Amélioration de l'intelligibilité de signaux audio de parole en contexte bruité automobile

Enguerrand Gentet

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAT008

Thèse de doctorat

Amélioration de l'intelligibilité des signaux audio de parole en contexte bruité automobile

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat: Signal, Images, Automatique et robotique

Thèse présentée et soutenue à Palaiseau, le 31/03/2021, par

ENGUERRAND GENTET

Composition du Jury :

Christophe d'Alessandro Directeur de Recherche CNRS, Sorbonne Université (Institut Jean le Rond ∂'Alembert)	Président
Yannis Stylianou Professeur, Université de Crète	Rapporteur
Étienne Parizet Professeur, INSA de Lyon (LVA)	Rapporteur
Maëva Garnier Chargée de Recherche CNRS, Université de Grenoble (GIPSA-LAB)	Examinatrice
Bertrand David Professeur, Télécom Paris	Directeur de thèse
Sébastien Denjean Docteur, Stellantis	Co-Directeur de thèse
Gaël Richard Professeur, Télécom Paris	Co-Encadrant de thèse
Vincent Roussarie Docteur, Stellantis	Co-Encadrant de thèse

Remerciements

Cette thèse a été réalisée dans le cadre d'une convention CIFRE gérée par l'Association Nationale de la Recherche Technique (ANRT) et établie entre Télécom Paris et la société Stellantis.

En premier lieu, je tiens à remercier toute l'équipe d'encadrement pour m'avoir permis de mener à bien ce travail de thèse. Au delà du cadre scientifique et du suivi rigoureux que vous avez maintenu tout au long de ma thèse, chacun d'entre vous m'a aussi grandement apporté sur de multiples aspects qui vous étaient propres. Merci à M. Bertrand David, pour m'avoir donné les outils organisationnels afin d'aborder et de développer mon travail de thèse dans de bonnes conditions, me facilitant alors grandement ces trois longues années de travail. Merci à M. Sébastien Denjean, pour m'avoir accompagné dans ma découverte du monde de l'entreprise ainsi que pour son oreille attentive lorsque les doutes et remises en question commençaient à se présenter. Merci à M. Gaël Richard, pour avoir toujours su dégager du temps pour des relectures très détaillées de mes livrables et pour ses conseils permettant de m'améliorer sans cesse sur la présentation de mes travaux. Merci à M. Vincent Roussarie, pour son regard strict et pédagogique sur le bon déroulement d'un travail de recherche au sein d'une entreprise.

Je tiens aussi à remercier M. Claude Barras et M. Christophe d'Alessandro pour avoir accepté de faire partie du comité d'évaluation à mi-parcours, vos retours pertinents et encourageants m'ont permis d'aborder la seconde partie de cette thèse dans de très bonnes dispositions. J'étais d'autant plus touché que M. d'Alessandro ait bien voulu continuer le suivi en acceptant de faire partie des membres du jury au sein duquel il sera nommé président permettant alors à la soutenance de se dérouler dans les meilleures conditions. Merci à tous les autres membres du jury qui ont alloué de leur précieux temps à l'évaluation de ces travaux de thèse. En commençant par les rapporteurs : merci à M. Yannis Stylianou pour avoir relu le manuscrit qui n'était pourtant pas rédigé dans sa langue natale et merci à Etienne Parizet pour son travail méticuleux de relecture ayant permis de mettre le doigt sur quelques incohérences et de nombreuses coquilles ! Merci à Mme Maëva Garnier pour avoir accepté d'examiner les travaux de thèse, cela a été un honneur de compter parmi les membres de mon jury l'auteure du premier manuscrit de thèse que j'ai lu dans son intégralité. J'en profite également pour remercier le GIPSA-LAB de l'Université de Grenoble, dont Mme Maëva Garnier fait partie, pour m'avoir accueilli une semaine dans leur locaux afin de présenter mes travaux et d'interagir avec les chercheurs passionnants du laboratoire.

Je termine par une profonde pensée à ma famille, mes amis et mes collègues pour leur soutien sans faille durant ces trois ans. Enfin, je remercie tout particulièrement ma compagne, Anaïs Vaquieri, qui a partagé ma vie durant cette épopée, merci pour ton soutien indéfectible et pour ton affection qui ont très largement participé au bon déroulement de ces travaux.

