'utilisation d'un pont de conférence.

Codage & transcodage	Facteur de dégradation Ie en note ΔR G.113 [55]
G.711	0
G.711-G.726	7
G.726	7
G.711-GSMEFR	3
GSMEFR	3
G.726-GSMEFR	7 + 3 = 10
GSMEFR-GSMEFR	3 + 3 = 6
G.729	11
G.729-G.729	11 + 11 = 22
G.711-G.729	11
G.711-GSMFR	20
GSMFR	20
GSMEFR-G.729	3 + 11 = 14
G.729-GSMEFR	11 + 3 = 14
GSMEFR-G.729-GSMEFR	3 + 11 +3 = 7
G.729-G.729-G.729	11 x 3 = 33
G.729-GSMFR	11 + 20 = 31
GSMFR-G.729	20 + 11 = 31

Tab. V.1 Facteurs de dégradation *Ie* (UIT G.113 [55]) correspondant au type de codage et transcodage

Les indicateurs *Ie* sont soumis à une transformation linéaire afin de correspondre à la deuxième dimension de l'espace perceptif (cf. Fig. V.1) :

$$DIM_2 = 1{,}046 - 0{,}110 \cdot Ie \qquad \textbf{Eq. V.1}$$

Dans le cas du transcodage, les valeurs des indicateurs *Ie* sont ajoutées entre elles. La corrélation obtenue entre la dimension calculée par l'indicateur *Ie* et la dimension perceptive est de $r_{Ie} = 0{,}88$, $p < 0{,}001$ (cf. Fig. V.1). La dégradation de la qualité vocale provoquée par le codage de la parole peut être assimilée à cette deuxième dimension, malgré certaines dispersions des codecs selon la dimension perceptive. Nous remarquons aussi une sous-estimation du codage G.726 par l'indicateur *Ie*, dans le cas de notre base sonore. Les dégradations dues au codage de la parole influencent directement l'évaluation de la qualité vocale ou encore le naturel de la voix (cf. §.I.4.6).

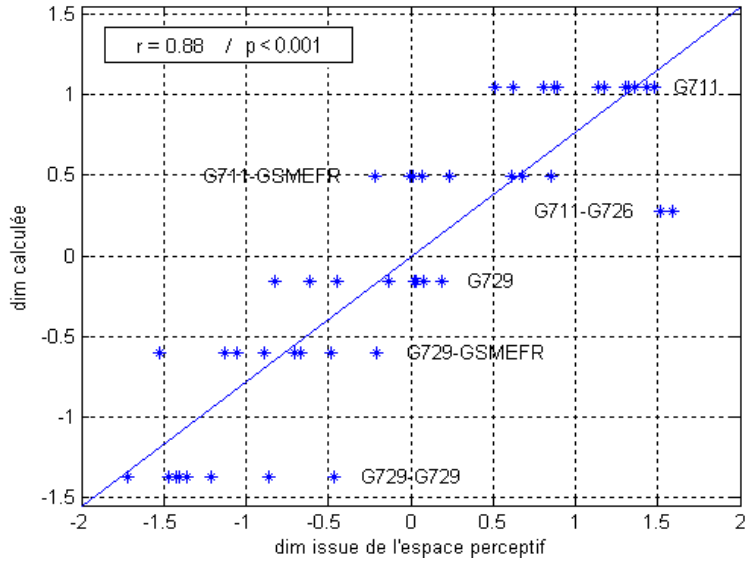


Fig. V.1 Performance de la modélisation de la dimension codage de la parole, à partir de l'indicateur paramétrique *Ie* (G.113 [55]), en prenant en compte le transcodage, pour les 46 stimuli (voix de femme et d'homme)

La dimension codage de la parole du modèle DESQHI peut être représentée par le facteur de dégradation *Ie* issu de la recommandation UIT G.113 [55]. Cet indicateur est pratique pour représenter notre deuxième dimension car il a été défini pour un large panel de codec non utilisé dans cette partie (p. ex. les codecs G.723, G.728, les conditions MNRU…).

L'inconvénient majeur de cet indicateur est qu'il est difficile de connaître tous les codecs successifs utilisés, dans un contexte de contrôle de la qualité vocale en temps réel. En pratique, seul le dernier codec est considéré par les modèles d'évaluation de la qualité vocale. Ce problème est inexistant lors de l'utilisation d'indicateurs basés sur le signal.

V.1.2. Indicateur basé sur le signal

L'attribut coloration semble être un attribut pertinent pour la représentation de la dimension du codage de la parole (cf. §.III.3.3.2). Les indicateurs de centre de gravité spectral (centroïde), son équivalent dans le domaine perceptif (sharpness) et le point de coupure à 95 %, tous ont donc été calculés pour les 46 stimuli prononcés par les deux voix.

D'autres indicateurs comme par exemple le flux spectral (cf. §.IV.3.1.2.B) ont aussi été testés pour représenter la dimension du codage de la parole. Cet indicateur semble pertinent pour représenter l'attribut grésillement ou bien le sifflement, comme le montre l'arbre de décision utilisé lors de la classification automatique des différents types de bruits (cf. §.IV.3.1.3).

Les indicateurs centroïde et flux spectral montrent les meilleures performances (parmi les autres). Ils sont performants pour les deux espaces déterminés à partir de la voix d'homme, et à partir de la voix de femme. Cependant, lorsqu'ils sont utilisés pour représenter la dimension de l'espace global (locuteurs femme et homme), les corrélations diminuent (cf. Tab. V.2).

	Femme	Homme	Global
Flux spectral	*r = -0,84 / p < 0,001*	*r = -0,93 / p < 0,001*	*r = -0,68 / p < 0,001*
Centroïde	*r = 0,63 / p < 0,001*	*r = 0,84 / p < 0,001*	*r = 0,35 / p = 0,12*

Tab. V.2 Coefficients de corrélation Pearson entre les différents indicateurs et la seconde dimension pour les voix de femme, d'homme et de l'espace global (femme et homme)

Dans le cas de la réalisation d'un modèle intrusif, la différence de coloration peut facilement être déterminée en comparant le signal original et le signal dégradé. Dans le cas d'un modèle non-intrusif, ces indicateurs ne sont pas adaptés principalement à cause de la différence de timbre entre les voix. On remarque aussi ce phénomène lors d'une communication réelle : l'auditeur peut difficilement juger d'une différence de coloration du signal de parole, sans avoir accès au signal de référence. La mesure de la coloration dans l'absolu (sans comparaison avec le signal de référence) ne semble pas être un attribut assez pertinent pour représenter la dimension liée au codage de la parole.

Les dégradations provoquées par le codage de la parole correspondent à des imperfections de la transcription numérique des signaux vocaux. Ces dégradations sont présentes sur la totalité du signal (zones actives ou non-actives), néanmoins la perception de ces défauts concerne en priorité les zones actives de la parole. L'indicateur basé sur le signal proposé dans cette partie est calculé uniquement sur les zones actives de la parole, en utilisant l'algorithme de Détection d'Activité Vocale (DAV) décrit dans l'annexe B, UIT G.729 [60]. Il est calculé à partir du résidu entre le signal dégradé et le signal reconstruit par les coefficients LPC. Suivant le type de langue utilisée (français, anglais, japonais, allemand…), l'attaque des syllabes est très différente. C'est pourquoi un indicateur relatif au temps d'attaque est combiné à l'indicateur résiduel des coefficients LPC, afin de s'affranchir de l'effet de la langue. Le signal de parole est analysé et reconstruit par trames de 256 échantillons équivalant à une durée de 32 ms, avec un recouvrement de 50%. L'indicateur *Ind* est alors calculé suivant les quatre étapes détaillées par la suite :

- Une première étape consiste à calculer les coefficients LPC à l'ordre dix du signal de la parole (cf. §.I.4.6). Les dix coefficients LPC obtenus permettent de reconstruire le signal de parole à un résidu près.

$$y(i) = -a(2) \times x(i-1) - a(3) \times x(i-2) - \ldots - a(p+1) \times x(i-p), \qquad \textbf{Eq. V.2}$$

  avec *y(i)* le signal reconstruit pour l'échantillon *i*, *x(i)* le signal de la parole, *a* les coefficients LPC et *p* l'ordre des coefficients LPC.

- La deuxième étape consiste à déterminer le résidu entre le signal de parole dégradé et ce même signal reconstruit pour tous les échantillons, en considérant le recouvrement de 50% pour chaque trame de 256 échantillons.

$$res(i) = x(i) - y(i), \qquad \textbf{Eq. V.3}$$

- ensuite, l'attaque *att* du signal reconstruit est déterminée à partir de la moyenne de la valeur absolue de la dérivée première du signal reconstruit

$$att = \frac{1}{N} \sum_{i=1}^{N-1} \left| y_{i+1} - y_i \right|, \qquad \textbf{Eq. V.4}$$

  avec *N* le nombre d'échantillons du signal reconstruit y_i. La dérivée première permet de compenser la différence d'erreurs de codage LPC en fonction des différentes langues utilisées. En effet, le codage prédictif se prête par exemple mieux au français qu'au japonais, l'indicateur appelé « attaque du signal » compense en partie cet effet.

- L'indicateur *ind* est constitué du rapport entre le résidu et l'attaque du signal reconstruit.

$$Ind = \frac{\frac{1}{N}\sum_{i=1:N}|res(i)|}{att}, \qquad \textbf{Eq. V.5}$$

avec N le nombre d'échantillons.

La relation entre cet indicateur et la dimension codage de la parole est déterminée à partir de la base sonore de l'expérience 1 du supplément 23 UIT-T [96]. Cette base sonore est choisie car elle est constituée uniquement de conditions relatives au codage de la parole (G.729, G.726, G.728, G.711, GSM-FR, IS-54 (North American VSELP), JD-HR (half-rate Japanese digital cellular), MNRU) et car les notes de qualité vocale correspondantes sont disponibles. Cette base sonore est prononcée par des locuteurs utilisant des langues différentes (français, anglais américain, japonais). Les indicateurs *Ind* sont calculés sur les 528 stimuli de cette base sonore, puis ils sont moyennés suivant les locuteurs et les phrases pour les 44 conditions de codage et de transcodage. Les moyennes des 44 indicateurs correspondant aux 44 conditions de codage et de transcodage sont représentées selon les notes de qualité vocale correspondantes (cf. Fig. V.2).

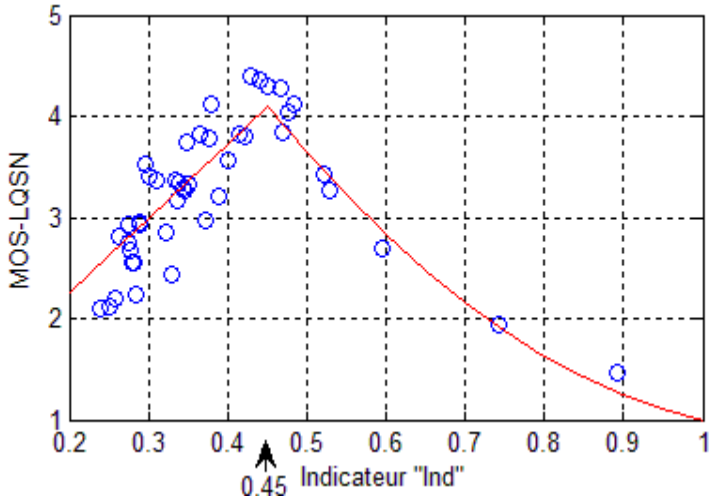


Fig. V.2 Notes de qualité vocale en fonction de l'indicateur *Ind* pour les 44 conditions de dégradation relatives au codage de la parole provenant de la base sonore du supl.23 exp.1 [96].

On remarque que les conditions de dégradation de type "MNRU" ne suivent pas la même relation que les autres codages testés : les conditions MNRU sont représentées pour des valeurs de *Ind* supérieures *à 0,45*, tandis que les différents codecs (G.729, G.726, G.728, G.711, GSM-FR, IS-54, JD-HR) sont représentés par des valeurs de *Ind* inférieures à *0,45* (cf. Fig. V.2). La prédiction de la qualité vocale par l'indicateur *Ind* est réalisée en distinguant deux groupes grâce à la valeur seuil *Vs*. Cette valeur seuil a été déterminée de manière à commettre le moins d'erreurs possible lors de la prédiction de la qualité vocale (*Vs = 0,45*). Lorsque la condition *Ind* < *Vs* est vérifiée, la relation entre l'indicateur *Ind* et les notes de qualité vocale est linéaire, sinon la relation est polynomiale d'ordre 2 (cf. Eq. V.6).

$$\begin{cases} MOS-LQON = 7{,}34 \cdot Ind + 0{,}79 & \text{si } Ind < 0{,}45 \\ MOS-LQON = 7{,}07 \cdot Ind^2 - 15{,}89 \cdot Ind + 9{,}82 & \text{si } Ind \geq 0{,}45 \end{cases} \qquad \textbf{Eq. V.6}$$

La performance de la prédiction des notes de qualité vocale par cette méthode est déterminée par le coefficient de corrélation de Pearson ($r = 0{,}89$, $p < 0{,}001$ cf. Fig. V.3).

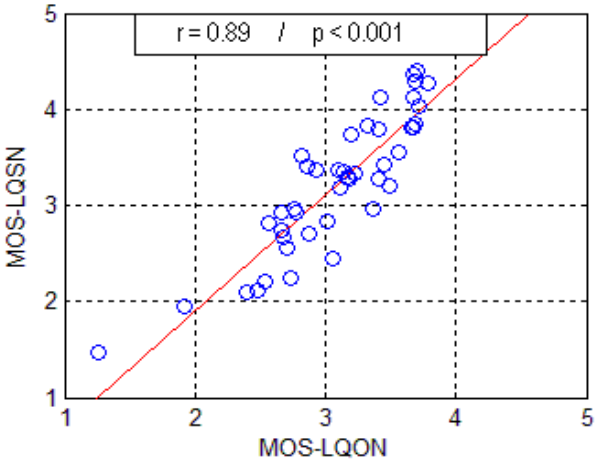


Fig. V.3 Performance de l'évaluation de la qualité vocale moyennée suivant les phrases par l'indicateur *Ind* pour la base sonore de l'expérience 1 du supplément 23

La dimension codage de la parole du modèle DESQHI est représentée par cet indicateur, en réalisant une transformation linéaire entre les notes MOS-LQON obtenues et les coordonnées des 46 stimuli suivant la dimension codage de la parole :

$$DIM_2 = -7{,}770 + 2{,}292 \cdot MOS - LQON \qquad \textbf{Eq. V.7}$$

Cet indicateur ne dépend pas du timbre du signal de la parole et représente précisément cette seconde dimension avec une corrélation entre la dimension issue de l'espace perceptif et l'indicateur proposé de $r = 0{,}86$, $p < 0{,}001$ (cf. Fig. V.4).

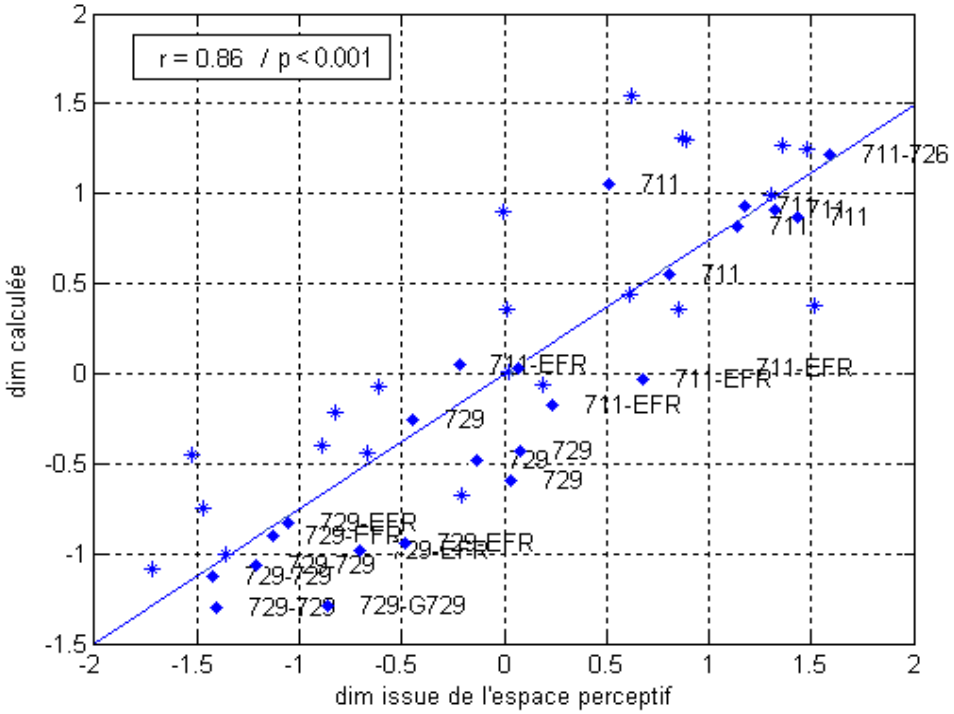


Fig. V.4 Performance de l'estimation de la dimension codage de la parole avec l'indicateur basé sur le signal, pour la voix d'homme ("*") et la voix de femme ("♦").

L'indicateur proposé dans cette partie a été développé à partir d'une base sonore différente de celle utilisée pour la construction du modèle DESQHI. Cet indicateur obtient une

bonne performance sur la base sonore utilisée pour la construction du modèle DESQHI ($r = 0,86$, $p < 0,001$ cf. Fig. V.4), et peut être utilisé pour de nombreux types de codages et transcodages.

L'indicateur *Ind* peut aussi être utilisé pour identifier le codage employé lors de la télécommunication. Cela est utilisé pour la tâche de diagnostic avancé de la qualité vocale (cf. §.VI.1.2). L'identification du codage ou du transcodage est réalisée grâce à un arbre de décision (Breiman *et al.* [91]). Six classes sont déterminées à partir de l'indicateur *Ind*, en regroupant différents types de codages et transcodages (cf. Fig. V.5).

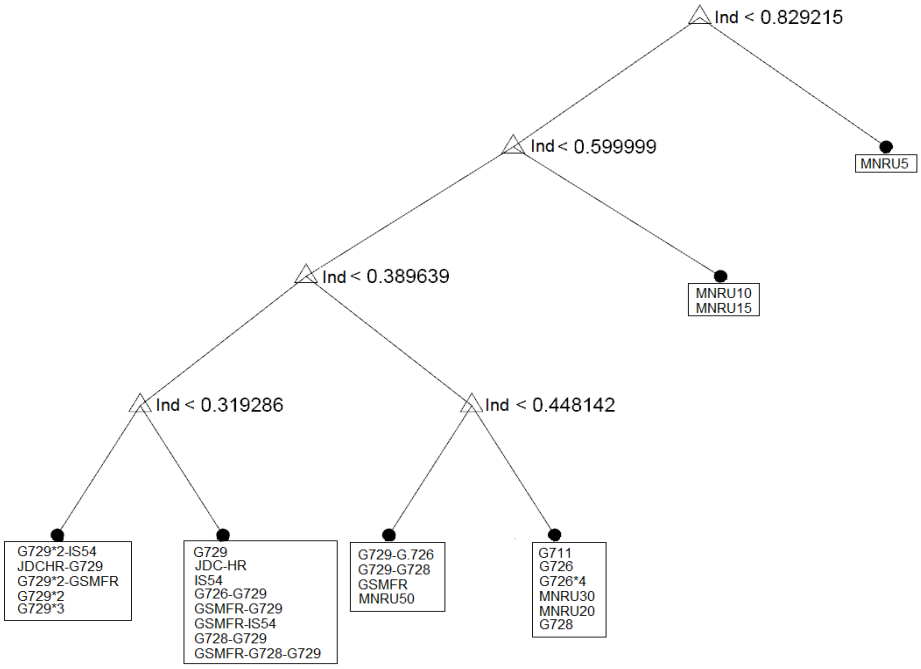


Fig. V.5 Arbre de classification des codages et transcodages en fonction de l'indicateur *Ind*

Cet arbre de décision permet principalement de donner une idée du type de codage ou de transcodage utilisé. Lorsque l'indicateur est supérieur à 0,60, il s'agit de conditions MNRU (classe 5 et 6). Si l'indicateur *Ind* est inférieur à 0,60, la valeur de l'indicateur est directement corrélée à l'évaluation de la qualité vocale des codecs correspondants.

Lorsque l'indicateur est compris entre 0,60 et 0,45, les codecs appartiennent à la première classe. Ce type de codec ne dégrade que peu le signal vocal. Quand l'indicateur est compris entre 0,45 et 0,39, les codecs appartiennent à la deuxième classe, si l'indicateur est compris entre 0,39 et 0,32, les codecs appartiennent à la troisième classe, tandis que lorsque l'indicateur est inférieur à 0,32, les codecs correspondent à la quatrième classe. Dans ce cas, les codecs dégradent fortement la qualité vocale.

Cette technique permet de distinguer les principaux types de codages (MNRU, G.711, G.729, JDC-HR, IS54, GSMFR, transcodage multiple). La distinction entre certains codecs n'est pas possible avec cette méthode (G.711 et G.726, JDC-HR et IS-54…). Il est difficile de déterminer un pourcentage de bonne classification car les catégories regroupent plusieurs codages et transcodages.

V.2. Modélisation de la continuité

La continuité de la parole correspond aux coupures temporelles du signal global (cf. §.I.4.3). Elle est provoquée par la présence de pertes de paquets ou encore d'erreurs de bits sur le signal (cf. Fig. III.20). Le niveau de continuité du signal peut être représenté par le taux d'erreurs de bits ou/et le taux de pertes de paquets, par la durée des interruptions du signal, par l'intensité sonore provoquée par les erreurs de bits, et aussi par le type de pertes (par rafale ou aléatoire). Trois types d'indicateurs sont proposés dans cette partie pour modéliser la continuité, correspondant à des indicateurs paramétrique, hybride et basé sur le signal.

V.2.1. Indicateur paramétrique

L'indicateur $I_{e,eff}$ défini dans la partie I.4.3 est utilisé par le modèle E afin de représenter les dégradations dues à la présence de pertes de paquets sur un signal pour un codec donné. Cet indicateur est calculé en fonction du facteur de dégradation *Ie*, du facteur de robustesse aux pertes de paquets *Bpl*, et du pourcentage de pertes de paquets présents sur le signal *pl*. Cet indicateur est efficace pour représenter la continuité du signal en présence de pertes de paquets, cependant il ne prend pas en compte la présence d'erreurs de bits.
Il est proposé dans cette partie un nouvel indicateur paramétrique de continuité prenant en compte toutes les dégradations en rapport avec la discontinuité du signal.

Le taux de pertes de paquets est disponible dans les statistiques du réseau IP, l'indicateur proposé ici suppose que le taux d'erreurs de bits peut être disponible à partir des statistiques du réseau mobile.

La dimension continuité est estimée par les taux de pertes de paquets et d'erreurs de bits issus des informations contenues dans les statistiques du réseau. Les pourcentages réels de pertes de paquets et d'erreurs de bits ont été déterminés dans la partie III.1.1 (2%, 4%, 6%, 8%, 9,6%, 11,4% de pertes de paquets et 0,2%, 0,41% et 0,65 % d'erreurs de bits).

Ces pourcentages sont pondérés afin d'obtenir un pourcentage global de discontinuités appelé *pgd* qui soit représentatif de la troisième dimension de l'espace perceptif (cf. §.III.3.3.3) :

$$\begin{cases} pgd = 12 \times pb & \textbf{si erreurs de bit} \\ pgd = 3{,}2 \times pl & \textbf{si pertes de paquets sans PLC} \\ pgd = pl & \textbf{sinon} \end{cases} \qquad \textbf{Eq. V.8}$$

Voici la relation entre les coordonnées des 46 stimuli projetés suivant la troisième dimension et cet indicateur *pgd* :

$$DIM_3 = -0{,}603 + 0{,}161 \cdot pgd \qquad \textbf{Eq. V.9}$$

La performance de l'estimation de cette troisième dimension par l'indicateur paramétrique *pgd* est déterminée grâce à la corrélation de Pearson entre la dimension de l'espace perceptif et la dimension déterminée par l'indicateur *pgd* : $r_{para} = 0{,}81$, $p < 0.001$ (cf. Fig. V.6).

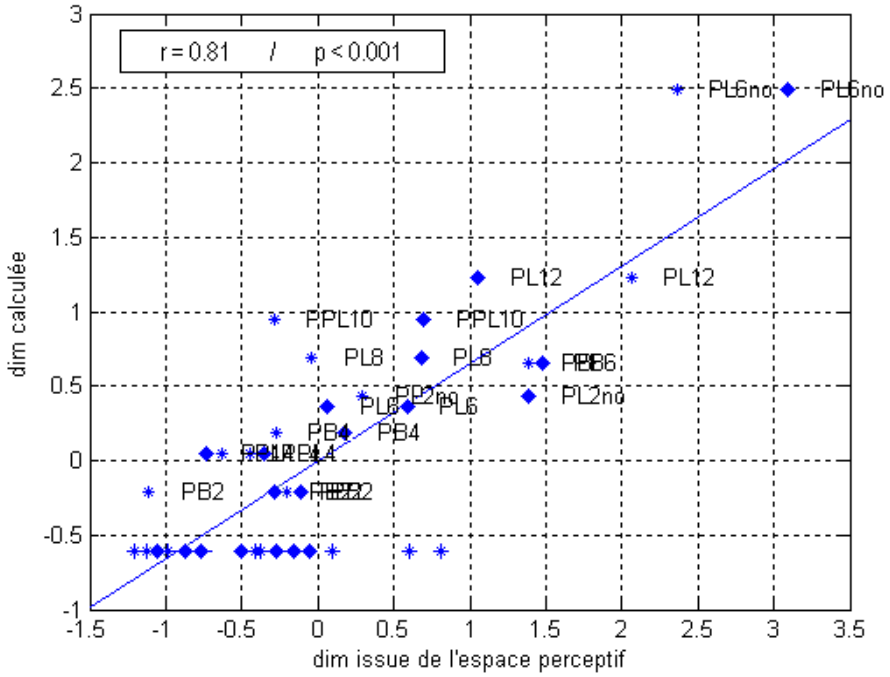


Fig. V.6 Performance de l'estimation de la dimension continuité avec l'indicateur paramétrique *pgd*, pour la voix d'homme ("*") et la voix de femme ("♦"). PL*X*→ *X*% de pertes de paquets, PL*X*no → *X*% de pertes de paquets sans PLC, PB*Y* → 0,*Y*% d'erreurs de bits.