Tables des matières

Liste des acronymes

AI *Articulation Index*. 24, 63

ANN *Artificial Neural Network*. 129–131, 133, 136, 137, 156, 157, 159

ANSI *American National Standards Institute*. 68–72, 82

API Alphabet Phonétique International. 32

ASA *American Standards Association*. 68–72, 82

BD bidirectionnel. 137, 156, 157, 159–164

BV Basse Vitesse. 106, 107, 109, 112, 115, 116

BV+P Basse Vitesse avec Pluie. 106, 107, 109, 112, 115

BV+PL Basse Vitesse avec Pluie Lissée. 112, 115, 116

CWT *Continuous Wavelet Transform*. 132, 140, 141, 155, 157, 159–162, 164, 166

DTW *Dynamic Time Warping*. 142–144, 149

FFNN FeedForward Neural Network. 129–131, 133, 136, 138, 155, 156, 159–162, 164, 166, 172

FIB Fonction d'Importance de Bande. 24, 25, 63, 71, 72, 76, 81, 82, 84, 94, 109, 116

FR *Fully Recurrent*. 156, 159, 160

FTM Fonction de Transfert de Modulation. 25, 93

GMM *Gaussian Mixture Model*. 127, 128, 130, 131, 133, 136, 155–157, 159, 160, 163, 172

GP *Glimpse Proportion*. 26, 123

GRU *Gated Recurent Unit*. 139, 140, 156, 159–164, 166

GV Grande Vitesse. 106, 109, 112, 115, 116

HINT *Hearing In the Noise Test*. 112

LPC *Linear Predictive Coding*. 34–36, 124, 132

LSP *Line Spectral Pairs*. 35, 124, 131, 132

LSTM *Long Short-Term Memory*. 138, 139, 156, 157, 159, 160, 163, 164, 166

MCD *Mel-Cepstral Distance*. 142, 157, 160

MFCC Mel Frequency Cepstral Coefficients. 131, 132, 142, 144, 154, 157, 159, 160, 164

MLPG *Maximum Likelihood Parameter Generation*. 128, 155

p.p. point de pourcentage. 48, 49, 82, 84, 96–100, 123–126, 160

ReLU *Rectified Linear Unit*. 138, 156

RMSE *Root Mean Squared Error*. 157, 159, 160, 164

RNN *Recurrent Neural Network*. 131, 132, 136–140, 150, 155, 156, 159, 160, 163, 164, 166, 172

RSB Rapport Signal sur Bruit. 4, 5, 20, 24–26, 62, 63, 68, 79–88, 92–97, 100, 103, 106–112, 116, 124, 171–173

RSP Rapport Signal sur Perturbation. 24, 71, 103

SII *Speech Intelligibility Index*. 5–8, 24–26, 63, 64, 68, 72, 73, 76, 78, 79, 81–88, 92–94, 96, 97, 102, 103, 106, 107, 109–111, 116, 117, 132, 171, 172

SIIB *Speech Intelligibility In Bits*. 26, 132

SMN *Speech Modulated Noise*. 20, 62, 125, 126

SPL *sound pressure level*. 14, 70, 94, 106, 112

SRP Seuil de Réception de la Parole. 22, 57, 100–102, 112–117

SSN *Speech Shaped Noise*. 20, 23, 43, 48, 49, 62, 70, 79–85, 87, 88, 109, 123–125, 174

STI *Speech Transmission Index*. 25

STRAIGHT *Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum*. 6, 36–39, 48, 132, 142, 144, 154, 166

TD-PSOLA *Time Domain Pitch Synchronous Overlap and Add*. 34, 123, 125

TFCT Transformée de Fourier à Court-Terme. 31–36, 49, 93

UD unidirectionnel. 137, 156, 159, 160

VBGMM *Variational Bayesian GMM*. 128, 156

WSOLA *Waveform Similarity Overlap and Add*. 34, 125

Introduction

Contexte

Dans les habitacles automobiles, conducteur et passagers sont exposés à une diffusion croissante de signaux de parole, qu'il s'agisse de télécommunications, de radiodiffusion ou encore d'informations transmises par l'agent conversationnel et le système de navigation. Malgré les avancées mécaniques et aérodynamiques pour diminuer les bruits dans les habitacles, il reste tout de même beaucoup de ces perturbations acoustiques. Cela peut conduire à une augmentation trop importante du volume ou à une mauvaise compréhension de messages importants (information, guidage) pouvant, de plus, entraîner des situations potentiellement dangereuses (dégradation de l'audition, perte d'attention sur la conduite).