Cet indicateur obtient une performance acceptable pour représenter la dimension continuité, cependant il n'est pas efficace pour distinguer les différences entre les deux locuteurs remarqués dans la partie III.3.3.3. Quel que soit le locuteur, l'indicateur est le même car les conditions de dégradations sont identiques.

Lorsque les deux stimuli contenant 6% de pertes de paquets sans utilisation de l'algorithme de PLC sont retirés de l'analyse, la corrélation *r* entre la dimension issue de l'espace perceptif et la dimension calculée diminue à $r = 0{,}69$, $p < 0{,}001$.

V.2.2. Indicateur hybride (paramétrique et basé sur le signal)

Deux hypothèses sont données dans la partie III.3.3.3 pour justifier les différences de localisation des discontinuités sur le signal. Ces deux hypothèses sont utilisées pour déterminer deux indicateurs hybrides.

La première hypothèse est que les discontinuités présentes sur les zones non-actives de la parole peuvent être entendues s'il y a du bruit de fond, ou ne pas être détectées dans le cas où il n'y a pas de bruit de fond. La perception de discontinuité du signal serait plus importante dans le cas d'un signal bruité que dans le cas d'un signal ne contenant pas de bruit de fond.

Afin de tester cette hypothèse, nous avons développé un indicateur hybride basé sur le pourcentage global de discontinuité *pgd* (cf. §.V.2.1) et le rapport signal sur bruit *RSB*. La représentation de la dimension continuité est alors obtenue par la combinaison suivante :

$$Dim3_{hyb} = pgd \cdot [-a + b \cdot RSB] - c \cdot RSB^2, \qquad \textbf{Eq. V.10}$$

avec *pgd* le pourcentage global de discontinuité, *RSB* le rapport signal sur bruit calculé entre les niveaux sonores moyens de la parole et du bruit de fond. *a*, *b*, *c* sont des constantes :

$a = 0{,}2021$ $\quad b = 0{,}0078$ $\quad c = 0{,}0008$

Le niveau actif de la parole est fixé à 79 dB SPL. Le rapport signal sur bruit est donc représentatif du niveau sonore du bruit de fond. Plus ce rapport est élevé, plus le niveau sonore du bruit est faible. Lorsque le terme *pgd* est nul (pas de pertes de paquets ni d'erreurs de bits), la dimension est représentée uniquement à partir du rapport signal sur bruit (terme $c.RSB^2$). Dans ce cas, plus le niveau sonore du bruit est élevé, plus la perception de discontinuité augmente. Lorsque le terme *pgd* devient non nul, le terme *b.RSB* a tendance à annuler l'effet de l'influence du niveau sonore du bruit de fond. La performance mesurée avec cet indicateur montre une corrélation de Pearson de $r = 0{,}89$, $p < 0.001$ (cf. Fig. V.7).

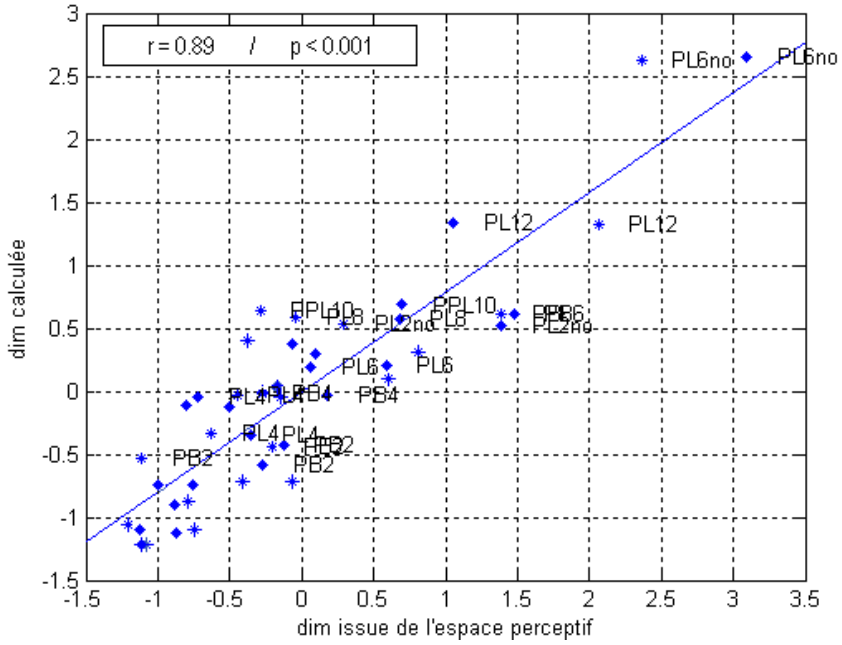


Fig. V.7 Performance de l'estimation de la dimension continuité avec l'indicateur hybride *pgd & RSB*, pour la voix d'homme ("*") et la voix de femme ("♦"). PL*X*→ *X*% de pertes de paquets, PL*X*no → *X*% de pertes de paquets sans PLC, PB*Y* → 0,*Y*% d'erreurs de bits.

Nous obtenons une meilleure performance par rapport à l'utilisation de l'indicateur paramétrique, principalement grâce aux conditions qui ne présentent pas de pertes ou d'erreurs ($-1{,}5 < dim3 < 0$). Ces conditions sont estimées par l'indicateur comme ayant un certain niveau de discontinuité. Il apparaît que les stimuli les plus continus correspondent aux conditions non bruitées, tandis que les stimuli présentant de faibles discontinuités correspondent aux conditions bruitées.

Le phénomène remarqué concernerait donc uniquement la mesure de la continuité pour les conditions de dégradations ne présentant pas de pertes ou d'erreurs. L'hypothèse concernant l'influence du bruit de fond lors de la présence de discontinuités n'a pas été vérifiée par cette étude. Cet indicateur n'est donc pas utilisé par la suite.

La deuxième hypothèse (cf. §.III.3.3.3) est testée en supposant que les discontinuités pourraient être plus gênantes lorsqu'elles sont situées sur les zones actives de la parole plutôt que sur les zones non-actives (cf. §.I.4.3). Un algorithme de détection d'activité vocale (G.729 [60]) est utilisé pour calculer le paramètre *pgd* uniquement sur les zones actives du signal. L'indicateur est hybride car il utilise le signal pour distinguer les zones actives des zones non-actives, et les paramètres du réseau pour déterminer le pourcentage global de discontinuité (cf. §.V.2.1). Les résultats de l'estimation de la troisième dimension par l'indicateur hybride sont présentés sur la Fig. V.8. Dans ce cas, la relation entre les coordonnées des 46 stimuli projetés suivant la troisième dimension et cet indicateur hybride est la suivante :

$$DIM_3 = -0{,}616 + 17{,}912 \cdot pgd \qquad \textbf{Eq. V.11}$$

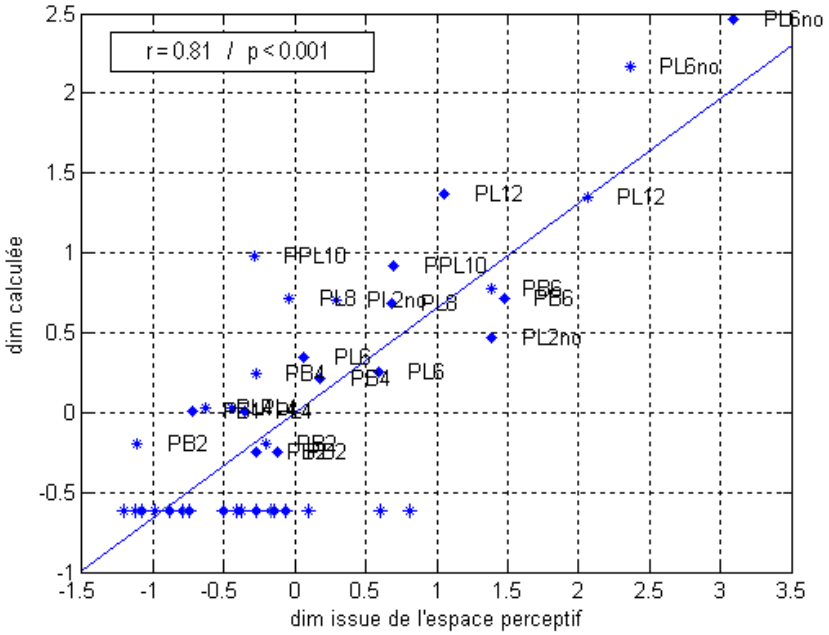


Fig. V.8 Performance de l'estimation de la dimension continuité avec l'indicateur hybride *DAV* & *pgd*, pour la voix d'homme ("*") et la voix de femme ("♦"). PL*X*→ *X*% de pertes de paquets, PL*X*no → *X*% de pertes de paquets sans PLC, PB*Y* → 0,*Y*% d'erreurs de bits.

On ne remarque pas d'amélioration significative de l'estimation de la continuité par l'indicateur hybride, comparativement à l'indicateur paramétrique (r_{para} = *0,81 et* r_{hyb} = *0,81*). Pourtant, cet indicateur hybride permet de relever des différences entre les deux locuteurs comme par exemple pour la condition *PL6no*. Cette dernière hypothèse est déjà utilisée par certains modèles existants (cf. §.I.4.3). Cet indicateur hybride est retenu pour représenter la dimension continuité.

V.2.3. Indicateur basé sur le signal

Les discontinuités génèrent des coupures plus ou moins nettes dans le signal sonore suivant leurs caractéristiques : pertes de paquets, erreurs de bit, pertes aléatoires ou par rafales, utilisation d'un algorithme de PLC,... Cela engendre des sauts irréguliers dans le signal vocal, perçus comme des "clips". Les discontinuités correspondent à la chute ou à l'augmentation soudaine de toutes les composantes fréquentielles du signal. Ce phénomène est observable par des représentations spectro-temporelles du signal, sous forme de raies dans le domaine temporel. Un exemple est donné sur la Fig. V.9 pour la condition de dégradation 11 présentant 6%

de pertes de paquets sans l'utilisation de l'algorithme de PLC, prononcée par la voix de femme (cf. Tab. III.2).

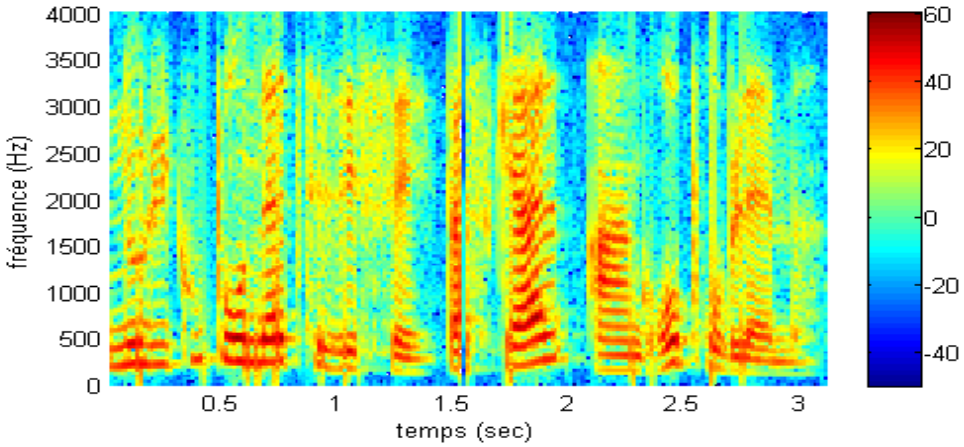


Fig. V.9 Spectrogramme de la condition de dégradation 11 prononcée par la voix de femme, présentant 6% de pertes de paquets sans l'utilisation de l'algorithme de PLC, en dB.

Les raies présentes aux endroits des discontinuités sont difficilement exploitables car elles sont masquées par le signal de la parole. Ces raies sont par contre bien visibles dans la zone spectrale où l'énergie de la parole est faible ($f < 100$ *Hz*).

La représentation de la continuité du signal vocal est basée sur l'analyse des composantes fréquentielles inférieures à 100 Hz, dans la zone où le signal de parole n'est pas représenté.

L'analyse est réalisée uniquement sur les zones actives de la voix en supposant que c'est sur la parole que les discontinuités sont les plus gênantes. On utilise l'algorithme de DAV déjà utilisé lors de la modélisation de la bruyance (cf. §.IV.3).

Le signal de parole est soumis à un filtre passe-bas d'ordre dix à la fréquence de coupure de $f_c = 80$ *Hz*. Le signal filtré est alors échantillonné à la fréquence d'échantillonnage *fe = 224 Hz* et représenté sur la Fig. V.10. Ces deux traitements sont réalisés afin de pouvoir appliquer un algorithme de détection des discontinuités présenté par la suite.

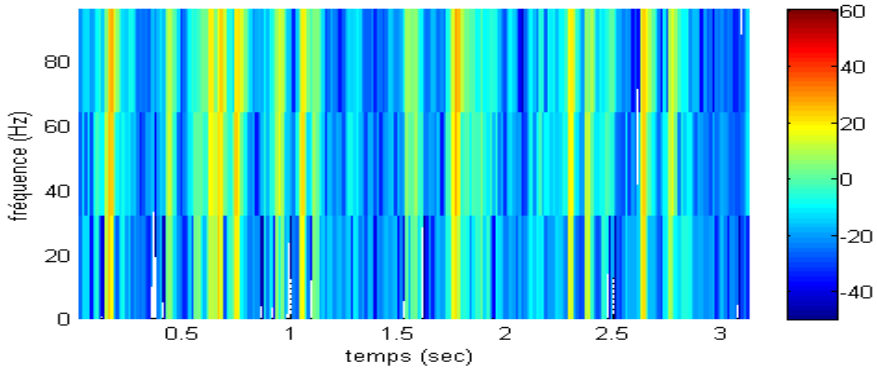


Fig. V.10 Spectrogramme du signal de parole filtré de la condition de dégradation 11 prononcée par la voix de femme, en dB pour les fréquences inférieures à 96 Hz, Δf = 32 Hz

En comparant les spectrogrammes des signaux filtrés des conditions de dégradation, nous remarquons que les stimuli codés en GSM-EFR ont des énergies plus importantes que celles des codecs G.711 et G.729 dans cette bande de fréquences inférieures à 100 Hz (cf. Fig. V.11). Les zones de discontinuités sont déterminées par rapport au niveau global du signal filtré afin de s'affranchir de l'effet du codage utilisé et de l'effet du filtrage causé par le terminal.

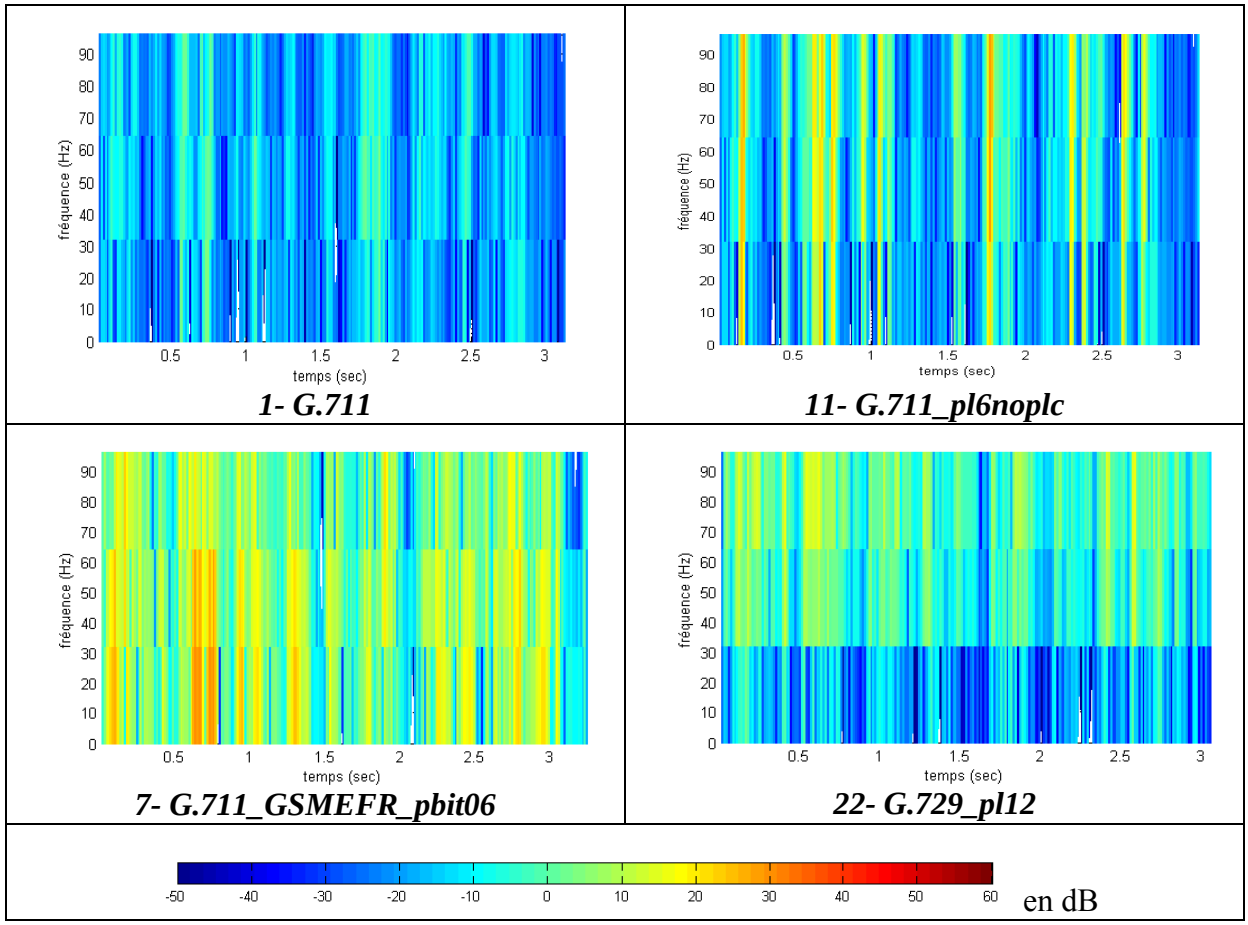


Fig. V.11 Spectrogrammes de 4 conditions de dégradations prononcées par la voix de femme en dB (conditions 1, 11, 7 et 22 détaillées dans la partie III.1.1) pour les signaux reconstitués. Δf = 32 Hz.

Les zones de discontinuité observées dans le cas du codage G.711 sans utilisation d'algorithme de PLC (condition 11) ont des énergies plus élevées que lors de l'emploi d'un codage utilisant un algorithme de PLC (condition 22). Plus la discontinuité est forte, plus la puissance des raies et la différence de puissance entre les zones continues et discontinues sont élevées. Ces variations de puissance semblent bien représenter la perception des discontinuités.

Un algorithme de détection des discontinuités est alors appliqué sur le signal filtré. Cet algorithme a été développé à partir de l'algorithme de détection de l'activité vocale détaillée dans P.56 [80], en supposant que les zones de discontinuité peuvent être assimilées aux zones de parole. Le rapport signal sur bruit est calculé comme la différence de niveau sonore entre les zones de discontinuité et les zones continues. Cet algorithme utilise le signal filtré par trame de 64 ms équivalent à 14 échantillons, avec un recouvrement de 50 %. Le niveau sonore est tout d'abord évalué sur chaque trame du signal filtré. Ensuite, cinq niveaux sonores correspondant aux seuils de discontinuité (niveaux vocaux actifs) sont déterminés en utilisant la méthode B définie dans P.56 [80] suivant les cinq marges M additionnées aux seuils de discontinuité, respectivement -12, -8, -4, -2 et 0 dB suivant que le rapport signal sur bruit est fort (12 dB) ou très faible (0 dB). La marge M à appliquer est choisie lorsque la valeur du RSB correspondante est située dans une bande prédéfinie. Cette technique a l'avantage d'analyser le signal en différence de niveau sonore entre les zones continues et discontinues du signal, et non en niveau absolu. Cela est nécessaire afin de s'affranchir des différences de niveau sonore des signaux analysés, provoquées par l'utilisation de différents types de codage de la parole ou encore suivant les terminaux utilisés (filtre IRS).

Une vingtaine d'indicateurs a été déterminée sur le signal filtré pour représenter le niveau de continuité du signal vocal. Parmi eux, quatre indicateurs ont été sélectionnés à partir de la méthode appelée "Stepwise regression" (Draper et Smith [97]). Cette méthode consiste à déterminer le nombre de termes prédictifs (indicateurs) et la régression linéaire multiple optimale des termes prédictifs, pour représenter la variable explicative (la dimension perceptive dans notre cas). A chaque étape, le terme prédictif le plus pertinent est sélectionné à partir du test F (test de Fischer) et de la valeur *p* correspondante. La valeur *p* représente la probabilité pour laquelle l'hypothèse nulle est rejetée. En d'autres termes, la valeur *p* donne une information sur le niveau de significativité du résultat statistique obtenu. Le terme prédictif retenu est celui comportant la valeur *p* la plus faible. Lorsqu'aucune des valeurs *p* n'est inférieure à une valeur fixée, la régression linéaire multiple est retenue.

Quatre indicateurs ont été retenus à partir de la méthode susmentionnée afin de représenter la dimension continuité :

Le premier indicateur I_1 représente la moyenne de la densité spectrale de puissance (*DSP*) du signal filtré, sur la bande de fréquence centrée sur 64 Hz.

$$I_1 = \frac{\sum_{i=1}^{N} DSP(i)}{N}, \qquad \textbf{Eq. V.12}$$

avec N le nombre d'échantillons.

Le deuxième indicateur I_2 représente la moyenne des valeurs maximales des zones discontinues du signal.

$$I_2 = \frac{1}{n}\sum_{j=1}^{n}[\max(|Sd_j|)], \qquad \textbf{Eq. V.13}$$

avec Sd_j le j^{ème} signal filtré identifié comme étant la j^{ème} zone de discontinuité et *n* le nombre de zones de discontinuité.

Le troisième indicateur I_3 représente l'écart-type des valeurs de la densité spectrale de puissance DSP du signal reconstruit, sur la bande de fréquence centrée sur 64 Hz.

$$I_3 = \frac{1}{N}\sum_{i=1}^{N}((DSP(i) - \langle DSP \rangle)^2)^{\frac{1}{2}}, \qquad \textbf{Eq. V.14}$$

avec *N* le nombre d'échantillons, *DSP* la Densité Spectrale de Puissance, <*DSP*> la moyenne de la *DSP* sur les *N* échantillons.

Le quatrième indicateur représente la différence de niveau sonore moyen entre les zones du signal discontinu (*Sd*) et du signal continu (*Sc*) :

$$I_4 = 10 \cdot \log_{10}(Sd_i^{\,2}) - 10 \cdot \log_{10}(Sc_i^{\,2}), \qquad \textbf{Eq. V.15}$$

Le niveau de discontinuité N_{disc} est alors obtenu par la combinaison linéaire de ces quatre indicateurs :

$$N_{disc} = 0{,}491 \cdot I_1 - 0{,}017 \cdot I_2 - 0{,}023 \cdot I_3 + 0{,}196 \cdot I_4 \qquad \textbf{Eq. V.16}$$

La troisième dimension est donnée par la relation suivante :

$$DIM_3 = -2{,}548 + 0{,}769 \cdot N_{disc} \qquad \textbf{Eq. V.17}$$

La performance mesurée de l'estimation de la troisième dimension avec cet indicateur est de $r_{sign} = 0{,}74$, $p < 0.001$ (cf. Fig. V.12).

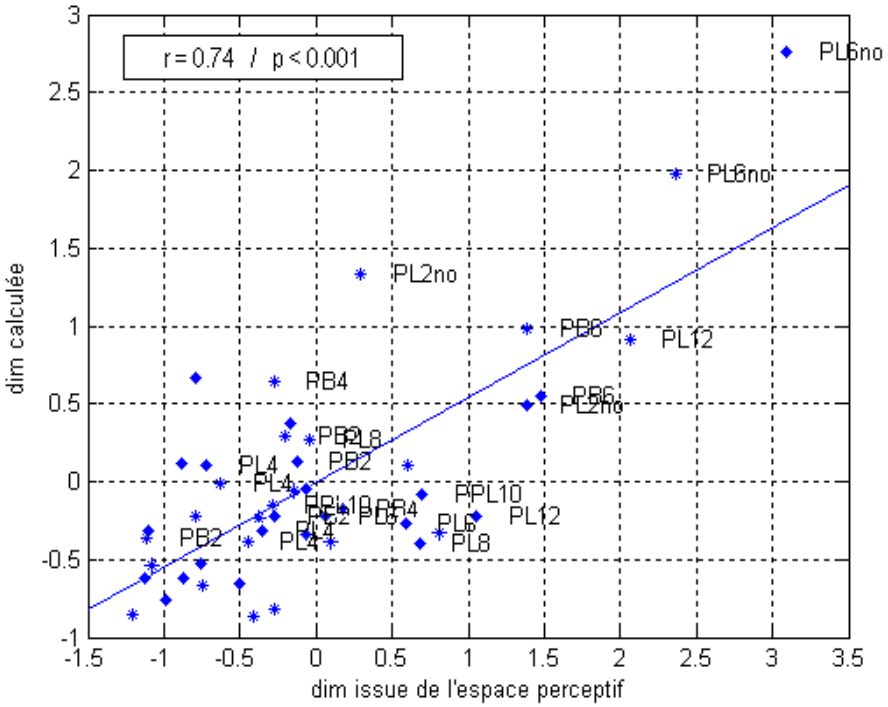


Fig. V.12 Performance de l'estimation de la dimension continuité avec l'indicateur basé sur le signal, pour la voix d'homme ("*") et la voix de femme ("♦").