Lors d'une conversation naturelle dans un tel environnement d'écoute dégradé, un locuteur peut bien sûr parler plus fort mais il utilise, souvent de manière réflexe, différentes stratégies pour modifier sa production de parole afin de mieux se faire comprendre. La parole Lombard et la parole claire en sont deux exemples. La parole Lombard fait référence à des modifications naturelles de la prosodie introduites par un locuteur afin d'améliorer son intelligibilité lorsqu'il s'exprime dans un environnement bruyant. La parole claire fait référence aux modifications naturelles de la prosodie introduites par un locuteur pour améliorer son intelligibilité dans un environnement non perturbé mais en tenant compte de conditions d'écoute dégradée dont il a préalablement conscience. Il peut par exemple s'agir d'un canal de transmission bruité ou d'un auditeur atteint de déficience auditive. Il a été montré que ces deux types de parole procurent des gains d'intelligibilité très significatifs dans des environnements bruyants, même sans augmentation du volume de la voix. S'appuyant sur ces résultats, notre approche est de proposer des traitements numériques susceptibles d'améliorer sensiblement l'intelligibilité de la parole dans le contexte particulier du bruit automobile

Du rehaussement au renforcement de la parole dans les habitacles automobiles

Bien que le bruit soit souvent le facteur incriminé dans la dégradation de l'intelligibilité, les signaux de parole sont dégradés à de multiples niveaux. D'abord à la source qui peut se situer dans un environnement bruyant ou réverbérant, et le matériel de prise audio peut potentiellement introduire des artefacts, de la distorsion voir du bruit d'enregistrement. Puis la chaîne de transmission peut ajouter du bruit de transmission ou des échos de ligne, le matériel de restitution audio peut aussi introduire des artefacts ou de la distorsion. Enfin, l'environnement d'écoute automobile est, effectivement, bruyant mais aussi réverbérant. Ce sont tous ces facteurs qui dégradent l'intelligibilité des signaux de parole et réduisent significativement l'expérience de l'utilisateur. De nombreuses pistes de recherche cherchent à traiter ces différents aspects et elles peuvent être regroupées en deux domaines d'étude :

- Le rehaussement de la parole, ou *speech enhancement*, qui consiste à traiter des signaux de parole qui ont déjà été dégradés préalablement à leur diffusion dans le but d'améliorer leur intelligibilité et leur qualité, la parole et la dégradation sont donc directement présents dans le signal à traiter. La réduction de bruit est un sous-domaine du rehaussement de la parole visant à améliorer les signaux de parole dégradés par un bruit additif. Ce type de dégradation se retrouve dans quasiment toutes les applications concernant des signaux de parole ce qui en fait un problème majeur auquel de nombreuses propositions de traitement ont été faites. La réduction de bruit a pris une telle ampleur qu'il arrive souvent que l'on y fasse référence lorsque l'on parle de rehaussement de la parole. C'est ainsi que LOIZOU fusionne les deux termes dans son ouvrage de référence *Speech enhancement: theory and practice* [112]. Cependant le rehaussement de la parole englobe deux autres types de traitements, le contrôle d'écho et la dé-réverbération, prenant en compte les dégradations restantes couplées au signal. Les algorithmes de rehaussement de la parole commencent à atteindre un degré de maturité suffisant pour que ce ne soit pas une source majeure de diminution de l'intelligibilité pour une écoute classique en contexte automobile.
- Le renforcement de la parole consiste à prendre en compte les conditions acoustiques de diffusion du signal et à traiter ce dernier en prévision des effets qui viendront le dégrader. Les applications sujettes à cette problématique sont nombreuses et peuvent concerner des domaines très différents allant de la téléphonie

mobile aux systèmes de diffusion d'information publique dans les gares et aéroports. Une grande attention a été portée à ce domaine de recherche cette dernière décennie avec des travaux de plus en plus nombreux. L'organisation de la première évaluation internationale de renforcement de la parole par COOKE et al. en 2013, le challenge *Hurricane* [45], atteste de cet intérêt. De même que pour le rehaussement de la parole, lorsque l'on parle de renforcement, on fait souvent référence aux traitements visant à traiter les signaux de parole qui vont être diffusés dans du bruit comme le suggèrent SAUERT et al. [172]. Cependant, le renforcement de la parole peut aussi englober un spectre plus large de traitements prenant en compte d'autres dégradations introduites par l'environnement d'écoute. Par exemple, l'écoute de signaux de parole dans un environnement réverbérant est une problématique bien présente en renforcement de la parole à tel point que la deuxième édition du challenge *Hurricane*, organisée en 2020 [162], a introduit la réverbération dans ses paramètres d'études.

La réverbération pourrait sembler être un facteur majeur responsable de la dégradation de l'intelligibilité dans les habitacles automobiles qui sont des espaces clos, cependant l'utilisation de matériaux acoustiques absorbants, devenue habituelle chez les constructeurs automobiles, permet de minimiser grandement le temps de réverbération des signaux audio [50] et de rendre ce facteur quasi-négligeable du point de vue de l'intelligibilité [51]. En revanche, la directivité du signal reste un facteur important et plus particulièrement lors d'échanges entre les passagers d'un véhicule [213]. En renforcement de la parole, les signaux traités sont diffusés dans l'habitacle par des haut-parleurs dont le positionnement et la commande panoramique permettent d'obtenir une bonne exposition pour tous les passagers, la directivité n'est donc pas un problème dans ce cas. Ainsi, la présence de bruit est bien le facteur principal de la dégradation de l'intelligibilité en voiture. On notera tout de même que la charge cognitive subie par le conducteur est aussi un facteur qu'il faudrait prendre en compte [8] mais la difficulté pour la mesurer fait qu'elle est généralement négligée dans les études actuelles.

En rehaussement de la parole, la problématique principale de la réduction de bruit est de réduire l'information relative au bruit vis-à-vis de l'information liée à la parole et donc de traiter un signal bruité afin d'augmenter son Rapport Signal sur Bruit (RSB). En renforcement de la parole dans le bruit, le signal de parole, considéré non-bruité, est traité avant sa diffusion dans le bruit, l'augmentation du RSB peut donc se faire en augmentant directement le niveau de présentation du signal de parole. Cette approche naïve est efficace mais atteint rapidement ses limites que ce soit au niveau des limitations matérielles ou du confort d'écoute pour l'auditeur. C'est pourquoi les approches de renforcement de la parole dans le bruit sont majoritairement menées sous contrainte énergétique qui reflète ces limites présentes dans des conditions d'écoute réelles. La contrainte énergétique classique systématiquement utilisée dans les études actuelles consiste à maintenir constante l'énergie moyenne des signaux.

Il existe tout de même des cas de figure spécifiques pour lesquels l'utilisation de la contrainte peut être relaxée, par exemple lors de l'apparition ou l'augmentation d'un bruit. Des travaux proposent donc d'amplifier les signaux de parole en se basant sur des modèles perceptifs [131, 129] afin de rétablir le niveau sonore perçu existant avant la perturbation [179, 178]. Ce principe, appelé contrôle automatique de gain, est très largement utilisé dans de nombreux domaines applicatifs où le niveau de bruit peut soudainement changer, comme l'automobile ou l'équipement audio nomade, mais aussi en l'absence de bruit afin de réduire la dynamique entre différentes sources, lors d'un changement de station de radio par exemple.

Renforcement de la parole dans le bruit sous contrainte énergétique

Il existe deux grandes familles d'approches en renforcement de la parole dans le bruit. Les approches directes qui consistent à utiliser des méthodes de manipulation et de filtrage classiques des signaux audio. Elles diffèrent des approches paramétriques qui se basent sur l'utilisation d'algorithmes d'analyse-modifications-synthèse de la parole, ou vocodeurs, permettant d'extraire des paramètres vocaux puis de les modifier afin de manipuler la prosodie de manière naturelle.