La performance de cette combinaison d'indicateurs basés sur le signal est moins efficace que les indicateurs paramétriques ou hybrides (cf. §.V.2.1 et §.V.2.2). Cette méthode de prédiction du niveau de continuité du signal est tout de même très intéressante lorsque les informations issues du réseau ne sont pas accessibles au point de mesure.

Les indicateurs basés sur le signal permettent aussi d'identifier la cause de la discontinuité du signal de parole, à savoir s'il s'agit d'erreur de bits, de pertes de paquets, et si un algorithme de PLC ("Packet Loss Concealment") est utilisé ou non.

Trois classes sont proposées suivant le type de dégradations :

- La première classe comprend les stimuli qui ne présentent pas de coupure nette du signal : les signaux sans perte de paquet, et les signaux avec des pertes de paquets atténuées par un algorithme de PLC.
- La deuxième classe correspond aux stimuli présentant des coupures nettes du signal de parole, c'est-à-dire les stimuli comprenant des pertes de paquets sans utilisation de PLC.
- La troisième classe comprend les stimuli dégradés par des erreurs de bits.

Un arbre de décision est proposé à partir de ces trois classes (Breiman *et al.* [91]).

La base sonore utilisée pour réaliser ce diagnostic avancé est constituée de 222 stimuli. 46 d'entre eux correspondent aux stimuli utilisés pour la construction du cœur du modèle DESQHI (cf. §.III.1.1). Parmi les 176 stimuli restants, 64 ont été construits à partir de 4 phrases prononcées par 2 locuteurs et contiennent des pourcentages de pertes de paquets correspondant à 0% et 3% pour les codecs G.711 et G.729. Les 112 derniers correspondent à 4 autres phrases prononcées par 2 locuteurs. Chacune des 4 phrases est soumise à des dégradations de pertes de paquets avec et sans PLC, pour des pourcentages de 2%, 4%, 6%, 8% et 12%, pour les codecs G.729 et G.711. Les phrases ont aussi été présentées avec des erreurs de bits pour des pourcentages correspondant à 0,2%, 0,4%, 0,6%, 0,8%, 0,10%, avec les transcodages G.711-GSMEFR et G.729-GSMEFR.

L'étape d'apprentissage de l'obtention de l'arbre de décision (Breiman [91]) a été réalisée à partir de 170 stimuli choisis aléatoirement. Les 52 autres stimuli sont conservés afin de tester la classification sur une base sonore inconnue.

Un indicateur supplémentaire est utilisé pour ce diagnostic. Il représente la moyenne du niveau sonore du signal filtré *S* :

$$I_5 = \log_{10}(S^2) \qquad \textbf{Eq. V.18}$$

Lorsque l'indicateur $I_2 < 87{,}5889$ (indicateur correspondant à la moyenne des valeurs maximales des zones discontinues du signal filtré), le signal ne présente pas de coupure nette. Le stimulus peut ne pas contenir de perte de paquets ou bien contenir des pertes de paquets

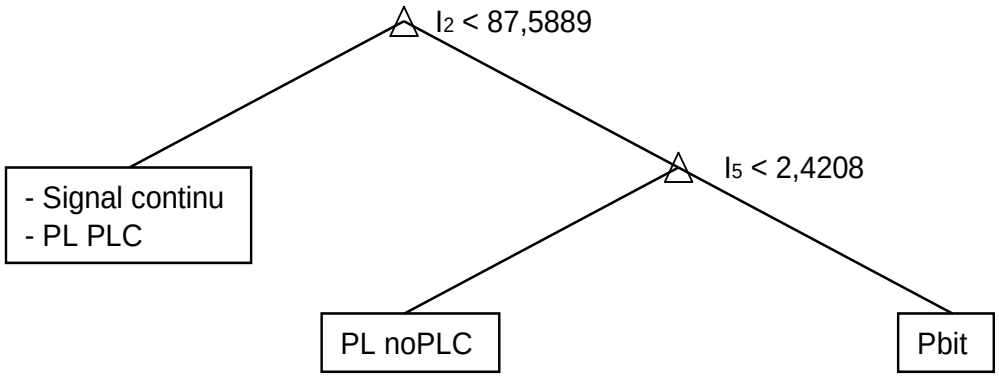


Fig. V.13 Arbre de décision pour le diagnostic avancé de la dimension continuité

Les proportions correctement classifiées correspondant aux différentes combinaisons de la base sonore (apprentissage, inconnue, totale) sont présentées dans le Tab. V.3.

Base sonore	170 BDF Apprentissage	52 BDF inconnue	222 BDF totale	46 BS cœur du modèle DESQHI
proportions correctement classifiées (en %)	81	77	80	94

Tab. V.3 Proportions correctement classifiées en % pour 4 combinaisons différentes de la base sonore

L'arbre de décision présente une proportion correctement classifiée de 80 % sur la totalité de la base sonore et de 94 % sur les 46 stimuli de la base sonore utilisée lors de la construction du cœur du modèle DESQHI.

Cette technique permet d'identifier le type de discontinuité présent sur le signal de la parole. Elle est utilisée pour fournir un diagnostic avancé de la télécommunication (cf. §.VI.1.2).

V.3. Structure globale du modèle DESQHI

Le Chapitre IV et le Chapitre V ont permis de proposer différents types d'indicateurs pour chacune des trois dimensions.

Pour la dimension bruyance, l'indicateur peut seulement utiliser des indicateurs basés sur le signal car les statistiques du réseau ne permettent pas d'obtenir les informations du niveau sonore ou encore du type de bruit de fond. La bruyance est prédite par l'indicateur basé sur le signal avec une performance de $r = 0,97$ correspondant au coefficient de corrélation de Pearson entre les résultats du test subjectif et la prédiction de la bruyance.

La dimension liée au codage de la parole peut être prédite soit par un indicateur basé sur le signal ($r = 0,86$) soit par un indicateur paramétrique ($r = 0,88$).

La dimension continuité peut être représentée par les trois types d'indicateurs : paramétrique, hybride ou encore signal, avec des performances respectivement de $r = 0,81$, $r = 0,81$ et $r = 0,74$. Cette troisième dimension comporte une part d'incertitude pour les conditions ne présentant pas de perte ou d'erreur. Il est possible que cette dimension fasse appel à d'autres attributs perceptifs.

La Fig. V.14 présente la structure globale du modèle et donne les performances de la prédiction de chacune des trois dimensions, par rapport aux résultats subjectifs de la détermination de l'espace perceptif.

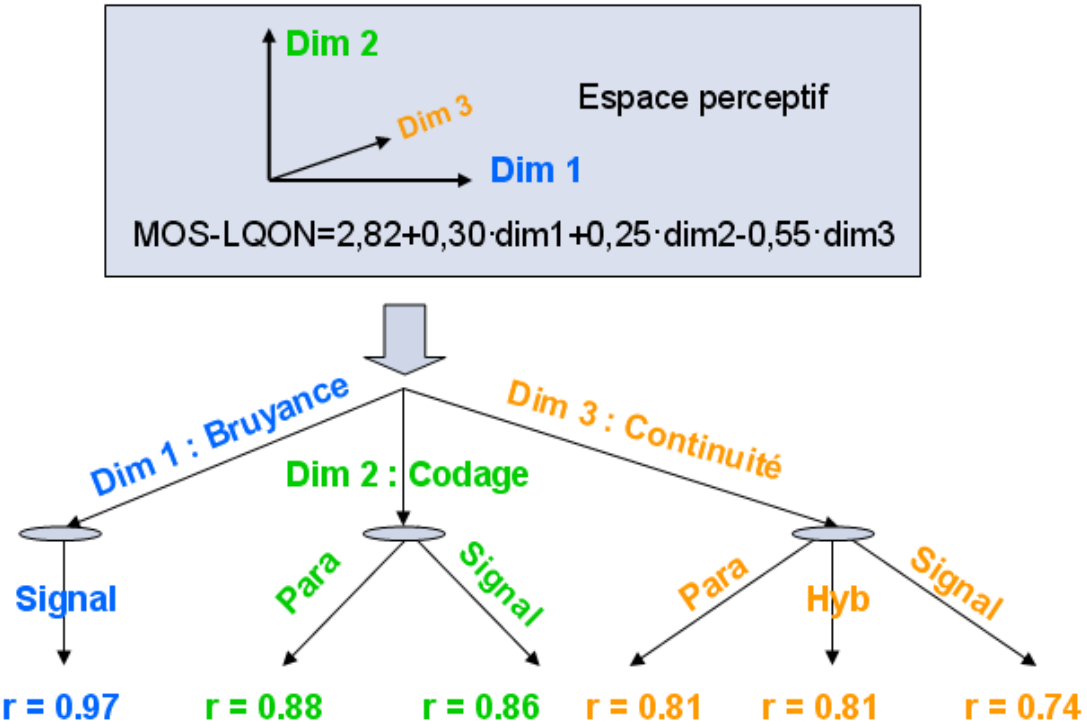


Fig. V.14 Structure globale du modèle hybride, basée sur 3 dimensions perceptives prédites par différents types d'indicateurs (paramétrique, signal et hybride). Les performances de chacun des indicateurs sont données par la corrélation de Pearson *r* entre les dimensions issues de l'espace perceptif et les dimensions prédites.

Le modèle DESQHI permet de s'adapter en fonction du contexte de mesure et des informations disponibles au point de mesure. Il peut être entièrement basé sur des indicateurs paramétriques si l'on ne tient pas compte de la dimension bruyance, ou entièrement basé sur le signal, ou encore hybride (cinq versions hybrides disponibles). En tout, sept adaptations du modèle DESQHI sont disponibles à partir d'un même cœur multidimensionnel. Les performances globales du modèle DESQHI sont présentées dans le dernier chapitre suivant le type d'indicateurs utilisés.

La modélisation des trois dimensions par les différents indicateurs basés sur le signal permet aussi d'identifier le type de dégradation suivant chacune des trois dimensions. Ces informations supplémentaires constituent le diagnostic avancé de la télécommunication proposé par la suite (cf. §.VI.1.2).

Chapitre VI. Performances du modèle DESQHI

La qualité vocale est estimée par la relation linéaire reliant les trois dimensions perceptives (cf. §.III.4). Ces dimensions correspondent à la bruyance, au codage de la parole et à la continuité. Les deux chapitres précédents ont permis de définir les indicateurs paramétriques, basés sur le signal et hybrides, afin de représenter chacune de ces trois dimensions.

Le modèle DESQHI (***Diagnostic and Evaluation of Speech Quality using Hybrid Indicators***) est adaptatif car il peut utiliser différents types d'indicateurs pour chacune des trois dimensions. L'expérimentateur a la possibilité de choisir le type de modèle (paramétrique, signal ou hybride) suivant les informations disponibles au point de mesure, suivant la performance du modèle, suivant les contraintes de temps CPU utilisé pour une application en temps réel ou de planification, et encore suivant le type de dégradation testé (cf. Fig. V.14). En tout, sept versions du modèle sont disponibles pour un unique cœur multidimensionnel.

Dans ce dernier chapitre, les performances du modèle DESQHI sont déterminées à partir du coefficient de corrélation de Pearson *r* et de l'erreur absolue moyenne *EAM* :

$$EAM = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - y_i\right|, \qquad \textbf{Eq. VI.1}$$

avec *N* le nombre de stimuli, *x* les notes MOS-LQSN issues du test subjectif et *y* les notes MOS-LQON prédites par le modèle testé.
Les résultats de ce chapitre sont présentés par une combinaison linéaire entre les MOS-LQSN et les MOS-LQON, en appliquant une fonction de mapping linéaire[16] aux MOS-LQON par rapport aux MOS-LQSN.

Les performances du modèle DESQHI sont déterminées à partir de la base sonore utilisée lors de la détermination de l'espace perceptif. Ensuite, le modèle DESQHI est appliqué à plusieurs bases sonores inconnues du modèle.

VI.1. Performances globales du modèle DESQHI

Les performances de DESQHI sont calculées sur la base sonore constituée des 46 stimuli décrits dans le Tab. III.2. Ces performances sont déterminées pour chacune des sept versions disponibles du modèle en comparant les notes prédites par DESQHI et les notes issues du test subjectif ACR.

Le diagnostic de la qualité vocale proposé par DESQHI est ensuite présenté dans une deuxième partie. La troisième partie compare les performances du modèle DESQHI aux modèles existants.

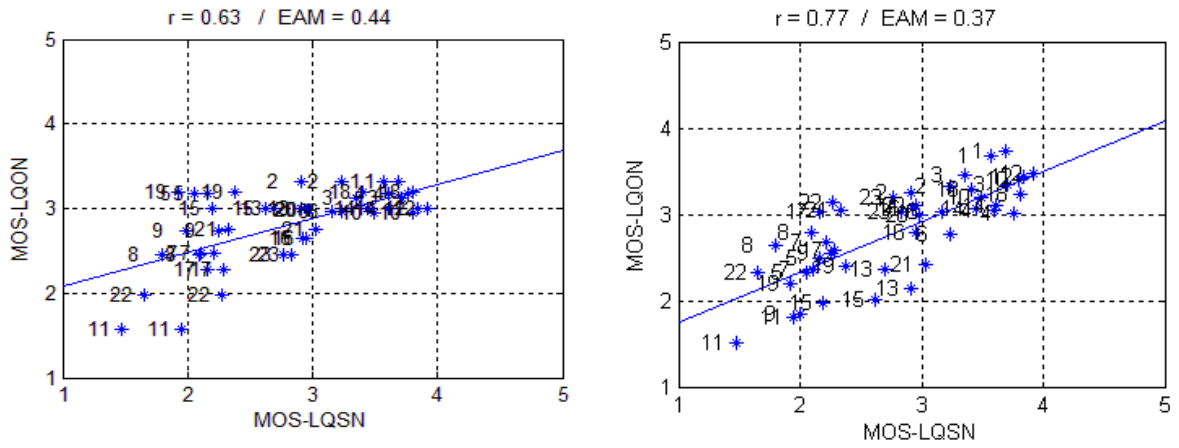
VI.1.1. Prédiction de la qualité vocale

Le modèle DESQHI utilisant uniquement des indicateurs paramétriques est disponible en considérant que les conditions de dégradations testées ne sont pas bruitées. Pour cela, la première dimension bruyance est renseignée par sa valeur par défaut (*DIM_1 = 1,2*). La deuxième dimension codage de la parole est représentée par le facteur *Ie* en considérant uniquement le dernier codage utilisé car le transcodage n'est actuellement pas disponible dans les

[16] Une fonction de mapping linéaire est une régression linéaire qui est appliquée afin de faire correspondre les MOS-LQON aux MOS-LQSN, en s'affranchissant de l'effet des différences de point d'ancrage entre les différents tests subjectifs.

informations issues du réseau (cf. §.V.1.1). La troisième dimension continuité est représentée par l'indicateur *pgd* (cf. §.V.2.1). Dans ce cas, la performance de DESQHI est de $r = 0{,}63$, comme le montre la Fig. VI.1 de gauche.

La version de DESQHI utilisant uniquement des indicateurs basés sur le signal obtient quant à elle une performance de $r = 0{,}77$ (cf. Fig. VI.1 de droite).



DESQHI paramétrique DESQHI signal

Fig. VI.1 Performances du modèle DESQHI utilisant uniquement des indicateurs paramétriques (à gauche) et uniquement basés sur le signal (à droite). Les stimuli sont numérotés de 1 à 23, ce qui correspond aux conditions de dégradations (cf. Tab. III.2).

Le modèle paramétrique obtient une performance moindre car il ne dispose pas des informations relatives à la bruyance des stimuli, et qu'il ne permet pas de prendre en compte le transcodage. Le modèle basé sur le signal obtient une meilleure performance, mais il nécessite des ressources de calcul un peu plus importantes (cf. Tab. VI.1).

Cinq versions hybrides du modèle DESQHI sont ensuite appliquées aux 46 stimuli. Les performances des sept versions sont résumées dans le Tab. VI.1. Le temps de calcul CPU correspondant à la prédiction de l'ensemble de la base sonore (46 stimuli) est précisé en secondes pour chacune des versions du modèle.

Bruyance	Codage	Continuité	*r*	*EAM*	*tps*
Paramétrique	Paramétrique	Paramétrique	0,63	0,44	4,7
Signal	Paramétrique	Paramétrique	0,88	0,25	7,0
Signal	Paramétrique	Hybride	0,89	0,24	7,5
Signal	Paramétrique	Signal	0,76	0,37	16,0
Signal	Signal	Paramétrique	0,90	0,24	11,1
Signal	Signal	Hybride	0,91	0,22	11,0
Signal	Signal	Signal	0,77	0,37	20,2

Tab. VI.1 Performances des 7 versions du modèle DESQHI sur la base sonore connue du modèle, à partir des coefficients de corrélation *r* et des erreurs absolues moyennes *EAM* entre les MOS-LQSN et les MOS-LQON. Le temps de calcul CPU *tps* des différentes versions est précisé en secondes pour les 46 stimuli.

Quatre versions de DESQHI obtiennent de bonnes performances ($r \approx 0{,}9$). Elles correspondent aux versions qui n'utilisent pas l'indicateur basé sur le signal de la dimension continuité.

Par exemple, il est présenté sur la Fig. VI.2 la performance de la version de DESQHI utilisant les indicateurs basés sur le signal pour les dimensions bruyance et codage de la parole, et l'indicateur hybride pour la continuité.

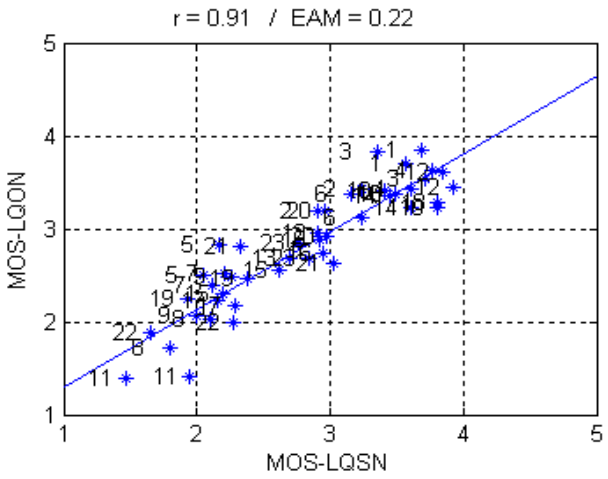


Fig. VI.2 Performance du modèle DESQHI hybride (basé sur le signal pour les dimensions bruyance et codage de la parole, et hybride pour la continuité), à partir du coefficient de corrélation de Pearson entre les notes MOS-LQSN et MOS-LQON. Les conditions de dégradation sont numérotées de 1 à 23 (cf. Tab. III.2).

La dimension codage de la parole est mieux représentée par l'indicateur basé sur le signal que l'indicateur paramétrique car le transcodage n'est pas pris en compte par ce dernier. Si le transcodage pouvait être disponible par l'indicateur paramétrique, les différentes versions hybrides de DESQHI obtiendraient alors des performances similaires à celles obtenues lorsque l'indicateur du codage de la parole est basé sur le signal. Il serait alors préférable de choisir l'indicateur paramétrique pour des raisons de temps de calcul CPU.

Lorsque le transcodage n'est pas disponible et que les ressources CPU le permettent, il est préférable d'opter pour l'indicateur basé sur le signal.

Nous remarquons systématiquement une baisse des performances de DESQHI lors de la modélisation de la dimension continuité par l'indicateur basé sur le signal. Ce résultat avait déjà été relevé dans la partie V.2.3. Il est nécessaire d'approfondir les recherches afin d'améliorer la représentation de cette dimension par des indicateurs basés sur le signal.

Nous avons formulé l'hypothèse que la différence de perception des discontinuités remarquée dans le paragraphe III.2 est due à la différence de localisation des pertes suivant les deux phrases prononcées par les locuteurs homme et femme (cf. §.III.3.3.3). Plus précisément, cette hypothèse concerne la localisation des pertes entre les zones actives et non-actives de la parole. L'indicateur hybride de la dimension de la continuité prend en compte cette différence en comptabilisant le taux d'erreurs uniquement sur les zones actives de la parole, contrairement à l'indicateur paramétrique qui détermine le taux d'erreurs sur le signal entier. Cependant, cette hypothèse n'est pas vérifiée car l'utilisation de l'indicateur hybride n'améliore que très légèrement les performances du modèle DESQHI par rapport à l'utilisation de l'indicateur paramétrique (de $r = 0{,}88$ à $r = 0{,}89$; cf. Tab. VI.1).

VI.1.2. Diagnostics de la qualité vocale

Les modèles actuels permettent de prédire des notes globales de qualité vocale. Dans une situation où l'estimation de la qualité vocale est similaire entre deux stimuli, il est possible que les dégradations perçues soient complètement différentes. La note globale MOS-LQO ne permet pas d'identifier la cause de la dégradation et ne permet pas d'orienter l'expérimentateur afin de proposer des solutions d'amélioration de la qualité vocale.

Le modèle DESQHI propose, en plus de l'évaluation de la qualité globale, deux types de diagnostics de la qualité vocale, dans le but de cibler la cause de la diminution de la qualité vocale et de pouvoir proposer des solutions, dans un contexte de contrôle en temps réel :

- Le premier consiste à exprimer les trois notes MOS relatives à chacune des trois dimensions.
- Le deuxième appelé **diagnostic avancé** consiste à identifier les causes physiques de la dégradation de la qualité vocale, comme par exemple, le type de bruit de fond présent sur le signal vocal ou le type de codage employé ou encore le type de discontinuité.

Le **diagnostic** est réalisé grâce à la structure multidimensionnelle du modèle en déterminant une note relative de qualité vocale pour chacune des trois dimensions perceptives. Ces notes relatives seront appelées par la suite *RMOS* (score moyen d'opinion relatif). Ces trois notes sont déterminées à partir de la relation linéaire qui relie les trois dimensions à la note globale de qualité vocale. La note relative à chacune des trois dimensions est calculée en fixant les deux dimensions non considérées à leurs valeurs par défaut (sans dégradation) (cf. Eq. VI.2).

```
RMOS_DIM_1=2.81+0.297*DIM_1+(0.254*2-0.548*(-1.5));
RMOS_DIM_2=2.81+0.254*DIM_2+(0.297*1.2-0.548*(-1.5));
RMOS_DIM_3=2.81-0.548*DIM_3+(0.297*1.2+0.254*2);
```

Eq. VI.2

Autrement dit, ces trois notes relatives correspondent à la qualité vocale perçue d'un échantillon sonore, en considérant uniquement les dégradations causées par la dimension perceptive analysée.

Un exemple de diagnostic est présenté pour trois conditions de dégradation ayant des notes globales de qualité vocale similaires entre elles (*MOS-LQSN* ≈ *2,2* cf. Fig. III.6), mais avec des conditions de dégradations différentes. Les conditions de dégradation retenues sont les conditions 5, 7 et 16 prononcées par la voix de femme (cf. Tab. III.2). Elles correspondent respectivement à des dégradations de bruit de fond, d'erreurs de bits et d'une combinaison entre le codage et les pertes de paquets.

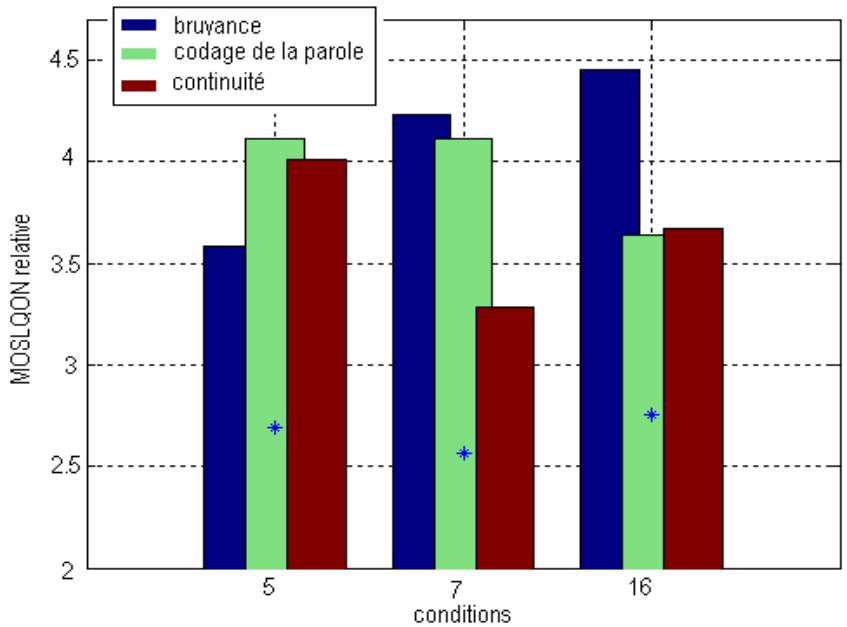


Fig. VI.3 Diagnostic de la qualité vocale représenté par les notes MOS-LQON relatives à la bruyance, au codage de la parole et à la continuité, pour les conditions 5, 7 et 16. La note de qualité globale prédite par DESQHI est représentée par l'étoile (*)

Dans le cas des trois conditions de dégradation 5, 7 et 16, les notes globales de qualité vocale obtenues par DESQHI sont similaires (MOS-LQON ≈ 2,6). Le diagnostic permet alors d'observer les différences entre ces trois conditions de dégradation. La condition de dégrada-

tion *5* obtient une note relative à la bruyance de *RMOS = 3,55* et d'environ *RMOS = 4* pour les dimensions codage de la parole et continuité. Le principal problème pour ce stimulus est la présence d'un bruit de fond sur le signal de la parole.