La contrainte énergétique empêche d'augmenter le volume global du signal de parole mais il reste possible de redistribuer son énergie dans différentes zones temporelles. Une idée très exploitée en renforcement direct de

la parole est de chercher à réduire la dynamique du signal [181] en ré-équilibrant le niveau entre les segments de faible amplitude, principalement les consonnes, et ceux de forte amplitude, principalement les voyelles. D'autres approches directes cherchent à amplifier directement les transitoires qui ont un rôle majeur pour l'intelligibilité [220, 198, 161]. Une redistribution de l'énergie spectrale est aussi envisagée dans de nombreuses approches directes qui se concentrent sur des filtrages visant à rehausser les formants dont l'importance pour l'intelligibilité de la parole est avérée. Cela peut se faire par filtrage fixe, par l'utilisation de filtres passe-haut [202, 147, 35] ou de filtres spécifiques [74], mais aussi par filtrage variable en suivant l'évolution des formants par des méthodes d'analyse conditionnant alors un filtrage adaptatif [82]. Il est tout à fait possible de combiner les deux principes en procédant à des redistributions spectrales et temporelles indépendantes [79, 221, 69, 167]. Certaines approches proposent même de traiter ces redistributions conjointement à partir d'observations connues sur l'intelligibilité de la parole [171, 169, 194, 196] ou en s'appuyant sur des mesures objectives de la parole en cherchant à maximiser un critère d'intelligibilité [160, 170, 195, 193, 183].

Les approches paramétriques proposent des libertés de traitement supplémentaires en permettant de manipuler la prosodie des signaux de parole. Bien souvent elles s'inspirent des modifications introduites par la parole Lombard ou la parole claire. Ainsi on retrouve des approches qui utilisent des modèles de parole afin d'introduire des manipulations empiriques du fondamental [151, 124, 79, 211, 215], de l'enveloppe spectrale [123, 79, 145] ou encore du débit [44, 79, 10]. La flexibilité procurée par les méthodes paramétriques permet d'aller plus loin en couplant ces méthodes avec de l'apprentissage statistique et en apprenant automatiquement les modifications prosodiques à apporter aux signaux de parole à partir d'exemples de parole Lombard, ou de parole claire. Cela peut permettre d'introduire des transformations contextuelles qui ne sont pas recensées dans les études des styles de parole mais qui peuvent être utiles à l'intelligibilité. En transformant conjointement les paramètres de la parole, on peut aussi s'attendre à une meilleure cohérence entre eux et un rendu plus naturel de la parole synthétisée qu'avec les approches empiriques.

Contributions de la thèse

Maximisation exacte d'un critère d'intelligibilité dans le bruit sous contrainte énergétique pondérée

Une approche de renforcement direct dans le bruit très appréciée dans la littérature est l'optimisation d'un critère d'intelligibilité par l'utilisation d'un égaliseur fréquentiel. Le *Speech Intelligibility Index* (SII) est une mesure très performante basée sur le calcul de RSB sur différents canaux fréquentiels et est parfaitement adaptée à notre domaine d'étude du fait que le bruit présent dans un habitacle automobile peut être mesuré par des microphones ou estimé à partir des paramètres connus du véhicule (modèle, vitesse, rapport…). La stratégie d'optimisation du SII a déjà été suivie dans la littérature, que ce soit par approximations linéaires [170] ou non-linéaires [193, 183] de la mesure, engendrant alors une augmentation très significative de l'intelligibilité lors de tests subjectifs dans des bruits d'études classiques. Dans tous ces travaux, la contrainte énergétique utilisée n'était pas pondérée et, au vu des spectres traités obtenus, il est avéré que le niveau perçu a été augmenté en concentrant l'énergie spectrale du signal dans des zones où l'oreille est plus sensible.

Ainsi, en plus de proposer une méthode de résolution exacte du problème de maximisation du SII qui permet d'interpréter et de comparer les résultats de l'optimisation de façon inédite, nos travaux proposent d'introduire une contrainte énergétique nouvelle basée sur l'utilisation d'une échelle perceptive afin de maintenir les signaux à leur niveau perçu d'origine. Des évaluations subjectives ont alors été menées permettant d'analyser l'intérêt de ces apports. D'abord par une participation à la deuxième édition du challenge *Hurricane* [162], afin d'étudier les gains d'intelligibilité obtenus par l'optimisation exacte du SII. Puis par la mise en place d'un test d'intelligibilité dans plusieurs contextes bruités automobiles sous la nouvelle contrainte perceptive [66, 67], afin d'analyser le comportement de l'approche sous cette contrainte novatrice.

Renforcement par conversion du style de parole et aspects temporels

Quelques travaux abordent la piste du renforcement paramétrique par conversion de la parole neutre vers la parole Lombard et obtiennent des scores de similarité du style très intéressants [175, 176, 108]. Cependant, les gains d'intelligibilité sont encore peu étudiés et, lorsqu'ils le sont, les résultats n'atteignent pas les performances procurées par des modifications naturelles [176]. Les auteurs avancent que les raisons principalement responsables sont les dégradations introduites par le vocodeur et le manque de naturel de la prosodie convertie. Nous proposons une explication complémentaire en remarquant que ces travaux mettent de côté de nombreux aspects temporels de la parole, que ce soit par les faibles contraintes imposées sur les trajectoires temporelles des paramètres modifiés, ou tout simplement par l'utilisation de modifications trop simplistes du débit de parole.

Nos travaux proposent alors de se concentrer sur les aspects temporels par l'exploitation de modèles d'apprentissage, et l'utilisation de traitements des caractéristiques, adaptés à l'analyse de séquences temporelles longues [65]. Nous proposons aussi une modélisation nouvelle des modifications du débit de parole directement intégrable dans l'apprentissage machine, ce qui n'avait alors jamais été fait auparavant en conversion de la parole, donnant ainsi lieu au dépôt d'un brevet d'invention [68].

Structure du document

Le document est structuré en trois parties regroupant chacune plusieurs chapitres. La figure 1 présente un schéma des principales dépendances entre les chapitres.

Première partie :

Dans la première partie sont présentés plusieurs acquis et connaissances portant sur la parole et son intelligibilité sur lesquels s'appuieront nos études et contributions.

Le chapitre 1 commence par expliquer les aspects perceptifs qui sont mis en jeu lors d'une écoute dévoilant ainsi les différents facteurs qui peuvent influencer l'intelligibilité de la parole. Ces notions nous permettent alors de détailler et de justifier les différents outils à notre disposition pour mesurer l'intelligibilité de la parole. D'abord par des mesures subjectives, pour lesquelles des tests d'intelligibilité doivent être mis en place avec des protocoles rigoureux prenant en compte les différents facteurs qui auront été introduits. Puis par des mesures objectives, qui facilitent l'estimation de l'intelligibilité en proposant des critères mathématiques basés sur des résultats d'études perceptives. Ce chapitre est un socle de connaissances sur lesquelles s'appuieront de nombreux choix et interprétations de notre étude.

Le chapitre 2 s'oriente plus du côté des mécanismes de production de la parole et des interprétations acoustiques que l'on peut faire à partir d'une analyse court-terme. Une bonne compréhension du modèle source-filtre et des caractéristiques acoustiques de la parole observables nous permet alors d'introduire les principaux algorithmes d'analyse-modifications-synthèse de la parole, ou vocodeurs. Nous nous concentrerons principalement sur le vocodeur STRAIGHT et sur les raisons pour lesquelles nous l'utiliserons pour nos contributions de renforcement paramétrique dans la partie 3.

Le chapitre 3 présente les deux principales stratégies d'adaptations naturelles de la parole permettant une amélioration de l'intelligibilité dans le bruit : la parole Lombard et la parole claire. Leur principe, les principales modifications recensées et leurs intérêts vis-à-vis du gain d'intelligibilité dans le bruit, sont détaillés dans ce chapitre. Il est important d'étudier et comprendre ces adaptations naturelles car elles sont à la base de nombreuses approches de renforcement de la parole dans le bruit, et plus particulièrement pour les approches paramétriques.

Deuxième partie :

La deuxième partie est consacrée au renforcement direct de la parole dans le bruit et, plus particulièrement, à l'adaptation et à l'approfondissement d'une approche par maximisation d'un critère d'intelligibilité, le SII, sous