Dans le cas de la condition 7, les notes relatives à la bruyance et au codage sont acceptables, par contre, il apparaît clairement un problème de discontinuité sur le signal.

La condition 16 montre un problème lié à la fois au codage de la parole ainsi qu'à la présence de certaines discontinuités. Il n'y a aucun problème de bruyance.

Le modèle DESQHI propose aussi un **diagnostic avancé** pour chacune des trois dimensions, à partir des différents indicateurs basés sur le signal.

- Dans le cas de la dimension bruyance, un algorithme de classification automatique permet d'identifier le type de bruit de fond présent sur le signal vocal parmi l'une des quatre classes de bruits de fond (*intelligible*, *environnement*, *souffle* et *grésillement*) (cf. §.IV.3.1).
- Dans le cas de la dimension codage de la parole, un outil est proposé afin d'identifier le type de codec utilisé lors de la télécommunication. Cet outil ne permet pas d'identifier exactement le ou les codec(s) utilisé(s) mais permet d'orienter l'expérimentateur du type de codec utilisé selon six classes (cf. Fig. V.5).
- Dans le cas de la troisième dimension, le type de discontinuité peut être identifié parmi trois classes. La première correspond aux stimuli continus ou présentant des pertes de paquets atténués par un algorithme de PLC. La deuxième classe comprend les stimuli dégradés par des pertes de paquets qui ne sont pas atténuées par un algorithme de PLC. Enfin, la troisième classe correspond aux stimuli dégradés par des erreurs de bits (cf. §.V.2.3).

Par exemple, le diagnostic avancé est présenté sur le Tab. VI.2 pour les trois conditions utilisées précédemment (5, 7 et 16 prononcées par la voix de femme (cf. Tab. III.2)).

Dimension / Condition	Bruyance	Codage	Continuité
5	Souffle	Classe 2 (GSMFR)	Continu
7	Souffle	Classe 2 (GSMFR)	Erreur de bit
16	Souffle	Classe 4 (G.729*2)	Continu

Tab. VI.2 Diagnostic avancé de la qualité vocale des conditions 5, 7 et 16 prononcées par la voix de femme, selon les trois dimensions

Le diagnostic avancé de la bruyance identifie le bruit comme étant de la classe souffle, ce qui correspond bien aux trois conditions de dégradation (bruit rose et bruit résiduel).

Le diagnostic avancé du codage de la parole n'est pas bien adapté à notre base sonore car il a été réalisé sur une base sonore ne comprenant pas le codec GSM-EFR (cf. §.V.1.2). Les types de codec déterminés pour ces trois conditions sont néanmoins cohérents avec les notes MOS relatives à la deuxième dimension présentées précédemment.

Le diagnostic avancé de la continuité permet de préciser que la condition 7 présente des erreurs de bit, tandis que les conditions 5 et 16 ne comportent pas de coupures nettes du signal. En combinant les résultats des deux diagnostics de cette dimension, on peut préciser que la condition 5 ne comporte pas de discontinuité (*RMOS = 4* ; cf. Fig. VI.3), tandis que la condition 16 présente des pertes de paquets atténuées par un algorithme de PLC (*RMOS = 3,6* ; cf. Fig. VI.3).

VI.1.3. Comparaison avec les modèles existants

Trois versions du modèle DESQHI sont comparées avec les modèles existants, suivant le type d'indicateurs utilisés (modèle basé sur le signal P.563 [2] et modèle paramétrique G.107 [3]). Le modèle DESQHI est testé par sa version paramétrique, sa version basée sur le signal et sa version hybride (indicateur basé sur le signal pour la bruyance et le codage de la parole, et indicateur hybride pour la continuité).

Indicateur / Modèle	paramétrique	Signal	Hybride
Modèle E	r = 0,47 / EAM = 0,51 0,14 sec		
P.563		r = 0,51 / EAM = 0,48 33,1 sec	
DESQHI	r = 0,63 / EAM = 0,44 4,7 sec	r = 0,77 / EAM = 0,37 20,2 sec	r = 0,91 / EAM = 0,22 11,0 sec
Caractéristiques	-Pas de BDF - pas de transcodage	Ø	Ø

Tab. VI.3 Comparaison des performances du modèle DESQHI avec les modèles existants. Le temps de calcul CPU est donné en secondes pour chacun des modèles, pour la base sonore entière (46 stimuli).

Le modèle E paramétrique (G.107) est comparé à la version paramétrique de DESQHI en ne prenant pas en compte les dégradations liées à la bruyance, ni au transcodage. De ce fait, les deux modèles obtiennent des performances assez limitées avec cependant un avantage pour le modèle DESQHI.

Dans le cas des modèles basés sur le signal, DESQHI est comparé avec le modèle P.563. Il obtient une performance supérieure (*r = 0,77* contre *r = 0,51*) pour un temps de calcul CPU inférieur à celui nécessaire au modèle P.563 (20,2 sec contre 33,1 sec pour la prédiction des 46 stimuli). Le modèle DESQHI et le modèle E sont programmés sous Matlab, tandis que les modèles PESQ et P.563 sont programmés en C. Les temps indiqués ne prennent pas en compte cette différence qui joue en faveur des modèles PESQ et P.563.

DESQHI propose en plus de la version paramétrique et de la version basée sur le signal, une version hybride permettant d'améliorer la performance globale de la prédiction de la qualité vocale, tout en limitant le temps de calcul CPU par rapport au modèle basé sur le signal.

Pour information, le modèle PESQ a été appliqué à notre base sonore et il obtient seulement une performance de *r = 0,62 / EAM = 0,46* et un temps de calcul CPU de 29,7 secondes. Cela montre la difficulté du modèle PESQ à représenter les diverses combinaisons de dégradations présentes dans notre base sonore.

Le modèle DESQHI obtient globalement de bonnes performances lorsqu'il est appliqué à la base sonore utilisée pour sa construction. Il est nécessaire de vérifier les performances de DESQHI sur différentes bases sonores inconnues du modèle afin de tester sa résistance et sa fiabilité sur d'autres conditions de dégradations.

VI.2. Performance du modèle DESQHI sur des bases sonores inconnues

Les bases sonores retenues pour la validation du modèle DESQHI sont sélectionnées par rapport aux conditions de dégradation des stimuli, mais aussi par rapport aux types de notes subjectives d'évaluation de la qualité vocale (notes MOS cf. §.I.2). Deux bases sonores issues du supplément 23 de la série P de l'UIT [96] sont tout d'abord utilisées. Ensuite, la base sonore construite par Gautier-Turbin et Gros [92] est utilisée pour tester les performances de DESQHI. Plusieurs bases sonores utilisées lors de la validation du modèle PESQ et POLQA sont aussi appliquées à notre modèle DESQHI.

Les différentes versions du modèle DESQHI sont testées pour chacune des bases sonores. Cependant, l'indicateur hybride de la dimension continuité ne peut être appliqué en raison de l'indisponibilité du pattern d'erreurs des discontinuités sur le signal vocal pour toutes les bases sonores utilisées. Les deux versions de DESQHI utilisant cet indicateur ne sont donc pas représentées. La dimension bruyance est dans tous les cas représentée par des indicateurs basés sur le signal car les statistiques du réseau ne permettent pas pour le moment, d'obtenir des informations sur la présence de bruit de fond. Les quatre versions de DESQHI sont appliquées aux différentes bases sonores afin d'analyser les performances de chacune des versions du modèle DESQHI. Ces performances sont ensuite comparées aux modèles existants (modèle E (paramétrique) et modèle P.563 (basé sur le signal)).

VI.2.1. Bases sonores du supplément 23

Il existe trois expériences associées à trois bases sonores dans le supplément 23 [96]. La deuxième base sonore aurait été intéressante à utiliser car elle est constituée de diverses conditions bruitées, cependant les notes subjectives correspondantes sont données uniquement par la CMOS à partir d'un test CCR (cf. §.I.2.3). L'échelle de mesure de la qualité vocale ne correspond pas à celle utilisée par DESQHI. De plus, chacun des signaux de référence correspondant aux signaux dégradés sont aussi bruités. Les notes CMOS ne contiennent pas les informations relatives à la présence de bruits de fond. Cette base sonore n'est donc pas utilisée pour valider les performances de notre modèle.

Les bases 1 et 3 sont associées à des notes MOS-LQSN obtenues par des tests ACR (cf. §.I.2.1). Ces deux bases sont utilisées pour analyser les performances de DESQHI.

VI.2.1.1. Expérience 1 du supplément 23

La première expérience du supplément 23 est constituée de 44 conditions de dégradation, présentant différents codages et transcodages (G.729, G.726, G.728, G.711, GSM-FR, IS-54 (North American VSELP), JD-HR (half-rate Japanese digital cellular), MNRU). Ces 44 conditions sont prononcées par deux locuteurs femmes et deux locuteurs hommes. Ce test a été réalisé trois fois pour trois langues différentes (français, anglais américain, japonais), pour un total de 528 stimuli. Les notes subjectives d'évaluation de la qualité vocale de ces 528 stimuli ont été obtenues à partir du test subjectif ACR décrit dans le paragraphe I.2.1.

Cette base sonore a été utilisée pour construire l'indicateur basé sur le signal de la dimension du codage de la parole (cf. §.V.1.2). Elle n'est donc pas totalement inconnue pour les versions utilisant cet indicateur. Par contre elle reste une base sonore inconnue lorsque la dimension du codage de la parole est représentée par l'indicateur paramétrique.

Quatre versions du modèle DESQHI sont utilisées afin de prédire la qualité vocale de ces 528 stimuli. Les performances globales du modèle DESQHI sont présentées dans le Tab. VI.4.

Codage / Continuité	Paramétrique		Signal	
	r	***EAM***	***r***	***EAM***
Paramétrique	*0,49*	*0,57*	*0,71*	*0,45*
Signal	*0,20*	*0,64*	*0,27*	*0,62*

Tab. VI.4 Performances des différentes versions du modèle DESQHI sur l'exp. 1 du supl. 23, à partir des coefficients de corrélation *r* et des erreurs absolues moyennes *EAM* entre les MOS-LQSN et les MOS-LQON. La dimension bruyance est représentée par des indicateurs basés sur le signal.

Au sujet de la dimension du codage de la parole, nous constatons que les performances du modèle DESQHI utilisant l'indicateur paramétrique sont beaucoup moins bonnes que lors de l'utilisation de l'indicateur basé sur le signal (cf. Tab. VI.4). Cette baisse de performance est due à l'indicateur paramétrique qui ne prend pas en compte le transcodage. Lorsque l'indicateur paramétrique est renseigné par la somme des *Ie* suivant les différents codages successifs (transcodage), le modèle DESQHI obtient des performances similaires par rapport aux versions utilisant l'indicateur basé sur le signal. Dans ce cas de transcodage, le modèle DESQHI hybride (respectivement signal, paramétrique et paramétrique) obtient une performance de *r = 0,70 / EAM = 0,46*, tandis que le modèle DESQHI hybride (respectivement signal, paramétrique et signal) obtient une performance de *r = 0,32 / EAM = 0,62*.

Cette observation confirme la conclusion démontrée dans le paragraphe VI.1.1 précédent : lorsque les informations sur le codage et le transcodage de la parole sont disponibles, il est judicieux d'utiliser l'indicateur paramétrique. Dans le cas où l'information sur le transcodage n'est pas disponible mais seulement le dernier codage, il vaut mieux utiliser l'indicateur basé sur le signal, si les ressources en temps CPU le permettent.

L'indicateur basé sur le signal de la dimension continuité n'est pas adapté à cette base sonore, comme le montre la diminution des performances du modèle DESQHI lors de l'utilisation de cet indicateur (*r = 0,20 et r = 0,27*).

Le modèle P.563 (basé sur le signal) est appliqué à cette base sonore. Il obtient une performance de *r = 0,72 / EAM = 0,*45 (cf. Fig. VI.4). Ce modèle nécessite cependant de plus grandes ressources en temps de calcul CPU comparativement au modèle DESQHI, car il nécessite de calculer un grand nombre d'indicateurs.

Le modèle E (paramétrique) obtient une performance *de r = 0,49 / EAM = 0,57*. Lorsque nous appliquons ce modèle en prenant en compte le transcodage, sa performance devient alors de *r = 0,71 / EAM = 0,46.* Le modèle DESQHI hybride (respectivement signal, signal, paramétrique) obtient une performance similaire au modèle P.563 (*r = 0,71* cf. Fig. VI.4) et obtient une meilleure performance que le modèle E.

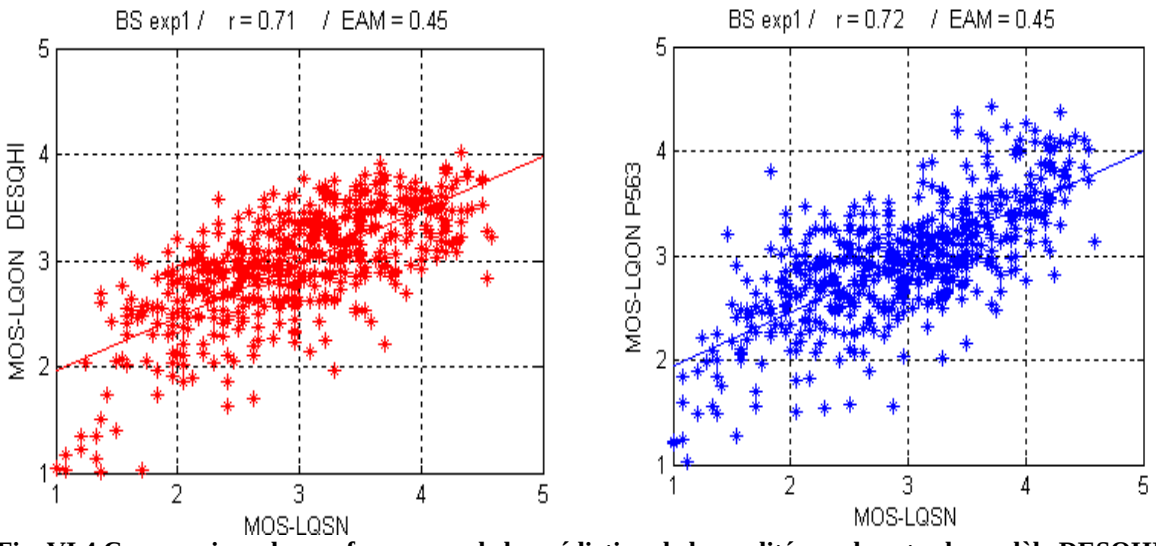


Fig. VI.4 Comparaison des performances de la prédiction de la qualité vocale entre le modèle DESQHI hybride (signal, signal, paramétrique) (à gauche) et le modèle P.563 (à droite)

Une analyse par la méthode bootstrap a aussi été réalisée afin de confirmer que les modèles DESQHI et P.563 sont bien fiables sur cette base sonore. L'analyse bootstrap a consisté à évaluer un grand nombre de fois les performances de ces modèles, à partir d'une sélection aléatoire avec remise d'un nombre défini de stimuli de la base sonore. Les modèles peuvent être considérés comme fiables si les différentes performances obtenues sont cohérentes entre elles. Nous avons choisi de sélectionner 60 échantillons aléatoirement avec remise afin d'évaluer les performances pour 30 combinaisons différentes.

Les résultats sont présentés à partir des moyennes et des écarts types des performances obtenues suivant les 30 combinaisons différentes (cf. Tab. VI.5). Les moyennes des performances montrent que les deux modèles sont fiables puisqu'ils obtiennent des valeurs similaires aux performances mesurées sur la totalité de la base sonore ($r \approx 0,7$). Les écarts types *ect* montrent que les modèles sont stables d'une combinaison à l'autre.

	Moyenne des performances	Ecart type
DESQHI	$r_{moy} = 0,69$	ect = 0,057
P.563	$r_{moy} = 0,71$	ect = 0,076

Tab. VI.5 Performances des modèle DESQHI et P.563 à partir d'une analyse bootstrap réalisée 30 fois à partir de 60 stimuli choisis de manière aléatoire avec remise, à partir de la base sonore de l'expérience 1 du supplément 23

Le modèle hybride DESQHI est aussi capable de préciser que les dégradations sont liées au codage de la parole, et non à la continuité du signal grâce au diagnostic de la télécommunication. Le bruit de fond généré par le codage est identifié par la classification automatique comme étant du bruit de souffle, ce qui permet d'orienter l'expérimentateur sur le type de codage utilisé. Par exemple, s'il y a un certain niveau de bruit de fond, il est plus probable que le codec utilisé soit le G.711 plutôt que le G.729 (cf. §.III.3.3.2). Ce point pourrait améliorer les performances du diagnostic avancé du codage de la parole.

VI.2.1.2. Expérience 3 du supplément 23

La base sonore de l'exp. 3 du supplément 23 [96] est constituée de 50 conditions de dégradation :

- codec G.729 proposé seul ou en combinaison,
- trois types de bruits de fond,
- différents taux de pertes de paquets FER diffusés aléatoirement ou par rafale,

- erreurs de bits de type BER.

Les trois bruits de fond correspondent à un bruit de véhicule, à un bruit de rue et à un bruit gaussien aléatoire de spectre fréquentiel similaire à celui de la parole ("hoth noise").

Les pertes de paquets sont présentées à des taux allant de 0 à 13 %, de manière aléatoire ou bien par rafales.

Les erreurs de bits de type BER correspondent à des erreurs de codage sur des transmissions IP. Ce type de dégradation n'est pas pris en compte par DESQHI. Les huit conditions de dégradations correspondantes sont donc exclues de notre analyse.

Les 42 conditions de dégradations sont appliquées à deux locuteurs femmes et deux locuteurs hommes. Quatre tests ACR ont été réalisés pour quatre langues différentes (français, anglais, japonais, allemand). Cette base sonore est constituée de 672 stimuli. Quatre versions de DESQHI sont appliquées à cette base sonore. Les performances sont présentées dans le Tab. VI.6.

Codage / Continuité	Paramétrique		Signal	
	r	***EAM***	***r***	***EAM***
Paramétrique	*0,70*	*0,50*	*0,72*	*0,49*
Signal	*0,49*	*0,60*	*0,45*	*0,62*

Tab. VI.6 Performances des différentes versions du modèle DESQHI sur l'exp. 3 du supl. 23, à partir des coefficients de corrélation *r* et des erreurs absolues moyennes *EAM* entre les MOS-LQSN et les MOS-LQON. La dimension bruyance est représentée par des indicateurs basés sur le signal.

Au sujet de la modélisation du codage de la parole, nous observons sur le Tab. VI.6 que les performances du modèle DESQHI obtenues par l'utilisation des deux types d'indicateurs (paramétrique et basé sur le signal) sont similaires. Lorsque le transcodage est pris en compte par l'indicateur paramétrique, les performances du modèle DESQHI sont meilleures : Par exemple, la performance de la version DESQHI (respectivement signal, paramétrique, paramétrique) augmente à $r = 0,77 / EAM = 0,44$. L'indicateur paramétrique *Ie* obtient de bonnes performances pour représenter les dégradations relatives au codage de la parole lorsque les différents codecs successifs sont connus. Sinon, il est préférable d'opter pour l'indicateur basé sur le signal.

La représentation de la continuité de la parole par l'indicateur basé sur le signal fait chuter considérablement les performances du modèle DESQHI ($r = 0,49$ et $r = 0,45$), ce qui porte à croire que cet indicateur n'est pas adapté pour des bases sonores différentes de celle utilisée lors de la construction du modèle. Il vaut mieux, dans ce cas utiliser l'indicateur paramétrique ($r = 0,70$ et $r = 0,72$).
Les performances de DESQHI obtenues sur ces deux bases sonores (expériences 1 et 3 du supplément 23) montrent qu'il est préférable d'utiliser les indicateurs basés sur le signal pour représenter les dimensions de la bruyance et du codage de la parole, et l'indicateur paramétrique pour la dimension de la continuité.

Le modèle P.563 (basé sur le signal) est appliqué à cette base sonore. Il obtient une performance de $r = 0,72 / EAM = 0,48$ (cf. Fig. VI.5). Le modèle E (paramétrique) obtient une performance *de* $r = 0,44 / EAM = 0,64$. Lorsque le transcodage est pris en compte, la performance du modèle E augmente à $r = 0,47 / EAM = 0,64$.

Le modèle E est moins performant car il ne tient pas compte du bruit de fond, ni du transcodage.

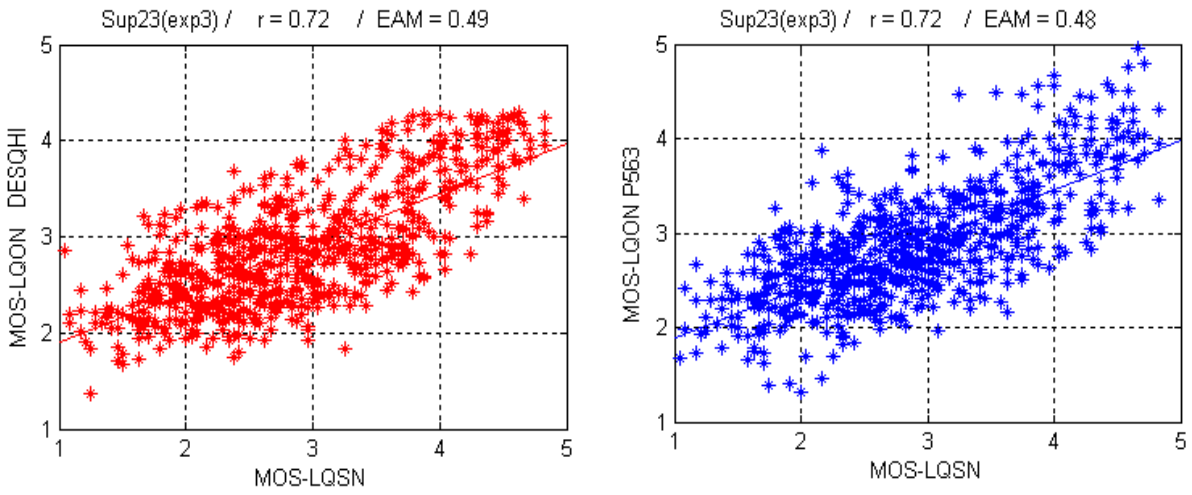


Fig. VI.5 Comparaison des performances de la prédiction de la qualité vocale entre le modèle DESQHI hybride (signal, signal, paramétrique) (à gauche) et le modèle P.563 (à droite)

Une analyse par la méthode bootstrap identique à celle réalisée avec la base sonore de l'expérience 1 est proposée afin de vérifier la fiabilité des modèles DESQHI et P.563. Soixante échantillons ont été sélectionnés aléatoirement avec remise afin d'évaluer les performances de 30 combinaisons différentes.

Les résultats sont présentés à partir des moyennes et des écarts types des performances obtenues suivant les 30 combinaisons différentes (cf. Tab. VI.7). Les moyennes des performances montrent que les deux modèles sont fiables puisqu'ils obtiennent des valeurs similaires aux performances mesurées sur la totalité de la base sonore ($r \approx 0,72$). Les écarts types *ect* montrent que les modèles sont stables d'une combinaison à l'autre.

	Moyenne des performances	Ecart type
DESQHI	$r_{moy} = 0,71$	ect = 0,062
P.563	$r_{moy} = 0,72$	ect = 0,064

Tab. VI.7 Performances du modèle DESQHI et du modèle P.563 à partir d'une analyse bootstrap réalisée 30 fois à partir de 60 stimuli choisis de manière aléatoire avec remise, à partir de la base sonore de l'expérience 3 du supplément 23

Le modèle DESQHI hybride (signal, signal, paramétrique) obtient des performances similaires à P.563 ($r = 0,71$ pour l'exp.1 et $r = 0,72$ pour l'exp.3) mais avec un temps de calcul CPU nettement inférieur. Le modèle DESQHI permet aussi de diagnostiquer la qualité vocale en indiquant si les dégradations proviennent d'un problème de bruyance, de codage de la parole ou bien s'il s'agit de la présence de discontinuité sur le signal.

VI.2.2. Base sonore bruitée

La base sonore bruitée de Gautier-Turbin et Gros [92] est appliquée au modèle DESQHI. Cette base sonore a déjà été utilisée lors de la validation du modèle de bruyance (cf. §.IV.4.2). Dans l'étude précédente, cette base sonore avait été utilisée en séparant les conditions de dégradation selon les différents algorithmes de débruitage. Cette distinction avait été faite car le modèle de bruyance ne prend pas en compte les dégradations présentes sur le signal de parole qui sont provoquées par l'utilisation de l'algorithme de débruitage.

Dans cette partie, la totalité de la base sonore est appliquée au modèle DESQHI, afin de tester les performances de prédiction de la qualité vocale en présence de dégradations relatives à différents types de bruits de fond et des dégradations de la parole causées par l'utilisation des algorithmes de débruitage de la parole.

Les performances de DESQHI sont présentées pour quatre versions différentes (cf. Tab. VI.8).

Codage / Continuité	Paramétrique		Signal	
	r	***EAM***	***r***	***EAM***
Paramétrique	*0,22*	*0,61*	*0,53*	*0,53*
Signal	*0,63*	*0,49*	*0,69*	*0,45*

Tab. VI.8 Performances des différentes versions du modèle DESQHI sur la base sonore bruitée [92], à partir des coefficients de corrélation *r* et des erreurs absolues moyennes *EAM* entre les MOS-LQSN et les MOS-LQON. La dimension bruyance est représentée par des indicateurs basés sur le signal.

Les performances de DESQHI montrent en général que les indicateurs utilisant au moins un indicateur paramétrique ne sont pas adaptés aux dégradations présentes dans cette base sonore (pas de perte de paquets, ni d'erreur de bit, ni de dégradation liée au codage de la parole). Cependant, les indicateurs basés sur le signal des trois dimensions perceptives du modèle DESQHI (bruyance, codage de la parole et continuité) permettent de mettre en évidence les dégradations présentes sur ces stimuli, qui sont causées par l'utilisation des algorithmes de débruitage de la parole. Le modèle DESQHI dans sa version entièrement basée sur le signal obtient une performance de *r = 0,69* (cf. Tab. VI.8 et Fig. VI.6).