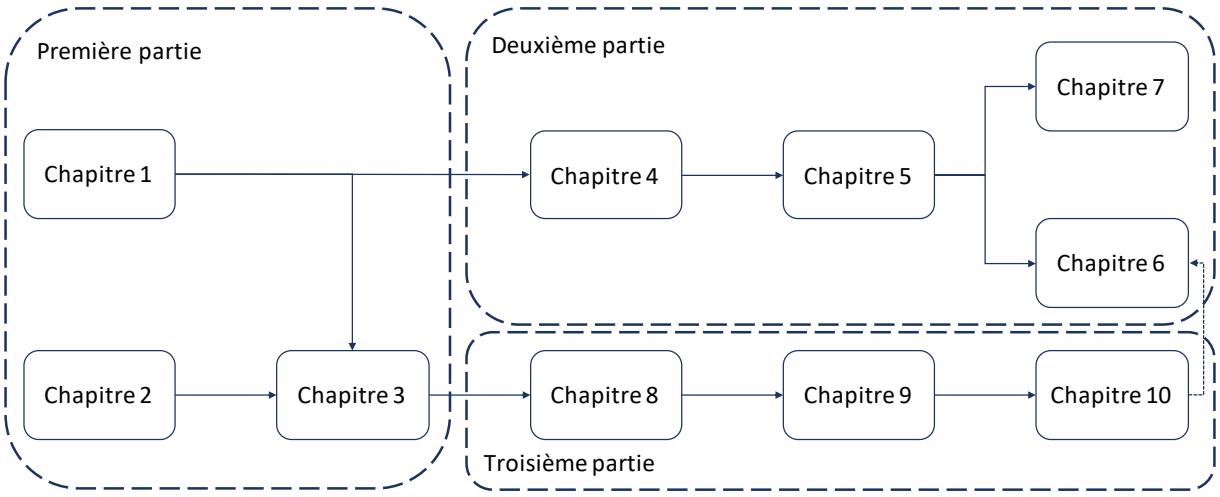


FIGURE 1 – Structure du document.

une nouvelle contrainte perceptive.

Le chapitre 4 présente les limites de la contrainte énergétique moyenne, classiquement utilisée en renforcement de la parole, pour certains cadres applicatifs comme l'environnement automobile : lorsqu'un auditeur choisi un niveau de présentation confortable, une conservation de l'énergie moyenne ne contraint pas le niveau perçu et une augmentation de ce dernier peut être gênant pour l'expérience utilisateur. En se basant sur les notions perceptives introduites chapitre 1, l'utilisation d'une pondération en dBA en guise de nouvelle contrainte énergétique est donc proposée. La suite du chapitre détaille alors les principales approches actuelles de renforcement direct de la parole dans le bruit avec une analyse au regard de la nouvelle contrainte perceptive proposée. Les méthodes de renforcement par maximisation du SII sont des approches très appréciées pour leurs performances mais, en redistribuant l'énergie spectrale dans les zones sensibles de l'oreille, nous remarquons qu'elles sont aussi les plus exposées à la nouvelle contrainte. De plus, toutes les approches actuelles sont basées sur des approximations du critère pour simplifier la résolution du problème d'optimisation. Les observations établies au cours de ce chapitre sont alors les fondements de notre motivation visant à approfondir les approches par maximisation du SII en proposant une résolution exacte du problème d'optimisation et en étudiant leurs performances face à la nouvelle contrainte perceptive.

Le chapitre 5 pose explicitement le problème d'optimisation visant à maximiser le SII et en propose une résolution exacte. Une analyse des résultats obtenus sur des bruits classiques (rose, blanc et conversation) est alors menée permettant d'étudier l'influence théorique de la nouvelle contrainte perceptive par rapport à la contrainte énergétique moyenne. Une adaptation des approches existantes, basées sur des approximations du SII, puis une comparaison de leurs résultats avec les solutions exactes, permettent d'analyser l'efficacité de ces solutions approximatives. Les observations ainsi obtenues nous amèneront à proposer une extension quasi-optimale avec un nombre d'opérations significativement inférieur à celui nécessaire pour la résolution exacte du problème.

Le chapitre 6 présente et analyse les résultats que nous avons obtenus lors de la seconde édition du challenge *Hurricane* proposant une comparaison à grande échelle d'une multitude d'algorithmes actuels de renforcement de la parole dans un bruit de conversation. Nous avons soumis trois participations à savoir : une basée sur la maximisation exacte du SII, une basée sur de la conversion de parole neutre vers parole Lombard (présentée chapitre 10) et une couplant les deux approches. Le système de conversion de parole neutre vers parole Lombard présente des artefacts très audibles et engendre une dégradation de l'intelligibilité, c'est pourquoi l'analyse se concentre exclusivement sur la méthode de maximisation du SII et cela justifie ainsi la place de ce chapitre dans la deuxième partie. Cette méthode propose des performances relativement équivalentes aux autres algorithmes et obtient même le meilleur gain d'intelligibilité dans une condition donnée. En revanche, les tests d'intelligibilité étant basés sur une contrainte énergétique moyenne, nous n'avons pas pu étudier l'effet de la nouvelle contrainte perceptive.