Le modèle P.563 (basé sur le signal) n'est pas adapté à ce type de dégradation. Il obtient une performance de *r = 0,32* / *EAM = 0,62* (cf. Fig. VI.6). Le modèle E (paramétrique) est inutilisable pour cette base sonore car les dégradations présentes dans cette base sonore ne sont pas disponibles dans les statistiques du réseau (bruit de fond et algorithme de débruitage).

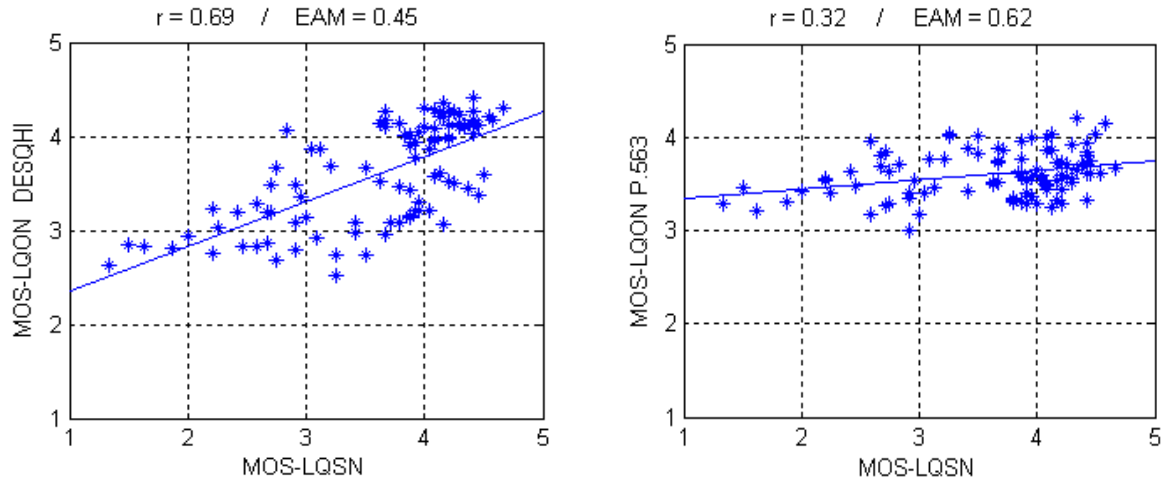


Fig. VI.6 Comparaison des performances de la prédiction de la qualité vocale entre le modèle DESQHI basé sur le signal (à gauche) et le modèle P.563 (à droite)

Les dégradations produites par l'utilisation des algorithmes de débruitage sont prises en compte par le modèle DESQHI dans sa version entièrement basée sur le signal. Ce résultat montre que ce modèle peut être appliqué à des dégradations qui n'ont pas été prises en compte lors de sa construction. Cela confirme que la structure tridimensionnelle du modèle DESQHI est bien adaptée à l'évaluation de la qualité vocale en présence de diverses dégradations présentes lors d'une communication téléphonique.

Les performances de la classification automatique du bruit de fond ont été mesurées en considérant que le bruit de rue est un bruit de souffle car il serait certainement classifié en tant que tel lors d'un test subjectif (bruit de vent qui souffle). Le bruit de bureau ressemble à un bruit rose, il est considéré comme du bruit de souffle. Le bruit de foule correspond à un bruit

d'environnement. Les résultats de la classification montrent d'excellentes performances avec seulement 7 erreurs sur 96 stimuli (7,3 % d'erreur de classification). Certains bruits de fond constitués initialement de bruit d'environnement (bruit de foule), sont identifiés comme du bruit de souffle car l'algorithme de débruitage modifie le bruit. Trois "erreurs" comptabilisées appartiennent à ce cas.

Lorsque l'algorithme de débruitage est appliqué au niveau "agressif" (cf. §.IV.4.2), il apparaît des dégradations sur le signal de parole. Il se peut qu'en voulant supprimer un bruit d'environnement qui n'est pas forcément gênant, les outils de débruitage génèrent d'autres dégradations qui peuvent être encore plus gênantes. Pour cela, un exemple de diagnostic est donné dans cette partie pour quatre conditions plus ou moins débruitées, dans le cas de la voix d'homme en présence du bruit de foule.

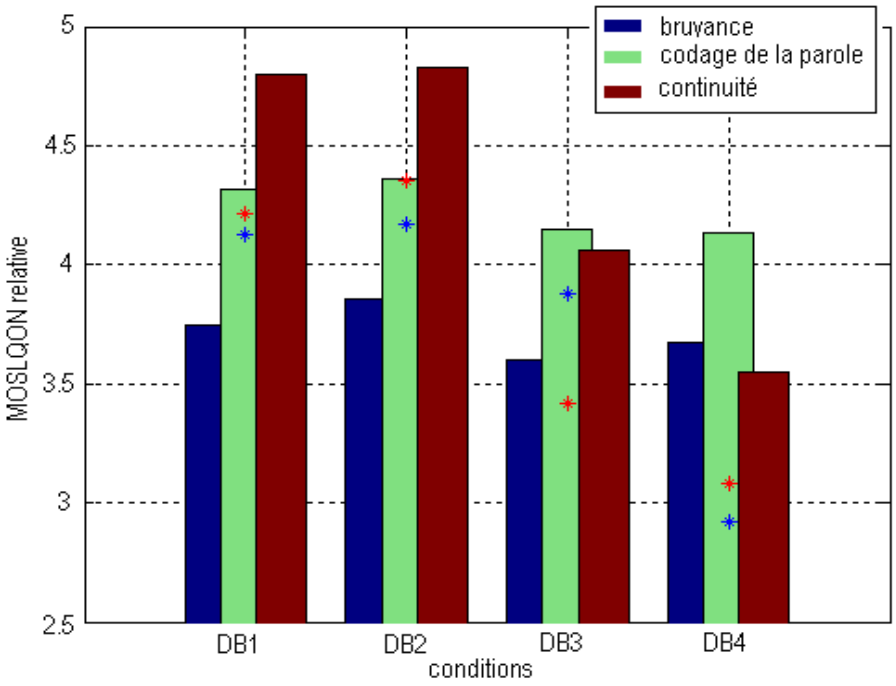


Fig. VI.7 Diagnostic de la qualité vocale (DESQHI basé sur le signal) de 4 conditions débruitées par les 2 algorithmes de débruitage utilisés de manière légère (DB1 et DB3) et agressive (DB2 et DB4). Le bruit de fond est celui de foule diffusé au niveau faible (RSB = 20 dB). Les 3 barres correspondent respectivement au RMOS de la bruyance, du codage de la parole et de la continuité. Les MOS-LQON prédites par DESQHI sont présentées en rouge (*), et les MOS-LQSN le sont en bleu (*).

On remarque pour ces quatre conditions que le deuxième algorithme de débruitage (DB3 et DB4) provoque de la discontinuité sur le signal, contrairement au premier algorithme de débruitage (DB1 et DB2). Cela se répercute sur les notes MOS-LQON et MOS-LQSN. L'utilisation des algorithmes de débruitage agressifs (DB2 et DB4) améliore les notes relatives à la bruyance, par rapport à l'utilisation des algorithmes légers (DB1 et DB3). Dans le cas du premier algorithme de débruitage, il n'y a pas d'impact sur les autres RMOS relatives au codage de la parole et à la continuité. Dans le cas du deuxième algorithme de débruitage, nous remarquons une diminution de la RMOS liée à la continuité. L'application de cet algorithme agressif à ce stimulus réduit donc le niveau sonore du bruit de fond, ce qui entraîne une amélioration de la RMOS de la bruyance, au détriment de la RMOS liée à l'ajout de discontinuité sur le signal.

Le diagnostic de la qualité vocale offre de nombreuses perspectives afin d'approfondir les diverses analyses de l'évaluation de la qualité vocale d'une transmission téléphonique.

VI.2.3. Base sonore "NB_LUC P563"

La base sonore construite en partenariat entre France Telecom, Lucent, Opticom Psytechnics et Swissqual [98] est appliquée à notre modèle DESQHI. Les conditions de dégradations sont au nombre de 50 :

- Un bruit de fond d'environnement diffusé à sept niveaux différents
- Deux terminaux disposés à deux positions différentes
- Cinq conditions MNRU,
- Erreurs de transmissions (taux de pertes de paquets entre 0% et 5% en aléatoire ou par rafales).

Les échantillons de parole ont été enregistrés par les quatre locuteurs en leur diffusant en même temps le bruit de fond au casque. Ce test prend donc en compte l'effet de l'adaptation de la voix des locuteurs en présence de bruit de fond d'environnement (effet Lombard Garnier [4]). Les 200 stimuli sont ensuite égalisés par rapport aux zones actives du signal à -26 dBov d'après la recommandation P.56 [80].
Tous les stimuli de cette base sonore sont codés en EVRC (Enhanced Variable Rate Codec). Le langage utilisé est l'anglais américain du nord.

Codage / Continuité	Paramétrique		Signal	
	r	***EAM***	***r***	***EAM***
Paramétrique	*0,86*	*0,33*	*0,65*	*0,48*
Signal	*0,42*	*0,64*	*0,33*	*0,66*

Tab. VI.9 Performances des différentes versions du modèle DESQHI sur la base sonore LUC [98], à partir des coefficients de corrélation *r* et des erreurs absolues moyennes *EAM* entre les MOS-LQSN et les MOS-LQON. La dimension bruyance est représentée par des indicateurs basés sur le signal.

DESQHI obtient une excellente performance dans sa version hybride (respectivement signal, paramétrique, paramétrique) avec une corrélation de $r = 0,86$ (cf. Tab. VI.9 et Fig. VI.8).

Les conditions de dégradations bruitées sont bien prises en compte par notre modèle. La classification automatique du bruit de fond obtient un pourcentage d'erreur de classification de 6%.

Lorsque la dimension du codage de la parole est représentée par l'indicateur basé sur le signal, la performance du modèle DESQHI diminue à $r = 0,65$. Cet indicateur ne prend pas en compte le codage EVRC, d'où cette diminution de performance.

Les discontinuités présentes sur cette base sonore ne sont pas bien représentées par l'indicateur basé sur le signal.

Le modèle E (paramétrique) et le modèle P.563 (basé sur le signal) sont appliqués à cette base sonore. Ils obtiennent respectivement des performances de $r = 0,22$ / $EAM = 0,69$ et $r = 0,75$ / $EAM = 0,45$ (cf. Fig. VI.8).

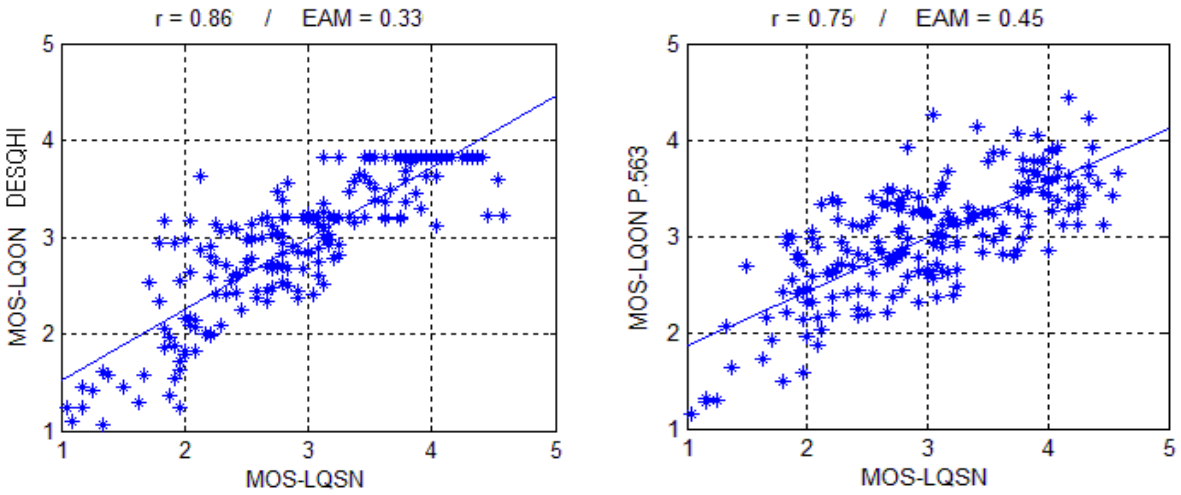


Fig. VI.8 Comparaison des performances de la prédiction de la qualité vocale entre le modèle DESQHI hybride (signal, paramétrique, paramétrique) (à gauche) et le modèle P.563 (à droite)

L'influence de l'utilisation des différents moyens de restitution (combiné, casque, différentes positions) semble bien être représentée par la dimension bruyance du modèle DESQHI (*r = 0,86*). La position du terminal et le type de terminal semblent influencer principalement le rapport signal sur bruit RSB.

VI.2.4. Base sonore "P862_BGN"

Cette base sonore a été construite par Ericsson pour l'IUT-T SG12 Q.13 [99], en vue de valider le modèle PESQ, puis réutilisée pour le modèle POLQA. Les conditions de dégradations sont au nombre de 49 et sont disponibles en anglais et en allemand :

- Un bruit gaussien aléatoire de spectre fréquentiel similaire à celui de la parole ("hoth noise") et un bruit d'environnement, diffusés à deux niveaux sonores différents
- Codecs GSM-EFR, TDMA EFR, G723.1, EVRC
- Six conditions MNRU,
- Erreurs de transmissions (erreurs de bits BER : rapport signal sur interférence C/I allant de 8 dB à 20 dB)

Le modèle DESQHI est appliqué à la base sonore prononcée en anglais, puis à la base sonore prononcée en allemand.

Codage / Continuité	Paramétrique		Signal	
	r	***EAM***	***r***	***EAM***
Paramétrique	*0,79*	*0,47*	*0,77*	*0,53*
Signal	*0,56*	*0,68*	*0,58*	*0,66*

Tab. VI.10 Performances des différentes versions du modèle DESQHI sur la base sonore BT en anglais [99], à partir des coefficients de corrélation *r* et des erreurs absolues moyennes *EAM* entre les MOS-LQSN et les MOS-LQON. La dimension bruyance est représentée par des indicateurs basés sur le signal.

La version de DESQHI la plus performante est obtenue pour la version hybride (respectivement signal, paramétrique, paramétrique), avec une performance de *r = 0,79* (cf. Tab. VI.10 et Fig. VI.9). Les conditions de dégradation ne comprennent pas de transcodage, ce qui justifie la similarité des résultats entre les indicateurs paramétriques et basés sur le signal de la dimension du codage de la parole.

L'indicateur basé sur le signal de la dimension continuité n'est pas adapté à cette base sonore (*r = 0,56*).

Le modèle E et le modèle P.563 sont appliqués à cette même base sonore. Ils obtiennent respectivement des performances de *r = 0,14 / EAM = 0,86* et *r = 0,78 / EAM = 0,50* (cf. Fig. VI.9).

Le modèle DESQHI dans sa version hybride (respectivement signal, paramétrique, paramétrique) obtient une performance similaire au modèle P.563, mais en utilisant un temps CPU inférieur à celui utilisé par P.563. De plus, le modèle DESQHI propose un diagnostic de la transmission en déterminant trois notes RMOS supplémentaires, et en identifiant certaines dégradations physiques, pour chacune des trois dimensions perceptives.

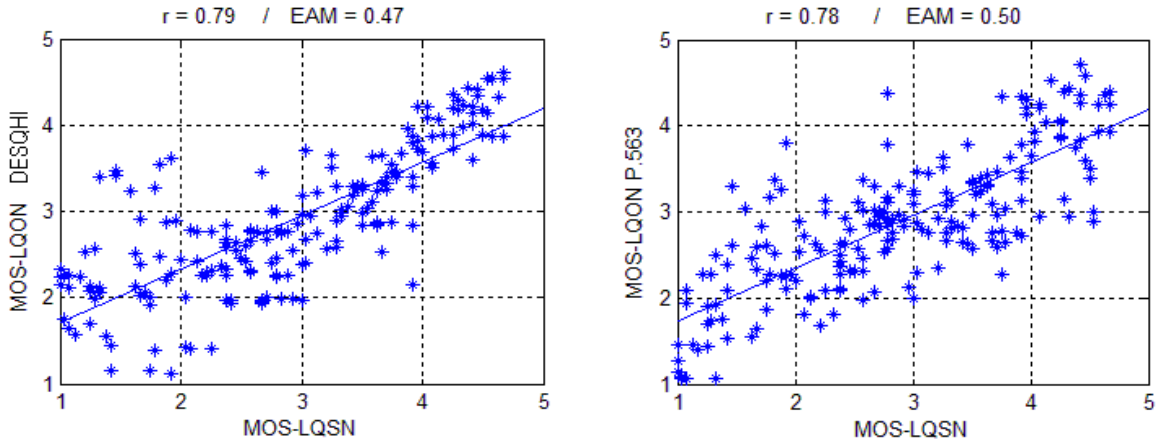


Fig. VI.9 Comparaison des performances de la prédiction de la qualité vocale entre le modèle DESQHI hybride (signal, paramétrique, paramétrique) (à gauche) et le modèle P.563 (à droite) pour la base sonore prononcée en anglais

Cette base sonore utilise un grand nombre de bruits de fond de différents types. La classification automatique du bruit de fond est assez précise avec un taux d'erreur de 14 % sur l'ensemble de la base sonore. Certains problèmes de classification ont été observés lors de fortes dégradations de la parole. Dans ce cas, la DAV est difficile à réaliser et entraîne des erreurs dans la classification automatique du bruit de fond.

Les performances globales du modèle DESQHI appliquées à la base sonore allemande sont présentées dans le Tab. VI.11.

Codage / Continuité	Paramétrique		Signal	
	r	***EAM***	***r***	***EAM***
Paramétrique	*0,77*	*0,51*	*0,72*	*0,89*
Signal	*0,48*	*0,75*	*0,50*	*0,74*

Tab. VI.11 Performances des différentes versions du modèle DESQHI sur la base sonore BT en allemand [99], à partir du coefficient de corrélation *r* et de l'erreur absolue moyenne *EAM* entre les MOS-LQSN et les MOS-LQON. La dimension bruyance est représentée par des indicateurs basés sur le signal.

Globalement, les performances du modèle DESQHI sont légèrement inférieures sur la base sonore allemande à celle en anglais, mais elles sont cohérentes avec les conclusions données précédemment.

Le modèle E et le modèle P.563 sont appliqués à cette base sonore prononcée en allemand. Ils obtiennent respectivement des performances de *r = 0,08 / EAM = 0,89* et *r = 0,83 / EAM = 0,45* (cf. Fig. VI.10). Nous remarquons que le modèle P.563 est plus efficace sur cette base sonore allemande (*r = 0,83*), que sur celle prononcée en anglais (*r = 0,78*).

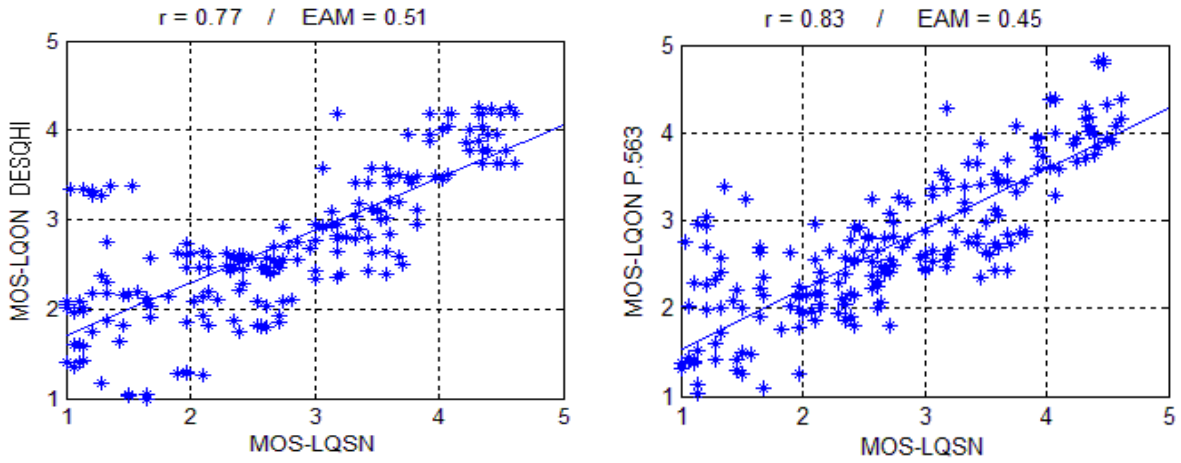


Fig. VI.10 Comparaison des performances de la prédiction de la qualité vocale entre le modèle DESQHI hybride (signal, paramétrique, paramétrique) (à gauche) et le modèle P.563 (à droite) appliqué à la base sonore prononcée en allemand

Les huit stimuli surestimés par les modèles DESQHI et P.563, et estimés lors du test subjectif par les valeurs comprises entre $1 < MOS\text{-}LQSN < 1{,}5$, correspondent à la condition de dégradation sans bruit de fond avec une forte discontinuité. La condition de dégradation correspondant au même taux de discontinuité mais avec du bruit de fond obtient une note MOS-LQSN similaire à celle enregistrée lorsqu'il n'y a pas de bruit de fond.

Ce résultat permet de faire l'hypothèse que lorsqu'une dégradation de discontinuité est présente à un niveau élevé, les autres dégradations sont alors masquées par celle-ci. Cette hypothèse devrait être prise en compte afin d'améliorer les performances du modèle DESQHI.

L'erreur de classification automatique du bruit de fond est de 15 % sur l'ensemble de la base sonore.

VI.3. Conclusions

Les performances des modèles existants (G.107 et P.563) et celles du modèle DESQHI sont résumées dans le Tab. VI.12 pour les sept bases sonores testées.

Modèle / Base sonore	Modèle E (G.107)	P.563	DESQHI
BS DESQHI	r = 0,47 / EAM = 0,51	r = 0,51 / EAM = 0,48	r = 0,91 / EAM = 0,22 (signal, signal, hybride)
Supl. 23 exp.1	r = 0,49 / EAM = 0,57	r = 0,72 / EAM = 0,45	r = 0,71 / EAM = 0,45 (signal, signal, para)
Supl. 23 exp.3	r = 0,44 / EAM = 0,64	r = 0,72 / EAM = 0,48	r = 0,72 / EAM = 0,49 (signal, signal, para)
BS bruitée	Ø	r = 0,32 / EAM = 0,62	r = 0,69 / EAM = 0,45 (signal, signal, signal)
NB_LUC_P563	r = 0,22 / EAM = 0,69	r = 0,75 / EAM = 0,45	r = 0,86 / EAM = 0,33 (signal, para, para)
P862_BGN_ENG	r = 0,14 / EAM = 0,86	r = 0,78 / EAM = 0,50	r = 0,79 / EAM = 0,47 (signal, para, para)
P862_BGN_DEU	r = 0,08 / EAM = 0,89	r = 0,83 / EAM = 0,45	r = 0,77 / EAM = 0,51 (signal, para, para)

Tab. VI.12 Comparaison des performances des modèles existants (G.107 et P.563) avec celles du modèle DESQHI pour 7 bases sonores; Les performances sont données par le coefficient de corrélation de Pearson *r* et l'erreur absolue moyenne *EAM*.

Le modèle paramétrique G.107 obtient systématiquement des performances inférieures aux deux autres modèles surtout car il ne prend en compte ni le bruit de fond ni le transcodage, mais seulement le dernier codage utilisé lors de la télécommunication. Les modèles P.563 et DESQHI obtiennent généralement des performances similaires, sauf dans le cas où les bases sonores sont constituées de conditions bruitées. Dans ce cas le modèle DESQHI est plus performant que le modèle P.563. Le modèle DESQHI nécessite moins de ressources de calcul comparativement au modèle P.563. Il est donc mieux approprié pour la mesure en temps réel. Le modèle DESQHI propose en plus de l'évaluation de la qualité vocale un diagnostic de la qualité vocale.

En ce qui concerne la modélisation de la première dimension bruyance, il a été remarqué que le modèle DESQHI est particulièrement efficace, grâce à l'utilisation de la classification automatique du bruit de fond. Les performances de DESQHI dépassent les performances des modèles existants (modèle E et P.563) pour des conditions de dégradations relatives à la présence de bruit de fond, comme le montrent les résultats obtenus sur les bases sonores "P862_BGN" (cf. §.VI.2.4), "NB_LUC_P563" (cf. §.VI.2.3) et base bruitée (cf. §.VI.2.2).

Le diagnostic avancé de la dimension bruyance proposé par le modèle DESQHI permet aussi d'identifier le type de bruit de fond présent sur le signal de la parole. Cette information peut très bien servir à décider automatiquement d'utiliser un algorithme de débruitage, et de décider du degré de puissance de l'algorithme de débruitage. Par exemple, si le bruit est très dérangeant (bruit classifié en grésillement) l'algorithme de débruitage peut être utilisé à son niveau maximal, tandis que s'il s'agit d'un bruit d'environnement ou intelligible, ce bruit n'est pas aussi gênant et peut être conservé dans le signal à transmettre. Une étude pourrait être réalisée dans ce sens.

La dimension du codage de la parole peut être représentée par un indicateur paramétrique ou un indicateur basé sur le signal. Il a été remarqué en général qu'il est préférable d'utiliser l'indicateur basé sur le signal qui prend en compte le transcodage, contrairement à l'indicateur paramétrique (cf. §.VI.2.1.1 et §.VI.2.2). Le temps de calcul CPU est légèrement supérieur pour l'indicateur basé sur le signal que pour l'indicateur paramétrique, néanmoins, il est tout à fait envisageable d'utiliser l'indicateur basé sur le signal lors de mesures en temps réel. L'indicateur paramétrique reste cependant un bon outil, surtout en absence de transcodage (cf. §.VI.2.3 et §.VI.2.4). L'ajout des coefficients *Ie* lors de la présence de transcodage permet d'obtenir des performances de prédiction de la qualité vocale similaires à l'utilisation de l'indicateur basé sur le signal, tout en diminuant le temps de calcul CPU. Dans l'avenir, cette solution paraît être la meilleure, dans le cas où l'information sur le transcodage est disponible dans les statistiques du réseau.
Nous avons aussi observé que l'indicateur basé sur le signal de la dimension du codage de la parole permet de mesurer certaines dégradations relatives à l'utilisation d'un algorithme de débruitage (cf. §.VI.2.2).
Cet indicateur permet de proposer un diagnostic avancé de cette dimension en proposant une identification globale du type de codage ou de transcodage utilisé lors de la télécommunication (cf. §.VI.2.2).

La dimension continuité peut être représentée par des indicateurs paramétriques, hybrides ou basés sur le signal. L'indicateur paramétrique est celui qui est le plus stable lors de l'application aux nouvelles bases sonores. Nous n'avons pas eu la possibilité de tester l'indicateur hybride de la continuité sur les bases sonores inconnues au modèle car l'information sur la localisation des pertes n'est pas disponible sur ces bases. Les hypothèses posées sur l'influence de la localisation des discontinuités lors de l'estimation de la qualité

vocale n'ont pas été vérifiées sur les bases sonores inconnues du modèle. D'autres études pourront être réalisées dans ce sens.
L'indicateur basé sur le signal utilise une technique originale, cependant elle devra être améliorée afin d'être applicable à d'autres bases sonores. Un diagnostic avancé de la télécommunication est proposé afin d'identifier le type de discontinuité parmi 3 classes (signal continu ou présentant des pertes de paquets atténué par un algorithme de PLC, perte de paquets sans utilisation d'un algorithme de PLC, erreurs de bits). Une étude pourrait aussi être réalisée afin d'améliorer le diagnostic avancé de la discontinuité du signal, en déterminant le taux de pertes, la durée des pertes ou encore le type de pertes (rafales ou aléatoires).

Conclusion

Les services de télécommunication sont de plus en plus nombreux et variés (RTC, RNIS, GSM, VoIP). Les opérateurs de téléphonie ont besoin de superviser en temps réel la qualité vocale des services qu'ils proposent et de planifier les nouvelles techniques en cours de développement. La qualité vocale peut être évaluée par des campagnes de tests subjectifs en demandant directement l'avis aux utilisateurs, cependant les méthodes existantes sont très coûteuses et peu adaptées à la supervision des services de télécommunication. Les modèles objectifs sont proposés afin d'évaluer la qualité vocale à moindre coût. Les modèles existants permettent d'estimer la valeur globale de l'évaluation de la qualité vocale, mais cela peut être insuffisant lorsque l'expérimentateur veut connaître l'origine du problème.

Les travaux réalisés dans cette thèse ont permis de construire un instrument de diagnostic et d'évaluation de la qualité vocale appliqué à la téléphonie, pour le contexte d'écoute (modèle DESQHI). Ce modèle est non-intrusif afin de pouvoir proposer un outil applicable en temps réel et en temps différé, et à n'importe quel point du réseau. Il permet de répondre à des besoins de contrôle et de planification de la qualité des services proposés aux utilisateurs.
Une des applications du modèle DESQHI est par exemple son intrusion dans les CPE (passerelles domestiques) sous forme de sonde afin de mesurer les performances de la qualité vocale et de diagnostiquer les éventuels problèmes de transmission.

L'état de l'art présenté dans le Chapitre I a permis de considérer l'évaluation de la qualité vocale comme un phénomène multidimensionnel faisant intervenir certains attributs perceptifs dépendant des dégradations induites par le type de communication utilisé (RTC/RNIS/VoIP/GSM). Le cœur du modèle DESQHI, détaillé dans le Chapitre III, utilise ce principe afin d'évaluer la qualité vocale et de proposer un diagnostic de la télécommunication. Ce cœur est fabriqué à partir d'un test d'évaluation de la qualité vocale (test ACR) et d'un test d'évaluation des dissimilarités par la méthode de comparaison par paires. Les mesures de dissimilarité ont été analysées par la méthode d'échelonnement multidimensionnel afin de déterminer un espace perceptif à trois dimensions identifiées comme la *bruyance*, le *codage de la parole* et la *continuité*. Une combinaison linéaire de ces trois dimensions permet de représenter les notes de qualité vocale issues du test subjectif avec une précision donnée par le coefficient de détermination entre les MOS-LQSN et les notes MOS-LQON ($R^2 = 0{,}86$ cf. §.III.4). Ce résultat confirme que la perception de la qualité vocale est bien un phénomène multidimensionnel faisant intervenir des attributs perceptifs cohérents avec ceux déterminés par Wältermann [14] et Mattila [19]. Cette technique permet de proposer, en plus de la note globale de qualité vocale, trois notes de qualité vocale relatives aux trois dimensions perceptives, ce qui constitue le premier type de diagnostic de la qualité vocale.

La particularité du modèle DESQHI réside dans l'utilisation d'indicateurs hybrides faisant intervenir soit des indicateurs paramétriques, soit des indicateurs basés sur le signal, soit les deux types d'indicateurs simultanément, afin de représenter chacune des trois dimensions perceptives. Cela a l'avantage de proposer un modèle adaptatif comportant différentes versions (une version paramétrique, une basée sur le signal et six hybrides). Il a été vérifié (cf. Tab. VI.3) que les performances du modèle DESQHI sont limitées dans le cas de la version paramétrique, que la version basée sur le signal obtient des performances satisfaisantes et que

l'utilisation de la version hybride améliore considérablement la performance du modèle (respectivement *r = 0,63; r = 0,77; r = 0,91*). L'expérimentateur a ainsi la possibilité de choisir la version du modèle suivant les informations disponibles au point de la mesure ou encore suivant le domaine d'application (contrôle différé ou temps réel, planification).

La modélisation de la bruyance exposée dans le Chapitre IV constitue une grande part du travail réalisé lors de cette thèse et a fait l'objet de deux demandes de dépôt de brevets d'invention ainsi que de deux contributions à l'UIT. Plusieurs tests subjectifs ont relevé une influence du niveau sonore du bruit de fond et une influence du contenu informationnel du bruit de fond lors de l'évaluation de la qualité vocale. Le niveau sonore est représenté par le rapport signal sur bruit. L'influence du contenu informationnel du bruit de fond est prise en compte en déterminant quatre classes de bruits de fond suivant la gêne provoquée lors de l'évaluation de la qualité vocale. Ces quatre classes correspondent respectivement aux *bruits intelligibles*, aux *bruits d'environnement*, aux *bruits de souffle* et aux *bruits de grésillement*. Un algorithme de classification automatique des bruits de fond est proposé à partir d'un arbre de décision utilisant seulement deux indicateurs basés sur le signal. La modélisation de la bruyance permet de représenter la première dimension de l'espace perceptif avec une excellente précision (*r = 0,97* cf. Fig. IV.18). L'algorithme de classification des bruits de fond constitue le diagnostic avancé de la dimension bruyance en permettant d'identifier le type de bruit présent sur le signal de parole. Par exemple, cela permet de prévoir si le bruit de fond est causé par un problème de transmission de la parole ou bien s'il provient de l'environnement du locuteur. Cet algorithme permet aussi de prendre la décision d'appliquer un débruitage ou non suivant le type de bruit de fond.
A ce jour, la dimension bruyance ne peut être représentée qu'à partir d'indicateurs basés sur le signal car les statistiques du réseau ne permettent pas de donner une quelconque information sur la présence du bruit de fond.

La dimension codage de la parole (cf. §.V.1) est représentée soit par un indicateur paramétrique, soit par un indicateur basé sur le signal. L'indicateur paramétrique utilise le facteur de dégradation *Ie* défini par l'UIT-T G.113 [55], utilisé également par le modèle E (G.107). L'indicateur basé sur le signal est issu d'une combinaison entre un indicateur résiduel issu des coefficients LPC et la détermination du temps d'attaque du signal. Cet indicateur permet de représenter le codage de la parole avec une précision de *r = 0,86* (cf. §.V.1.2). Il permet aussi d'orienter l'expérimentateur sur le type de codage utilisé grâce à une classification en six groupes, pour un diagnostic avancé du codage de la parole.

La dimension continuité (cf. §.V.2) peut être représentée soit par un indicateur paramétrique, soit par un indicateur hybride, soit par un indicateur basé sur le signal. L'indicateur paramétrique est appelé le pourcentage global de discontinuité (*pgd*). Il a été construit comme une pondération entre le pourcentage global de pertes de paquets et le pourcentage global d'erreurs de bits. L'indicateur hybride utilise ce même indicateur *pgd*, mais en comptabilisant les discontinuités uniquement sur les zones actives du signal à l'aide d'un algorithme de détection d'activité vocale. L'indicateur basé sur le signal utilise l'analyse du signal soumis à un filtre passe-bas à la fréquence de coupure de 80Hz afin de mieux déceler les discontinuités dans cette zone dépourvue de signal de parole. Un diagnostic avancé de la dimension continuité est proposé pour identifier le type de discontinuité à savoir, si le stimulus est continu (pas de discontinuité ou alors pertes de paquets atténuées par un algorithme de PLC), si le stimulus présente des coupures franches du signal (pertes de paquets sans utilisation d'algorithme de PLC) ou bien s'il s'agit d'erreurs de bits.

DESQHI propose en plus de la note globale d'évaluation de la qualité vocale, deux types de diagnostics de la télécommunication qui sont très utiles pour orienter

l'expérimentateur lors du contrôle des télécommunications. Le premier diagnostic utilise la structure multidimensionnelle du modèle afin de proposer trois notes relatives à chacune des trois dimensions (RMOS). Le deuxième appelé diagnostic avancé consiste à identifier les causes techniques de la dégradation de la qualité vocale.

Le Chapitre VI présente les performances du modèle DESQHI à partir de sept bases sonores inconnues du modèle. Ces performances sont comparées à celles obtenues par les modèles existants (P.563 et modèle E). Il a été vérifié que le modèle DESQHI est au moins aussi performant que ces deux modèles, en utilisant un temps CPU inférieur à celui du modèle P.563 (cf. Tab. VI.12). De plus, le modèle DESQHI propose deux types de diagnostic de la qualité vocale.

Les performances du modèle DESQHI permettent de faire certaines recommandations lors du choix de la version du modèle (basé sur le signal, paramétrique ou hybride). Lorsque les conditions de dégradation sont relatives à des dégradations classiques d'une communication IP et/ou mobile et/ou RTC, il est préférable d'opter pour le modèle hybride utilisant les indicateurs signal pour la bruyance, signal pour le codage de la parole et paramétrique pour la continuité. Dans le cas où l'information du transcodage est disponible dans les statistiques du réseau nous recommandons d'utiliser l'indicateur paramétrique pour représenter le codage de la parole (DESQHI hybride respectivement signal, paramétrique, paramétrique).

Lorsque les stimuli comprennent des dégradations liées à l'utilisation d'algorithme de débruitage, il est recommandé d'opter pour la version entièrement basée sur le signal afin de considérer le maximum de dégradations (cf. §.VI.2.2). De manière générale, lorsque les informations sur les dégradations sont indisponibles dans les statistiques du réseau, il est préférable d'utiliser les indicateurs basés sur le signal.

Nous recommandons d'appliquer dans tous les cas la première dimension relative à la bruyance, même si les conditions de dégradations ne sont pas bruitées car certains codecs génèrent du bruit de fond.

Perspectives

Le modèle DESQHI est limité par les dégradations physiques présentes dans la base sonore utilisée lors de la construction du modèle. Certaines dégradations non utilisées lors de la construction du cœur du modèle sont parfois assimilées à l'une des trois dimensions perceptives, comme dans le cas de l'utilisation des algorithmes de débruitage par exemple. Néanmoins, le modèle pourra être amélioré en considérant de nouvelles dégradations comme le niveau sonore du signal vocal, le time-warping, l'utilisation d'un kit main libre.

Les nouvelles techniques de communication sont axées sur l'élargissement de la largeur de la bande passante (bande élargie et super élargie correspondant respectivement à 50 Hz – 8 kHz et 50 Hz – 16 kHz). Les outils d'évaluation de la qualité vocale (et le modèle DESQHI) doivent être adaptés à ce nouveau type de télécommunication. Les études réalisées dans le cadre de la thèse de Côté [10] ont montré que la prédiction de la qualité vocale de conditions de dégradation en bande étroite et en bande élargie est possible à partir d'un espace perceptif à quatre dimensions. Trois de ces dimensions sont issues des études de Wältermann [14] et concordent avec notre espace perceptif. La quatrième dimension correspond à la sonie de la parole, cependant elle n'est pas orthogonale aux trois autres dimensions. Les dimensions perceptives correspondant à la bruyance, la continuité et au codage de la parole semblent couvrir la plupart des dégradations perceptives causées par les techniques actuelles de télécommunication. La combinaison reliant les différentes dimensions perceptives à la prédiction de la qualité vocale devra aussi être adaptée à de telles conditions. Il faudra définir l'échelle de l'évaluation de la qualité vocale et trouver un moyen de compatibilité entre l'échelle des notes obtenues pour des conditions en bande étroite et l'échelle correspondante à la bande élargie, ou encore une échelle qui serait invariante de la largeur de bande utilisée.

Les modélisations de chacune des trois dimensions perceptives du modèle DESQHI peuvent faire l'objet de nouvelles analyses afin d'améliorer les performances globales de la prédiction de la qualité vocale :

- A ce jour, la dimension bruyance ne peut être représentée qu'à partir d'indicateurs basés sur le signal car les statistiques du réseau ne permettent pas de donner une quelconque information sur la présence du bruit de fond. L'identification du codage utilisé lors de la télécommunication peut être utile afin de prévoir le niveau et le type du bruit généré par celui-ci, cependant cela ne couvre qu'une part minime de la présence de bruit de fond sur le signal vocal. Il est supposé que le diagnostic avancé d'identification du type de bruit de fond permettrait de décider d'utiliser un algorithme de débruitage.

- La modélisation de la dimension codage de la parole par les indicateurs paramétriques est limitée lorsque l'information concernant le transcodage utilisé n'est pas disponible dans les statistiques du réseau. Si cette information vient à être disponible, il serait alors avantageux d'additionner les facteurs de dégradation correspondant aux codages successifs. Une étude devra aussi être proposée afin de déterminer les facteurs de dégradation *Ie* pour les nouveaux codecs développés. Dans le cas de l'indicateur basé sur

le signal, le diagnostic avancé permettant actuellement d'identifier un type de codage ou de transcodage devrait être affiné afin de connaître précisément le codage ou le transcodage utilisé lors de la télécommunication.

- La modélisation de la dimension continuité utilisant l'indicateur basé sur le signal nécessite une amélioration. Une approche originale est proposée dans ce manuscrit, cependant les performances obtenues sont encore limitées. Le principal problème réside entre les différences remarquées entre des dégradations provoquées par les erreurs de bits et les pertes de paquets. Il pourrait être intéressant d'utiliser le diagnostic avancé de la continuité pour distinguer les dégradations liées aux pertes de paquets et aux erreurs de bits afin de proposer, pour ces deux types de dégradations physiques, différents indicateurs. Le diagnostic avancé pourrait aussi être amélioré en déterminant par exemple le type de pertes (aléatoires ou par rafales), le pourcentage de pertes, la durée des discontinuités. La différence de perception des discontinuités suivant les deux locuteurs n'a pas été clairement justifiée. L'hypothèse la plus probable concerne la localisation des pertes sur le signal de la parole, cependant, les indicateurs hybrides développés pour répondre à cette hypothèse ne sont pas concluants.

Bibliographie

[1] ITU-T Rec. P.862 : Perceptual Evaluation of Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs, Geneva, (2001)

[2] UIT-T Rec. P.563 : Méthode mono-extrémité pour l'évaluation objective de la qualité vocale dans les applications de la téléphonie à bande étroite, Genève, (2004)

[3] UIT-T Rec. G.107 : Le modèle E, Modèle de calcul utilisé pour la planification de la transmission, Genève, (2003)

[4] Garnier M., *Communiquer en environnement bruyant : de l'adaptation jusqu'au forçage vocal*, Université de Paris 6, 222 p., (2007)

[5] Möller S., *Assessment and Prediction of Speech Quality in Telecommunications*, Kluwer Academic Publisher, 244 p., ed. first, (2000)

[6] Greenberg S., *Temporal Properties of Spoken Language*, International Congres on Acoustics, Kyoto, Japan, (2004)

[7] Jekosch U., *Voice and Speech Quality Perception: Assessment and Evaluation Signals and Communication Technology*, Springer, Berlin, (2005)

[8] Raake A., *Speech Quality of VoIP : Assessment and Prediction*, Wiley, 336 p., ed. 1, (2006)

[9] French et Steinberg, *Factors Governing the Intelligibility of Speech Sounds*, Journal of the Acoustical Society of America, vol. 19, (1), p. 90-119, (1947)

[10] Côté N., *Integral and Diagnostic Intrusive Prediction of Speech Quality*, Université de Berlin, 238 p., (2010)

[11] UIT-T Rec. P.800 : Méthodes d'évaluation subjective de la qualité de transmission, Genève, (1996)

[12] UIT-R Rec. BS.1534 : Méthode d'évaluation subjective du niveau de qualité intermédiaire des systèmes de codage, Genève, (2003)

[13] Gabrielsson A. et Sjogren H., *Perceived sound quality of sound-reproducing systems*, Journal of the Acoustical Society of America, vol. 65, (4), p. 1019-1033, (1979)

[14] Wältermann M., Scholz K., Möller S., Huo L., Raake A. et Heute U., *An Instrumental Measure for End-to-end Speech Transmission Quality Based on perceptual Dimensions : Framework and Realization*, Interspeech 2008, Brisbane, Australia, (2008)

[15] Jekosch U., *Meaning in the Context of Sound Quality Assessment*, Acustica, vol. 85, p. 681-684, (1999)

[16] Petersen K.T., Hansen S.D. et Sorensen J.A., *Speech quality assessment of compounded digital telecommunication systems; perceptual dimensions.*, IEEE, p. 1375-1378, (1997)

[17] Hall J.L., *Application of multidimensional scaling to subjective evaluation of coded speech*, Journal of the Acoustical Society of America, vol. 110, (4), p. 2167-2182, (2001)

[18] Wältermann M., Raake A. et Möller S., *Underlying Quality Dimensions of Modern Telephone Connections*, Interspeech, Pittsburgh, Pennsylvania, (2006)

[19] Mattila V.-V., *Ideal point modelling of speech quality in mobile communications based on multidimensional scaling (MDS)*, Journal of the Audio Engineering Society, vol. 112, p. 1-14, (2002)

[20] Zimmer K., Ellermeier W. et Schmid C., *Using Probabilistic Choice Models to Investigate Auditory Unpleasantness*, Acta Acustica united with Acustica, vol. 90, p. 1019-1028, (2004)

[21] McGee V.E., *Determining perceptual spaces for the quality of filtered speech*, Journal of Speech and Hearing Research, vol. 8, p. 23-38, (1965)

[22] Bappert V. et Blauert J., *Auditory quality evaluation of speech-coding systems*, acta acustica, vol. 2, p. 49-58, (1994)

[23] Osgood C.E., Suci G.J. et Tannenbaum P.H., *The measurement of meaning*, Universtiy of Illinois Press., 342 p., (1957)

[24] Voiers W.D., *Diagnostic Acceptability Measure for Speech Communication Systems*, ICASSP, p. 204-207, Hartford, Connecticut, USA, (1977)

[25] Borg I. et Groenen P.J.F., *Modern multidimensional scaling, Theory and Applications, Second Edition*, Springer Sciences Business Media, 614 p., (2005)

[26] Susini P., Adams M. et Winsberg S., *A Multidimensional Technique for Sound Quality Assessment*, Acustica, vol. 85, p. 650-656, (1999)

[27] McDermott B.J., *Multidimensional Analyses of Circuit Quality Judgments*, Journal of the Acoustical Society of America, vol. 45, (3), p. 774-781, (1969)

[28] Lavandier M., *Différences entre enceintes acoustiques: Une évaluation physique et perceptive*, Université de Aix-Marseille, 199 p., (2005)

[29] Bernex E. et Barriac V., *Architecture of non-intrusive perceived voice quality assessment*, Measurement of Speech and Audio Quality in Networks, Prague, (2002)

[30] Escoufier Y., *Le positionnement multidimensionnel*, Revue de statistique appliquée, vol. 23, (4), p. 5-14, (1975)

[31] Torgeson W.S., *Theory and Methods of Scaling*, Wiley, 460 p., Oxford, England, (1958)

[32] Kruskal J.B., *Multidimensional Scaling by optimizing goodness of fit to a nonmetric hypothesis*, Psychometrika, vol. 29, (1), p. 1-27, (1964)

[33] Carroll J.D. et Chang J.-J., *Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition*, Psychometrika, vol. 35, (3), p. 283-319, (1970)

[34] Etame T.E., *Conception de signaux de référence pour l'évaluation de la qualité perçue des codeurs de la parole et du son*, Université de Rennes, (2008)

[35] Gabrielsson A., Rosenberg U. et Sjogren H., *Judgments and dimension analyses of perceived sound quality of sound-reproducing systems*, Journal of the Acoustical Society of America, vol. 55, (4), p. 854-861, (1974)

[36] Scholz K., Kuhnel C., Wältermann M., Möller S. et Heute U., *Assessment of the Speech-Quality Dimension "Noisiness" for the Instrumental Estimation and Analysis of Telephone-Band Speech Quality*, ICSLP, Brisbane, (2008)

[37] Zwicker E. et Zwicker U.T., *Binaural masking-level differences in non-simultaneous masking*, ScienceDirect, vol. 13, (3), p. 221-228, (1984)

[38] Zwicker E. et Fastl H., *Psychoacoustics : Facts and Models*, Springer, (1999)

[39] Festen J. et Plomp R., *Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing*, Journal of the Acoustical Society of America, vol. 88, p. 1725-1736, (1990)

[40] Bronkhorst A.W. et Plomp R., *Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing*, Journal of the Acoustical Society of America, vol. 92, (6), p. 3132-3139, (1992)

[41] Brungart D.S., Chang P.S., Simpson B.D. et Wang D., *Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation*, Journal of the Acoustical Society of America, vol. 120, (6), p. 4007-4018, (2006)

[42] Grataloup C., Hoen M., Pellegrino F. et Meunier F., *Influence des paramètres psycholinguistiques du cocktail party sur la compréhension d'un signal de parole cible*, Actes des XXVI journée d'études sur la parole, Dinard, (2006)

[43] Hoen M., Meunier F., Grataloup C.-L., Pellegrino F., Grimault N., Perrin F., Perrot X. et Collet L., *Phonetic and lexical interferences in informational masking during speech-in-speech comprehension.*, Speech communication, vol. 49, p. 905-916, (2007)

[44] Durlach N., Mason C.R., Kidd G., Arbogast T.L., Colburn H.S. et Shinn-Cunningham B., *Note on informational masking (L)*, Journal of the Acoustical Society of America, vol. 113, (6), p. 2984-2987, (2003)

[45] Rhebergen K.S., Versfeld N.J. et Dreschler W.A., *Release from informational masking by time reversal of native and non-native interfering speech (L)*, Journal of the Acoustical Society of America, vol. 113, (3), p. 1274-1277, (2005)

[46] Ellermeier W., Zeitler A. et Fastl H., *Impact of Source Identifiability on Perceived Loudness*, The 18th International Congress on Acoustics, Kyoto, Japan, (2004)

[47] Fastl H., *Neutralizing the Meaning of Sound for Sound Quality Evaluations*, 17th ICA Proceedings, Rome, (2001)

[48] Ellermeier W., Zeitler A. et Fastl H., *Predicting annoyance judgments from psychoacoustic metrics: Identifiable versus neutralized sounds*, The 33rd International Congress and Exposition on Noise Control Engineering, Prague, (2004)

[49] Moore et Glasberg, *A model for the prediction of the thresholds, loudness and partial loudness*, Journal of the Audio Engineering Society, vol. 45, (4), p. 224-240, (1997)

[50] Falk T.H. et Chan W.-Y., *Single-Ended Speech Quality Measurement Using Machine Learning Methods*, IEEE, vol. 14, (6), p. 1935-1947, (2006)

[51] Mattila V.-V., *Descriptive analysis and ideal point modelling of speech quality in mobile communication*, Journal of the Audio Engineering Society, vol. 113, p. 1-18, (2002)

[52] UIT-T Rec. G.711 App.1 : Algorithme simple de haute qualité pour le masquage des pertes de paquets en codage G.711, Genève, (1999)

[53] Voran S.D., *Perception of temporal Discontinuity Impairments in Coded Speech - A Proposal for Objective Estimators and Some Subjective Test Results*, Measurement of Speech and Audio Quality n Networks, Prague, République Tchèque, (2003)

[54] Huo L., Wältermann M., Heute U. et Möller S., *Estimation of the Speech Quality Dimension "Discontinuity"*, ITG-Conference on Speech Communication, Aachen, Germany, (2008)

[55] ITU-T Rec. G.113 : Transmission impairments due to speech processing, Geneva, (2007)

[56] Ding L., Lin Z., Radwan A. et El-Hennawey M.S., *Non-intrusive single-ended speech quality assessment in VoIP*, Science Direct, vol. 49, p. 477-489, (2007)

[57] Rabiner L.R. et Sambur M.R., *Application of an LPC distance measure to the voiced-unvoiced-silence detection problem*, IEEE, vol. 25, (4), p. 338-343, (1977)

[58] Kim D.-S. et Tarraf A., *ANIQUE+ : A New American National Standard for Non-Intrusive Estimation of Narrowband Speech Quality*, Bell Labs Technical Journal, vol. 12, (1), p. 221-236, (2007)

[59] Krenek J. et Holub J., *histogram based approach for non-intrusive speech quality measurement in networks*, Audio Engineering Society, République tchèque, (2005)

[60] ITU-T Rec. G.729 : Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP), Geneva, (2007)

[61] d'Alessandro C. et Demars C., *Représentations temps-fréquence du signal de parole*, Traitement du signal, vol. 9, (2), p. 153-173, (1992)

[62] Didiot E., *Segmentation parole/musique pour la transcription automatique de la parole continue*, Université de Henri Poincaré, 131 p., (2007)

[63] Istrate D.M., *Détection et reconnaissance des sons pour la surveillance médicale*, Université de Grenoble, 186 p., (2003)

[64] Kryter K.D., *Methods for the Calculation and Use of the Articulation Index*, JASA, vol. 34, (11), p. 1689-1697, (1962)

[65] ITU-T Handbook on Telephonometry, Geneva, (1992)

[66] Molla S., Boullet I., Meunier S., Rabau G., Gauduin B. et Boussard P., *Calcul des indicateurs de sonie : revue des algorithmes et implémentation*, 10éme Congrès Français d'Acoustique, Lyon, France, (2010)

[67] Scheirer E. et Slaney M., *Construction and evaluation of a Robust Multifeature Speech/Music Discriminator*, ICASSP-97, Munich, (1997)

[68] Huo L., Wältermann M., Heute U. et Möller S., *Estimation Model for Speech-Quality Dimension "Noisiness"*, Acoustics08, Paris, (2008)

[69] Guéguin M., *On the Evaluation of the Conversational Speech Quality in Telecommunications*, Journal on Advances in Signal Processing, vol. 2008, p. 15, (2008)

[70] Appel R. et Beerends J.G., *On the quality of hearing one's own voice*, Journal of the Audio Engineering Society, vol. 50, (4), p. 237-248, (2002)

[71] Guéguin M., *Evaluation objective de la qualité vocale en contexte de conversation*, Université de Rennes1, 178 p., (2006)

[72] UIT-T Rec. P.562 : Analyse et interprétation des mesures en service sans intrusion dans les services vocaux, Genève, (2004)

[73] UIT-T Rec. P.561 : Dispositif de mesure en service et sans intrusion - Mesures pour les services vocaux, Genève, (2002)

[74] ITU-T Rec. P.564 : Conformance testing for voice over IP transmission quality assessment models, Geneva, (2007)

[75] Malfait L., Berger J. et Kastner M., *P.563-The ITU-T Standard for Single-Ended Speech Quality Assessment*, IEEE Transaction on Audio, Speech, and Language Processing, vol. 14(6), p. 1924-1934, (2006)

[76] Berger J. UIT-T Contribution COM. 12-34: TOSQA Telecommunication Objective Speech-Quality Assessment, Geneva, (1997)

[77] Kim D.-S., *ANIQUE: An Auditory Model for Single-Ended Speech Quality Estimation*, IEEE, vol. 13, (5), p. 821-831, (2005)

[78] Grancharov V., Zhao D.Y., Lindblom J. et Kleijn W.B., *Low-Complexity, Nonintrusive Speech Quality Assessment*, IEEE Transaction on Audio, Speech, and Language Processing, vol. 14, (6), p. 1948-1956, (2006)

[79] Raja A. et Flanagan C., *Real-time, Non-intrusive Speech Quality Estimation: A Signal-Based Model*, EuroGP 2008, p. 1627-1634, Napoli, Italie, (2008)

[80] UIT-T Rec. P.56 : Mesure objective du niveau vocal actif, Genève, (1993)

[81] IUT-T Rec. G.711 : Pulse Code Modulation (PCM) of voice frequencies, Geneva, (1972)

[82] IUT-T Rec. G.726 : 40, 32, 24, 16 kbit/s Adaptative Differential Pulse Code Modulation (ADPCM), Geneva, (1990)

[83] Gilbert E.N., *Capacity of a burst noise channel*, Bell System Technical, vol. 39, p. 1253-1266, (1960)

[84] UIT-T Rec. P.48 : Spécification d'un système de référence intermédiaire, Genève, (1989)

[85] Boullet I., *La sonie des sons impulsionnels: Perception, Mesures et Modèles*, Université de Aix-Marseille, (2005)

[86] Hellbrück J., Fastl H. et Keller B., *Effects of meaning of sound on loudness judgements*, Forum Acusticum, Sevilla, (2002)

[87] Ma L., Smith D.J. et Milner B.P., *Context awareness using environmental noise classification*, Eurospeech, p. 2237-2240, Geneva, Switzerland, (2003)

[88] El-Maleh K., Samouelian A. et Kabal P., *Frame-level noise classification in mobile environments*, IEEE Conference Acoustics, speech, Signal Proc, p. 237-240, (1999)

[89] ITU-T Rec. P.341 : Transmission characteristics for wideband (150-7000 Hz) digital hands-free telephony terminals, Geneva, (1998)

[90] Istrate D., Vacher M. et Serignat J.f., *Détection et classification des sons : application aux sons de la vie courante et à la parole*, GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, Louvain-la-Neuve, Belgique, (2005)

[91] Breiman L., Frieman J., Olshen R. et Stone C., *Classification and regression trees*, Chapman and Hall, 358 p., (1993)

[92] Gautier-Turbin V. et Gros L., *On the perceived quality of noise reduced signals*, Interspeech 2008, p. 2062-2065, Brisbane, Australia, (2008)

[93] ITU-T Rec. P.835 : Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithm, Geneva, (2003)

[94] Nagle A., Quinquis C., Sollaud A., Battistello A. et Slock D., *Quality impact of diotic versus monaural hearing on processed speech*, Audio Engineering Society, New York, USA, (2007)

[95] ITU-T Rec. G.107 : The E-model, a computational model for use in transmission planning, Geneva, (2003)

[96] IUT-T Serie P, Supplement 23 : Telephone transmission quality, Telephone Installations, local line networks, Geneva, (1998)

[97] Draper N. et Smith H., *Applied Regression Analysis, Second Edition*, John Wiley and Sons, p. 307-312, (1981)

[98] IUT-T Contribution Com12 : Test Plan for single-Ended Assessment Models, Geneva, (2001)

[99] IUT-T SG12 Q.13 : Test plan for Background Noise Conditions Evaluation of Objective Speech Quality Measure, Geneva, (1997)

[100] Zwicker E. et Scharf B., *A model of loudness summation*, Psychological Review, vol. 72, p. 3-26, (1965)

Annexes

Annexe A. Méthode d'échelonnement multidimensionnel (EMD)

Les méthodes généralement utilisées lors de l'analyse des données de dissimilarités appliquées aux sciences humaines sont les méthodes d'échelonnement multidimensionnel (EMD) encore appelées MultiDimensional Scaling (MDS). Ces méthodes consistent à modéliser les dissimilarités par des distances (euclidiennes la plupart du temps) qui peuvent être représentées dans un espace perceptif constitué d'un nombre minimal de dimensions, avec une erreur d'approximation minimale des dissimilarités par les distances (Borg [25] et Escoufier [30]). Les méthodes de Torgeson, de Kruskal puis de Caroll et Chang sont présentées ci-dessous.

A. Calcul des distances euclidiennes

La distance euclidienne entre les deux stimuli *i* et *j* appelée d_{ij} peut être représentée dans un espace à *N* dimensions par la relation suivante :

$$d_{i,j} = [\sum_{k=1}^{N} (x_{ik} - x_{jk})^2]^{\frac{1}{2}}, \qquad \textbf{Équation 1}$$

où $d_{i,j}$ est la distance euclidienne entre les conditions *i* et *j* situées dans l'espace perceptif à *N* dimensions. Les distances euclidiennes calculées respectent les trois propriétés suivantes :

Propriété de symétrie :

La distance euclidienne entre les sons A et B doit être identique à celle qu'on trouve entre les sons B et A.

La matrice des distances obtenue est donc symétrique.

$$d_{i,j} = d_{j,i} \qquad \textbf{Équation 2}$$

Propriété d'identifiabilité :

Le jugement de dissimilarité entre deux sons identiques doit être nul dans le cas du test de dissimilarité, ou de 1 dans le cas du test de similarité.

$$d_{i,i} = 0 \qquad \textbf{Équation 3}$$

L'inégalité triangulaire :

Soit trois sons différents A, B et C. Les distances perçues entre ces sons sont $d_{a\text{-}b}$ / $d_{a\text{-}c}$ / $d_{b\text{-}c}$.

Dans ce cas, l'inégalité triangulaire doit vérifier ces 3 inégalités :

$$\begin{cases} d_{a-b} \le d_{a-c} + d_{b-c} \\ d_{a-c} \le d_{b-c} + d_{a-b} \\ d_{b-c} \le d_{a-c} + d_{a-b} \end{cases} \qquad \textbf{Équation 4}$$

Ces propriétés peuvent être utilisées à partir des jugements de dissimilarités afin de vérifier la cohérence des réponses de certains sujets.

Par exemple, dans le cas de la propriété de symétrie, des paires témoins de stimuli peuvent être ajoutées au test subjectif en présentant les deux ordres (A-B) et (B-A). Un sujet normal devrait juger les différences entre les deux paires de stimuli de la même manière.
Dans le cas de la propriété d'identifiabilité, on présente une paire composée de deux stimuli identiques ; le sujet devrait spécifier qu'il n'y a pas de différence entre les stimuli.
La propriété de l'inégalité triangulaire peut être évaluée sur l'ensemble des paires testées.

Ces trois propriétés peuvent être testées dans le but de rejeter certains sujets qui auraient des jugements complètement aberrants ou encore qui n'auraient pas compris la consigne.

B. Méthode EMD métrique de Torgeson

La première des méthodes EMD a été développée par Torgeson [31] en 1958. Elle est nommée la méthode EMD métrique ou Classical MDS (CMDS). Cette méthode utilise des données métriques de dissimilarités obtenues par le jugement moyen des sujets à l'aide d'une échelle métrique continue. La modélisation de ces dissimilarités est estimée par les distances euclidiennes introduites dans le paragraphe précédent, afin de représenter les données dans un espace euclidien à N dimensions. Plus le nombre de dimensions de l'espace euclidien est élevé, plus l'erreur de modélisation des dissimilarités sera faible, cependant certaines dimensions ne participent que très peu à la diminution de l'erreur commise et peuvent être exclues de l'espace perceptif final. La méthode de Torgeson utilise une analyse en composantes principales afin de réduire au maximum le nombre de dimensions en considérant une limite acceptable de l'erreur commise de modélisation des dissimilarités.

$$E = \sqrt{\sum_{i,j} (\delta_{ij} - d_{ij})^2} \,, \qquad \textbf{Équation 5}$$

avec d_{ij} les distances euclidiennes et δ_{ij} les mesures de dissimilarités. E représente alors l'erreur commise. La grandeur utilisée pour exprimer l'erreur de la modélisation des dissimilarités est le stress dit métrique, déterminant la différence entre les distances euclidiennes et les dissimilarités estimées. Le stress métrique peut être défini par la formule suivante :

$$stress\_Torgeson = \sqrt{\frac{\sum (\delta_{ij} - d_{ij})^2}{\sum d_{ij}^2}} \,, \qquad \textbf{Équation 6}$$

avec $d_{i,j}$ les distances euclidiennes et $\delta_{i,j}$ les mesures de dissimilarités.

La représentation de l'erreur commise lors de la modélisation des dissimilarités peut être représentée par deux graphiques appelés "Scree plot" et "diagramme de Shepard".

Le Scree plot représente la valeur du stress en fonction du nombre de dimensions constituant l'espace perceptif. Ce graphique est souvent utilisé pour déterminer le nombre de dimensions optimal à la représentation de l'espace. Dans le meilleur des cas, la courbe présente un coude au niveau du nombre de dimensions optimal à la représentation de l'espace.

La performance de la modélisation des dissimilarités peut aussi être visualisée par "le diagramme de Shepard" qui représente la corrélation entre les dissimilarités estimées par le test subjectif et les distances euclidiennes de l'espace perceptif considéré.

Ces deux représentations graphiques servent à choisir le meilleur compromis entre le nombre de dimensions et l'erreur commise lors de la modélisation des dissimilarités. En général, le nombre de dimensions est défini lorsque l'ajout d'une dimension supplémentaire n'apporte plus d'information significative à la modélisation des dissimilarités.

C. Méthode EMD non-métrique de Kruskal (NMDS)

Kruskal [32] a développé une méthode EMD utilisant des données non-métriques de dissimilarités en 1964. La modélisation des dissimilarités est ordinale en représentant l'ordre

de classement des dissimilarités pour les différents stimuli. Cette méthode est utilisée entre autres par Mc Dermott [27].

Contrairement à Torgeson, Kruskal cherche directement une représentation de l'espace euclidien pour un nombre défini de dimensions qui vérifie la propriété de l'ordre des distances euclidiennes par rapport à l'ordre des dissimilarités. L'algorithme EMD non-métrique est constitué de 2 phases.

Tout d'abord les distances euclidiennes d_{ij} doivent être traduites au mieux par une fonction croissante monotone et positive appelée *f*, afin de modéliser le plus précisément possible les dissimilarités δ_{ij} par la relation suivante :

$$f(\delta ij) = dij + E \qquad \textbf{Équation 7}$$

Dans ce cas, le but est de minimiser l'erreur commise représentée par la relation suivante :

$$E = \sqrt{\sum_{i,j} (f(\delta_{ij}) - d_{ij})^2} \qquad \textbf{Équation 8}$$

La deuxième phase consiste alors à déplacer les points de coordonnées d_{ij} par itération de manière à se rapprocher le plus possible des dissimilarités δ_{ij}. Lorsque la valeur de l'erreur (appelée stress) devient inférieure à une limite imposée par l'expérimentateur, l'espace perceptif est retenu.

L'erreur de représentation de la matrice des dissimilarités est donnée par le stress de la NMDS:

$$stress\_kruskal = \sqrt{\frac{\sum (f(\delta_{ij}) - d_{ij})^2}{\sum d_{ij}^2}}, \qquad \textbf{Équation 9}$$

Avec $d_{i,j}$ les distances euclidiennes et $\delta_{i,j}$ les mesures de dissimilarités.

Cette valeur d'erreur peut être représentée graphiquement de la même manière que le stress métrique grâce au graphique "scree plot" ainsi que par "le diagramme de Shepard".

D. Méthode EMD pondérée de Caroll et Chang (1970) [33]

Cette méthode, aussi appelée "Individual Difference Scaling" (INDSCAL) ou "weighted MDS" (WMDS), est une extension de la méthode de Kruskal en prenant en compte les différences interindividuelles des sujets. Elle est utilisée par Bappert et Blauert [22], Mattila [19], Wältermann *et al.* [14], Etame [34] ainsi que Gabrielsson et Sjogren [13]. Les données d'entrée utilisées par cette méthode peuvent être des matrices de dissimilarités métriques ou non-métriques, déterminées pour chaque sujet.

La configuration d'origine est tout d'abord déterminée par la méthode de Torgeson ou de Kruskal sur l'ensemble des résultats de dissimilarités moyennées sur les sujets.

La méthode INDSCAL considère que tous les sujets ont jugé les dissimilarités en considérant les mêmes dimensions perceptives, mais en leur attribuant des pondérations différentes selon les sujets. INDSCAL évalue donc les pondérations suivant chaque dimension pour chacun des sujets.

On suppose que l'on dispose d'une matrice de dissimilarités symétrique $\delta_{ij,k}$ désignant les dissimilarités entre les conditions *i* et *j* pour le sujet *k*.

A partir de ces matrices de dissimilarités, une matrice w_{kl} comportant le poids de la dimension *l* pour le sujet *k* est déterminée simultanément avec les coordonnées des *n* conditions dans un espace de dimension *p*, de manière à obtenir une matrice de distance $d_{ij,k}$ se rapprochant le plus de la matrice des dissimilarités $\delta_{ij,k.}$

Le calcul des distances euclidiennes est dans ce cas donné par la formule suivante :

$$d_{ij,k} = [\sum_{l=1}^{p} w_{kl}(x_{il} - x_{jl})^2]^{\frac{1}{2}}$$ **Équation 10**

$f(\delta_{ij,k}) = d_{ij,k} + E_k$, en non-métrique, et

$\delta_{ij,k} = d_{ij,k} + E_k$ en métrique,

avec E_k l'erreur de la reconstruction des dissimilarités.

La méthode INDSCAL utilise des données de dissimilarités métriques, mais elle peut traiter ces dissimilarités soit comme des données métriques, soit comme des données ordinales.

L'erreur commise est exprimée par le S-Stress :

$$S - Stress = \sqrt{\frac{1}{K}\sum_{k}(\frac{\sum(\delta_{ij,k}^2 - d_{ij,k}^2)^2}{\sum d_{ij,k}^4})}\,,$$ **Équation 11**

avec $d_{ij,k}$ les distances euclidiennes et $\delta_{ij,k}$ les mesures de dissimilarités entre les conditions i et j pour le sujet k.

Les méthodes d'échelonnement multidimensionnel présentées dans cette partie sont utilisées lorsque l'expérimentateur n'a aucune idée du nombre de dimensions à prendre en compte pour représenter l'espace perceptif. Dans ce cas le nombre de dimensions à considérer est choisi en analysant l'erreur commise donnée par la valeur du stress. En pratique le nombre de dimensions optimum est déterminé lorsque l'ajout d'une dimension n'ajoute que peu d'information supplémentaire à la reconstruction de la matrice distance.

Annexe B. Consigne du test d'évaluation de la qualité vocale

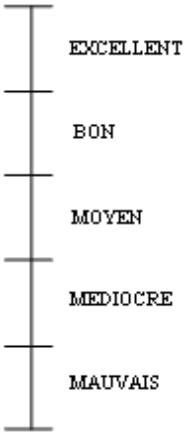
Bonjour,

Vous allez entendre à travers le casque qui est placé devant vous des échantillons de parole d'une durée de 4,5 secondes comme si vous étiez en communication avec une personne distante.

Chaque échantillon est constitué de deux phrases courtes séparées par un silence.

Pendant l'écoute, le bouton rouge qui est devant vous sera allumé. **Vous voudrez bien écouter chaque échantillon de parole complètement.**

Puis, quand le bouton vert s'allumera, **vous donnerez votre opinion sur la qualité de l'échantillon de parole que vous venez d'entendre en plaçant le curseur sur l'échelle continue suivante :**



Vous pouvez placer le curseur où vous voulez sur l'échelle. Les labels (Excellent à Mauvais) servent de repères pour vous.
Vous disposez de 5 secondes pour enregistrer votre réponse (temps pendant lequel le bouton vert restera allumé). Après les 5 secondes, se produira une courte pause avant l'échantillon suivant.

Dans cette expérience nous commencerons par un apprentissage formé de quelques échantillons. Viendra ensuite la séance d'une durée d'environ 15 minutes.

Merci pour votre participation et bon courage.

Annexe C. Consigne du test d'évaluation des dissimilarités

Test de dissimilarité

Vous allez entendre à travers le casque qui est placé devant vous des paires d'échantillon de parole qui correspondent à différents cas de communications téléphoniques. Chaque échantillon a une durée de 4,5 secondes et est constitué de deux phrases.

Pendant l'écoute, le bouton rouge qui est devant vous sera allumé. **Vous voudrez bien écouter chaque paire d'échantillons complètement.**

Puis, quand le bouton vert s'allumera, vous devrez évaluer la **différence** entre les deux échantillons de parole, en plaçant le curseur sur l'échelle allant **d'identique** à **très différent**. Vous pouvez placer le curseur où vous voulez entre les deux extrémités, de façon à ce que sa position reflète bien la différence que vous percevez entre les deux échantillons.

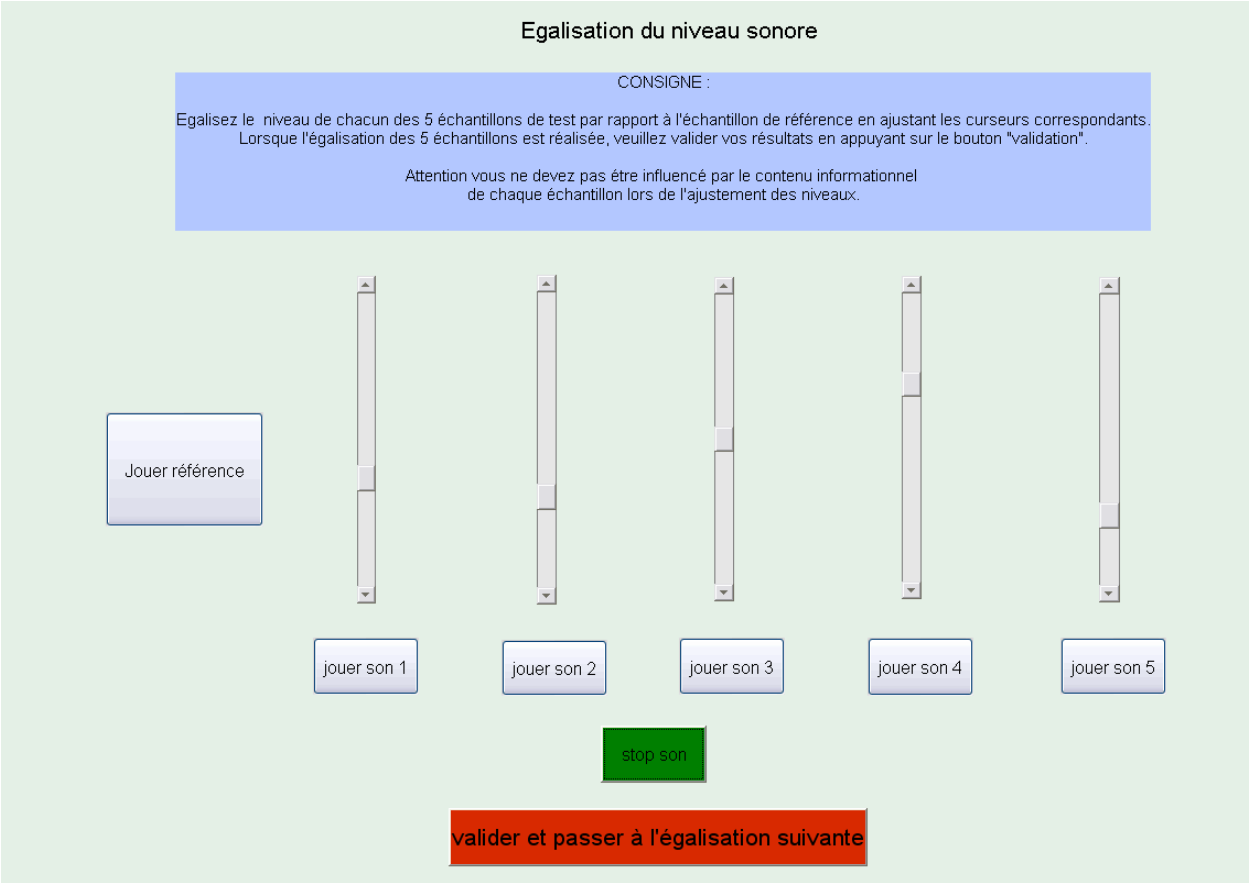
Vous disposez de 5 secondes pour enregistrer votre réponse (temps pendant lequel le bouton vert restera allumé).

Après ces secondes se produira une courte pause avant la paire suivante.

Dans cette expérience nous commencerons par un apprentissage formé de vingt paires d'échantillons. Viendront ensuite les séances d'une durée de 13 minutes chacune environ.

Merci pour votre participation et bon courage.

Annexe D. Interface graphique et consigne du test d'égalisation des niveaux sonores



Fiche de présentation du test d'égalisation des niveaux :

Test d'écoute :

Le but de ce test est d'égaliser en sonie les 5 sons par rapport à la référence.

Mode opératoire :

Pour écouter le son de référence, cliquez sur le bouton suivant : 

Pour écouter les sons à égaliser, cliquez sur les boutons : 

Ajuster le niveau des sons correspondants à l'aide des curseurs : 

Vous avez la possibilité d'arrêter la lecture de n'importe quel son en cliquant sur : 

Lorsque vous avez fini d'égaliser les 5 sons, vous pouvez passer à la deuxième partie en validant vos réponses en appuyant sur le bouton :

→ tournez SVP

1

Consignes du test :

Il est demandé de bien écouter la totalité de chaque son pour le jugement du niveau, spécialement pour ceux non stationnaires!!

Enfin, il vous est demandé de faire abstraction du contenu informationnel des sons, c'est-à-dire d'évaluer le niveau sans s'occuper des différentes natures des sons (que ce soit un son d'environnement, de conversation, ou bruité).

Ne réfléchissez pas trop sur les réponses formulées.
Seules les réponses spontanées nous intéressent.
Il n'y a pas de mauvaises réponses.

Bonne écoute et bon test !

2

Annexe E. Résultats des tests de Student

Des tests de Student ont été appliqués pour analyser les différences d'évaluation de la qualité vocale entre chacune des conditions bruitées. L'analyse est réalisée pour les trois niveaux sonores de bruit de fond. Voici, ci-dessous, les résultats des tests de Student donnés par la valeur *p*. Les valeurs en rouge / gras représentent les valeurs de *p* inférieures à 0,05 correspondant aux notes MOS-LQSN significativement différentes suivant deux conditions bruitées.

		rose	BPS	électrique	ville	restaurant	parole
		BDF1	BDF 2	BDF 3	BDF 4	BDF 5	BDF 6
rose	BDF 1	1	0,941	**0,001**	0,112	0,807	0,052
BPS	BDF 2	0,948	1	**0,001**	0,071	0,822	**0,024**
électrique	BDF 3	**0,001**	**0,001**	1	0,065	**0,001**	0,465
ville	BDF 4	0,113	0,071	0,065	1	0,071	0,370
restaurant	BDF 5	0,807	0,822	**0,001**	0,071	1	**0,004**
parole	BDF 6	0,052	0,024	0,465	0,370	**0,004**	1

Tableau 1 : Valeurs de p issues de tests de Student par paire bilatérale pour le niveau faible de bruit de fond (62 phone)

		rose	BPS	électrique	ville	restaurant	parole
		BDF1	BDF 2	BDF 3	BDF 4	BDF 5	BDF 6
rose	BDF 1	1	0,380	**0,012**	**0,001**	**0,005**	**0,001**
BPS	BDF 2	0,380	1	**0,001**	**0,005**	**0,030**	**0,005**
électrique	BDF 3	**0,012**	**0,001**	1	**0,000**	**0,000**	**0,000**
ville	BDF 4	**0,001**	**0,005**	**0,000**	1	0,604	0,495
restaurant	BDF 5	**0,005**	0,031	**0,000**	0,604	1	0,223
parole	BDF 6	**0,001**	**0,005**	**0,000**	0,495	0,223	1

Tableau 2 : Valeurs de p issues de tests de Student par paire bilatérale pour le niveau faible de bruit de fond (70,5 phone)

		rose	BPS	électrique	ville	restaurant	parole
		BDF1	BDF 2	BDF 3	BDF 4	BDF 5	BDF 6
rose	BDF 1	1	**0,047**	0,767	**0,000**	**0,000**	**0,000**
BPS	BDF 2	**0,047**	1	0,186	**0,005**	**0,000**	**0,000**
électrique	BDF 3	0,767	0,186	1	**0,000**	**0,000**	**0,000**
ville	BDF 4	**0,000**	**0,005**	**0,000**	1	0,469	**0,000**
restaurant	BDF 5	**0,000**	**0,000**	**0,000**	0,469	1	**0,000**
parole	BDF 6	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	1

Tableau 3 : Valeurs de p issues de tests de Student par paire bilatérale pour le niveau faible de bruit de fond (78 phone)

Annexe F. Base sonore utilisée pour la classification des bruits de fond

Quartier	Autoroute	Gd magasin	Parole2
Usine	Bar	Guignol	Peuple
Applaudissement	Bouteille	Int hélico	Place ville
Aspirateur	Bureau	ip02464	Plage
Atelier mécanique	Café	Marché	Voiture R21
Gare	Chien	Mouche	Réunion
Oiseaux	Cigale	Mouette	Route Nationale
Rue	Foire	Moustique	Séchoir à mains
Usine	Foret & rivière	Nature	Sèche cheveux
Vent plage	Gare	Orage	Trafic en ville
Vent fort	Gare2	Ventilateur	Quartier populaire
Ville	Ville2	Ville3	Voiture et pluie

Annexe G. Test préliminaire d'égalisation de la sonie

Les bruits de fond ont été présentés aux auditeurs à des niveaux d'isosonie équivalents afin de pouvoir s'affranchir de l'influence de leur niveau sonore. L'influence du type de bruit de fond lors de l'évaluation de la qualité vocale a pu alors être évaluée.

La méthode d'ajustement a été choisie pour le test préliminaire d'égalisation des sonies, grâce aux recherches effectuées par Boullet [85]. Cette méthode consiste à présenter en alternance un son de comparaison et le son dont on cherche à mesurer la sonie. L'auditeur doit ajuster le niveau du son de comparaison, à l'aide d'un potentiomètre.

Dans notre étude, on cherche à égaliser des sons de natures différentes. Le bruit rose a été choisi comme son de référence pour ses caractéristiques temporelles et spectrales, car la sonie peut être facilement calculée sur ce son stationnaire (Boullet [85]) et qu'il est plus facile d'égaliser un bruit par rapport à un son large bande qu'avec un son composé d'une fréquence pure. Les sons de comparaisons à ajuster sont les cinq autres bruits de fond sélectionnés dans la partie IV.1.2.2. Ce test préliminaire est réalisé en écoute monaurale.

Le principe d'étalonnage du système de restitution est présenté, puis la méthodologie du test préliminaire est décrite. La troisième partie présente les résultats et une dernière partie compare les résultats de l'égalisation de la sonie à des modèles existants.

A. Etalonnage du système de restitution

Notre système de restitution est composé d'une carte son "phase 26 terratec" reliée à un casque monaural "Sennheiser HD 25". Lors de l'étalonnage, le casque a été placé sur les oreilles du mannequin BRUEL & KJAER et nous avons relevé l'intensité acoustique mesurée par le microphone de l'oreille gauche de celui-ci.

Le bruit rose de référence est diffusé à -26 dBov, puis les volumes de la carte son et de Windows sont ajustés pour définir le niveau acoustique du signal vocal optimum correspondant à 79 dB SPL (UIT "Handbook on Telephonometry" [65]).

Figure 1 Etalonnage des bruits de fond à l'aide de la tête acoustique BRUEL & KJAER

Par la suite, nous raisonnerons en niveau physique (dB SPL) correspondant au niveau numérique (dBov). Un niveau de -26 dBov correspond donc à un niveau de 79 dB SPL.

B. Déroulement du test d'égalisation en sonie

Le bruit rose a été diffusé aux trois niveaux retenus (cf. §.IV.1.2.2.D) et il est demandé à vingt sujets d'ajuster les niveaux des cinq bruits de fond par rapport à cette référence pour les trois niveaux.

Nous avons choisi d'afficher les six bruits sur la même fenêtre pour que le sujet ait la possibilité de comparer la sonie des différents types de bruit de fond entre eux. Cela lui simplifie la tâche car il est plus facile d'ajuster des niveaux sonores de deux signaux ayant des caractéristiques spectro-temporelles similaires que deux signaux ayant des caractéristiques très différentes.

Nous avons choisi 20 sujets experts[17] pour ce test afin d'obtenir des résultats plus précis qu'avec des sujets naïfs. L'ordre de diffusion des trois niveaux sonores est aléatoire selon les 20 sujets.

Il y a en tout 100 paliers disponibles pour ajuster le niveau de chaque curseur, correspondant à des niveaux d'intensité sonore allant de $RSB = 46\ dB$ pour le niveau minimum et à $RSB = 9\ dB$ pour le niveau maximum. L'échelle des gains est logarithmique. Il a été vérifié que chaque palier correspond à une différence de niveau inférieure au seuil différentiel du niveau sonore (correspondant à la plus petite variation de niveau sonore détectée par l'oreille humaine).

Les consignes du test ont été données par écrit sous forme de fiche de présentation et oralement (cf. Annexe D).

C. Résultats du test préliminaire d'égalisation des bruits de fond

Les niveaux d'isosonie des six bruits de fond sont présentés suivant leurs niveaux physiques respectifs en dB SPL (cf. Figure 2).

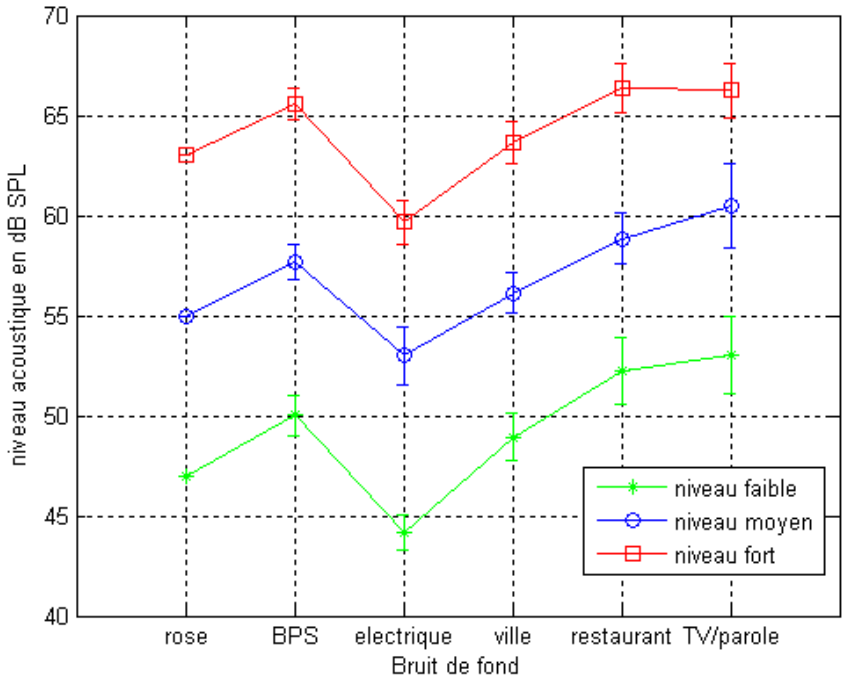


Figure 2 Niveaux isosoniques moyens et intervalles d'incertitude bilatérale à 5% des 6 BDF selon les 20 sujets

[17] Sujets experts : individus ayant l'habitude d'effectuer des tests subjectifs en rapport avec l'audio

Le Tableau 1 récapitule les niveaux physiques exprimés en dB SPL correspondant aux trois niveaux d'isosonie.

	Niveau 1 (dB SPL)	Niveau 2 (dB SPL)	Niveau 3 (dB SPL)
rose	47,0	55,0	63,0
BPS	50,0	57,6	65,5
ville	48,9	56,1	63,6
restaurant	52,2	58,8	66,3
parole	53,0	60,4	66,2
électrique	44,1	53,0	59,6

Tableau 1 Niveau sonore exprimé en dB SPL des 6 bruits de fond correspondant aux 3 niveaux d'isosonie

Pour chacun des trois niveaux d'isosonie, les niveaux sonores exprimés en dB SPL du bruit électrique sont inférieurs à ceux obtenus pour les autres bruits de fond. D'après Hellbrück *et al.* [86], l'effet de la signification d'un son n'a pas d'influence significative sur l'évaluation de la sonie pour des niveaux forts (supérieur à 70 dB SPL). Par contre, à des niveaux plus faibles, comme dans notre expérience, la signification du son peut augmenter le jugement de la sonie.

Nous pouvons déterminer la sonie équivalente des trois niveaux grâce au modèle développé par Zwicker [100] en se basant sur la sonie du bruit rose de référence. Ce bruit est choisi pour représenter la sonie globale des six bruits de fond parce qu'il est stationnaire, et parce que le modèle de sonie de Zwicker est performant pour ce type de bruit de fond, contrairement aux bruits non-stationnaires où il est encore difficile d'utiliser un modèle fiable (Boullet [85]).

Les valeurs isosoniques obtenues en phone peuvent alors être converties en sone par la relation suivante (cette relation est vraie pour des niveaux supérieurs à 40 phones) :

$$L = 40 + 10\log_2(S) \quad \rightarrow \quad S = 2^{\frac{L-40}{10}}, \qquad \textbf{Équation 12}$$

avec L la sonie en phone S la sonie en sones.

Le Tableau 2 représente les valeurs des trois niveaux isosoniques des six bruits de fond ainsi que la sonie du bruit résiduel appelée "niveau silence". Ces valeurs sont exprimées en phone et en sone.

	Niveau silence	Niveau 1	Niveau 2	Niveau 3
Sonie globale en phone	47,4	62,0	70,5	78,0
Sonie globale en sone	1,67	4,6	8,2	14,0

Tableau 2 Sonie du bruit résiduel et les 3 niveaux isosoniques des 6 bruits de fond utilisés dans le test d'évaluation de la qualité vocale, exprimés en phone et en sone

Les différences de sonie entre les niveaux 1-2 et 2-3 correspondent approximativement à un doublement de la perception du niveau sonore. Il est admis que le fait de multiplier par 2 la sonie d'un son provoque une sensation de doublement du niveau sonore. Cela correspond approximativement à ajouter 10 dB SPL ou encore 10 phone à un signal sonore.

Les conclusions données dans Boullet [85] à propos du test d'ajustement spécifient une précision de 4,7 phones. Cette précision est déterminée à partir des moyennes des écarts-types selon les 14 sujets pour différents types de sons.

4,7 phones est équivalent à 4,7 dB SPL pour un son pur de 1 kHz. La consigne était d'ajuster le niveau du son pur pour qu'il soit perçu de même niveau que les sons présentés. C'est pourquoi les écarts-types décrits en phones peuvent aussi être analysés en dB SPL.

Dans le cas de notre expérience, les écarts-types des niveaux sonores obtenus selon les 20 sujets ainsi que leurs moyennes sont exposés dans le tableau ci-dessous en dB SPL :

Ecart type en dB SPL	BPS	Elec-trique	Ville	Restau-rant	parole	Moyenne
Niveau 1 (47)	2,3	2,0	2,7	3,8	4,4	**3,0**
Niveau 2 (55)	2,0	3,3	2,3	2,9	4,9	**3,1**
Niveau 3 (63)	1,9	2,5	2,4	2,8	3,1	**2,5**
Moyenne	**2,0**	**2,6**	**2,5**	**3,2**	**4,1**	**2,9**

Tableau 3 Ecarts-types des niveaux sonores obtenus selon les 20 sujets lors du test d'ajustement ainsi que leurs moyennes en dB SPL

La moyenne globale des écarts-types de notre test est de 2,9 dB SPL, ce qui représente une meilleure précision que celle décrite dans la thèse de Boullet (4,7 phones).

Voici quelques hypothèses pouvant justifier la meilleure précision obtenue dans notre expérience :

- Le bruit rose de référence est plus ressemblant aux autres sons à égaliser en terme de contenu fréquentiel, par rapport à un son de fréquence pur de 1 kHz dans le cas du test d'Isabelle Boullet;
- La présentation des six bruits sur une interface dans notre cas, au lieu d'une comparaison par paires;
- La durée des stimuli (8 s dans notre test contre 1 s);
- 20 sujets dans notre cas, au lieu de 14 dans le cas du test d'Isabelle Boullet;
- Les auditeurs sont experts dans notre cas, contrairement à l'étude d'Isabelle Boullet qui utilise des sujets naïfs;
- Restitution au casque dans notre cas, et par une enceinte dans le cas d'Isabelle Boullet;
- Dans notre cas, les niveaux sonores des bruits à égaliser sont approximativement compris entre 45 et 65 dB SPL, contre des différences de niveaux allant de 40 à 89 dB SPL dans le cas du test d'Isabelle Boullet.

Nous observons la même tendance que l'étude d'Isabelle Boullet concernant l'augmentation de la précision des résultats selon les sujets pour des niveaux de plus en plus forts. Cela peut être expliqué par la meilleure sensibilité du seuil différentiel de l'oreille pour des niveaux forts que pour des niveaux plus faibles.

D. Comparaison des résultats du test d'égalisation en sonie avec les modèles existants

Plusieurs indicateurs d'estimation du niveau sonore sont testés sur les 18 bruits de fond (6 bruits diffusés aux 3 niveaux d'isosonie) selon qu'il s'agit de bruits stationnaires ou non stationnaires :

Les indicateurs de niveau sonore pour les bruits stationnaires sont

- le niveau en dB SPL
- le niveau en dB(A)
- la sonie stationnaire en phone (Zwicker ISO532B).

Lorsque les bruits sont non-stationnaires, il est préférable d'utiliser les estimateurs de sonie suivants :

- Sonie maximum en phones
- N5 en phones (Zwicker)
- N10 en phones (Zwicker)

Nous avons retenu l'indicateur de la sonie stationnaire par le modèle de Zwicker pour les bruits stationnaires et le *N5* pour les sons non stationnaires (Zwicker et Fastl [38]). Le *N5* correspond à la sonie dépassée pendant 5 % du temps total du stimulus. Les résultats sont présentés sur le Tableau 4.

Niveau / BDF	Niveau faible 62 phone	Niveau moyen 70,5 phone	Niveau fort 78 phone	Méthode "Zwicker"
Rose	62,1	70,4	77,9	Sonie
BPS	62,8	71,0	78,5	Sonie
Electrique	58,4	68,1	74,9	Sonie
Ville	63,6	70,1	77,2	N5
Restaurant	64,8	71,2	77,4	N5
Parole	66,0	72,3	77,6	N5

Tableau 4 Valeurs des niveaux sonores en phone, déterminées par les modèles de Zwicker existants (sonie stationnaire et N5)

Les six bruits de fond ont été égalisés à trois niveaux d'isosonie par rapport au bruit rose de référence lors du test préliminaire. Nous obtenons donc des sonies égales à 62 phone, 70,5 phone et 78 phone pour les six bruits de fond. Ces trois niveaux sont comparés à l'estimation de la sonie par les modèles proposés par Zwicker (cf. Figure 3).

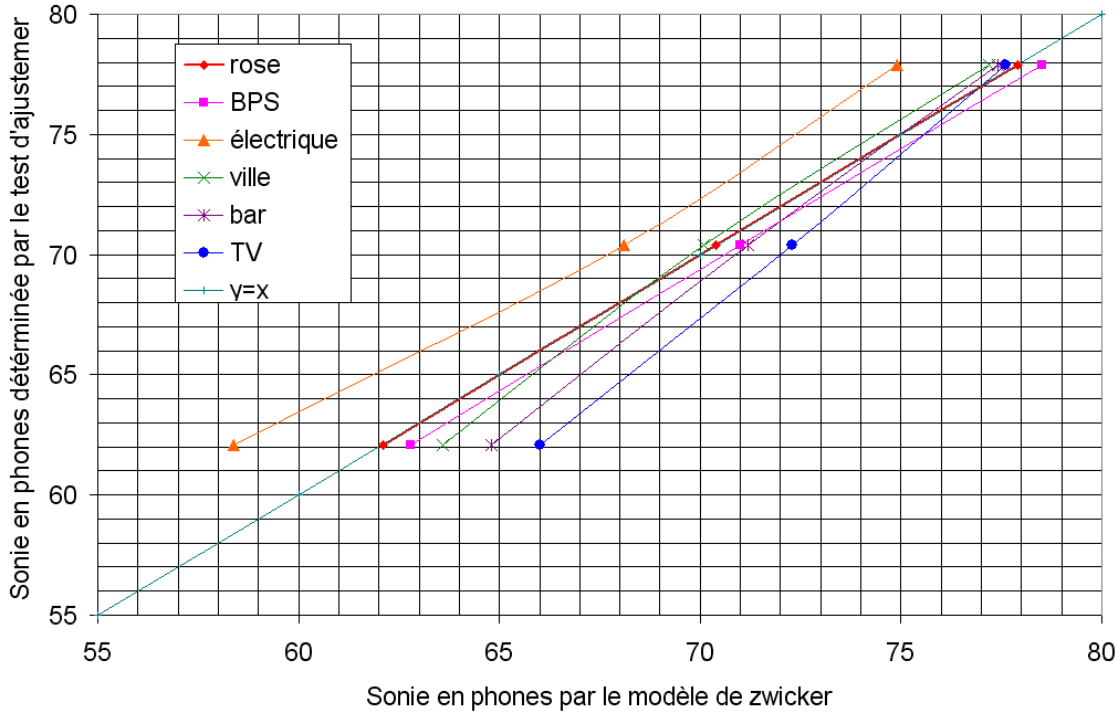


Figure 3 Comparaison des niveaux d'isosonie issus du test d'ajustement et issus des indicateurs de Zwicker (N5 et sonie stationnaire)

La sonie du bruit de fond électrique est sous-estimée par le modèle de Zwicker. Pour chacun des trois niveaux, nous déterminons les écarts-types des sonies obtenus par les indica-

teurs de Zwicker suivant les différents bruits de fond (avec et sans le bruit électrique). Les écarts-types sont présentés dans le tableau ci-dessous :

	Niveau 1	Niveau 2	Niveau 3
6 BDF	2,63	1,41	1,24
5 BDF (sans elect)	1,56	0,85	0,51

Tableau 5 Ecarts types des sonies suivant les différents bruits de fond en phone

L'écart type est de 1,24 pour le niveau fort contre 2,63 dans le cas du niveau faible. Les estimations des niveaux d'isosonie par les indicateurs sont de plus en plus précises lorsque la sonie des bruits augmente (cf. Figure 3).

Annexe H. Publications, contributions et brevets personnels

Publications :

- A. Leman, J. Faure, E. Parizet : "Diagnostic & Evaluation automatique de la qualité vocale à partir d'indicateurs hybrides" JJCAAS – Paris, 2010

- A. Leman, J. Faure, E. Parizet : "Hybrid model for non-intrusive speech quality evaluation in telephony applications" AES 38th – Piteå, Sweden, 2010

- A. Leman, J. Faure, E. Parizet : "Modèle hybride pour l'évaluation de la qualité vocale sans référence, appliqué à la téléphonie" CFA – Lyon, 2010

- A. Leman, J. Faure, E. Parizet : "Modèle basé sur le signal non-intrusif de l'évaluation de la qualité vocale, utilisant une méthode de classification automatique des bruits de fond" JJCAAS - Marseille, 2009

- A. Leman, J. Faure, E. Parizet : "A non-intrusive signal-based model for speech quality evaluation using automatic classification of background noises", Interspeech conference - Brighton, 2009

- A. Leman, J. Faure, E. Parizet : "Influence of informational content of background noise on speech quality evaluation for VoIP application" Acoustics08 – Paris, 2008.

Contributions à l'UIT :

- A. Leman, J. Faure, E. Parizet : " A non-intrusive signal-based model for speech quality evaluation using automatic classification of background noises " proposé en contribution à l'Union International des Télécommunication (UIT-T) COM 12, Q.15/9 – Genève, 2009.

- A. Leman, J. Faure, E. Parizet : "Influence of informational content of background noise on speech quality evaluation for VoIP application" proposé en contribution à l'Union International des Télécommunication (UIT-T) COM 12, Q.15/9/8/7/12 – Genève, 2009.

Dépôt de Brevets :

- A. Leman, J. Faure : Brevet d'invention "Procédé et dispositif d'évaluation objective de la qualité vocale d'un signal de parole prenant en compte la classification du bruit de fond contenu dans le signal" déposé le 17/04/2009, V/Ref : 07338-FR, N/Ref : 0952531, L059

- A. Leman, J. Faure : Brevet d'invention "Procédé et dispositif de classification du bruit de fond contenu dans un signal audio" déposé le 31/03/2009, V/Ref : 07314-FR, N/Ref : 0952053, L059

Résumé

Les services de télécommunication sont de plus en plus nombreux et variés avec l'apparition de nouvelles technologies (RTC, RNIS, GSM, VoIP). Les opérateurs de téléphonie ont ainsi besoin de superviser en temps réel la qualité vocale des services qu'ils proposent. La qualité vocale peut être évaluée par des campagnes de tests subjectifs en demandant directement l'avis aux utilisateurs, cependant les méthodes existantes sont très coûteuses et peu adaptées à la supervision. Les modèles objectifs sont ainsi proposés afin de prédire la qualité vocale à moindre coût.

Cette thèse propose un modèle de diagnostic et d'évaluation de la qualité vocale, utilisant les informations disponibles au point de mesure : le modèle DESQHI (Diagnostic and Evaluation of Speech Quality using Hybrid Indicators). Il se démarque des modèles existants par deux caractéristiques principales.

La première concerne la structure du cœur du modèle. Il est montré que l'évaluation de la qualité vocale peut être représentée comme un phénomène multidimensionnel faisant intervenir trois dimensions perceptives correspondant à la *bruyance*, au *codage de la parole* et à la *continuité*. Cette structure permet de diagnostiquer la qualité vocale en identifiant le ou les principale(s) cause(s) perceptive(s) de la dégradation de la qualité vocale.

La deuxième caractéristique concerne le type d'indicateur utilisé pour représenter chacune des trois dimensions perceptives, à savoir l'utilisation d'indicateurs basés sur le signal et paramétriques. Les indicateurs basés sur le signal utilisent les informations numériques pour représenter les caractéristiques du signal comme par exemple le rapport signal sur bruit qui donne une estimation du niveau sonore du bruit de fond. Les indicateurs paramétriques sont issus des statistiques du réseau, comme par exemple le pourcentage de pertes de paquets qui fournit une indication sur le niveau de discontinuité du signal de parole.

L'utilisation d'indicateurs hybrides composés à la fois des informations du signal numérique et à la fois des statistiques du réseau permet d'améliorer les performances globales de la prédiction de la qualité vocale, comparativement aux modèles uniquement basés sur le signal (p. ex. modèle P.563) et aux modèles utilisant les indicateurs paramétriques (p. ex. modèle E).

Mots clés : qualité vocale, perception sonore, diagnostic réseau

Abstract

With increasing development of new technologies (RTC, RNIS, GSM, VoIP), telecommunication services are becoming more and more diversified. To this end, telecommunication operators need to supervise in real-time the speech quality of the services they offer. Speech quality is usually evaluated from subjective experiments. Nevertheless, such experiments are time consuming and do not allow any supervisory control. So, accurate objective models are useful to estimate the speech quality

This thesis proposes a non-intrusive model for diagnosing and evaluating speech quality using information available at the measurement point: the DESQHI model (Diagnostic and Evaluation of Speech Quality using Hybrid Indicators). It differs from existing models in terms in two main characteristics.

The first one concerns the structure of the model. It is shown that speech quality can be represented as a multidimensional phenomenon incorporating three perceptual dimensions related to *noisiness*, *speech codec* and *continuity*. This multidimensional structure allows for a diagnostic of speech quality based on identifying the principal features affecting speech quality.

The second characteristic concerns the nature of indicators (signal-based and parametric) used to represent the three perceptual dimensions. Signal-based indicators use numeric information to represent the characteristics of the signal, for example, the loudness of the speech signal. Parametric indicators are obtained from the network statistics, for example, the percentage of packet loss, which gives information about the level of the discontinuity in the speech signal. This work proposes hybrid indicators (using both signal-based and parametric metrics). It is shown that they are better speech quality predictors than existing models, either parametric only (e.g. ITU-T Recommendation G.107, also known as the E-model) or signal-based only (e.g. ITU-T Recommendation P.563 model).

Keywords : speech quality, sound perception, network diagnostic tests.