Le chapitre 7 propose alors d'étudier l'effet de la nouvelle contrainte perceptive sur l'approche de maximisation exacte du SII dans trois bruits différents d'habitacle automobile. Une étude objective sur les trois bruits habitacles, similaire à celle effectuée sur les trois bruits classiques dans le chapitre 5, est d'abord présentée. Cette étude permet d'observer et interpréter le comportement de la méthode, ainsi que de comparer l'influence des deux contraintes énergétiques, dans trois nouveaux cas pratiques. De plus, nous obtenons une nouvelle fois des résultats quasi-optimaux avec l'extension proposée chapitre 5 permettant de valider ses performances dans trois nouveaux bruits. Enfin, la mise en place rigoureuse de tests perceptifs et l'analyse des résultats obtenus permet de montrer que la méthode de maximisation du SII, soumise à la nouvelle contrainte perceptive, conserve de très bonnes performances dans des bruits au spectre localisé. En revanche, nous observons que les gains d'intelligibilité ne sont plus significatifs pour des bruits au spectre plus étalé.

Troisième partie :

La troisième partie est consacrée au renforcement paramétrique de la parole dans le bruit, elle se concentre sur les approches par conversion du style de parole en cherchant à améliorer le traitement des aspects temporels.

Le chapitre 8 présente les différentes approches actuelles de renforcement paramétrique de la parole dans le bruit. L'étude des différentes approches met en évidence de nombreuses limitations provoquées par un manque de prise en compte du contexte des phonèmes et du traitement trop simpliste des aspects temporels. Les approches de renforcement par conversion du style de parole, relativement nouvelles, basées sur de l'apprentissage automatique améliore grandement la prise en compte du contexte phonétique. En revanche, nous remarquons que les approches actuelles négligent toujours grandement les aspects temporels que ce soit par l'absence de prise en compte du contexte temporel à grande échelle, ou encore par l'omission des modifications du débit de parole dans les modèles d'apprentissage.

Le chapitre 9 s'oriente alors sur des propositions d'amélioration des aspects temporels dans les méthodes de renforcement par conversion de la parole dans le bruit. Les deux premières améliorations proposées portent sur une adaptation des modèles de conversion et des caractéristiques acoustiques converties. Concernant les modèles de conversion, nous proposons l'utilisation d'architectures récurrentes de réseaux de neurones artificiels bien plus adaptées au traitement des séquences temporelles que les modèles classiquement utilisés (mélange gaussien ou réseau à propagation avant). Pour les caractéristiques acoustiques converties, nous proposons d'appliquer une transformée en ondelette continue afin de représenter l'évolution des caractéristiques sur différentes échelles temporelles. La dernière amélioration proposée porte sur une modélisation nouvelle des modifications du débit qui peut être intégrée dans le modèle d'apprentissage comme caractéristique à convertir.

Finalement, le chapitre 10 décrit la mise en place d'un système de conversion de la parole neutre vers la parole Lombard en introduisant les améliorations des aspects temporels proposées dans le chapitre 9. De nombreuses architectures ont été entraînées et une analyse des performances objectives d'apprentissage permet d'étudier l'intérêt des améliorations introduites. Les architectures récurrentes améliorent légèrement l'apprentissage pour l'ensemble des caractéristiques et la transformée en ondelette permet un apprentissage bien plus performant pour la fréquence fondamentale. De plus, l'intégration de la nouvelle modélisation des modifications du débit de parole dans la fonction de conversion procure un meilleur apprentissage que les méthodes empiriques actuelles pour les segments non-voisés et équivalent pour les segments voisés. Malgré des résultats objectifs très encourageants, les signaux convertis laissent entendre d'importants artefacts de synthèse qu'il faudra impérativement traiter avant d'envisager un quelconque gain d'intelligibilité dans le bruit.

Première partie

Acquis et connaissances portant sur la parole et son intelligibilité

Chapitre 1

Perception de la parole et mesures de son intelligibilité

Sommaire

Retour à la table des matières

Introduction du chapitre 1

La parole est une succession d'ondes acoustiques variant rapidement en intensité et en fréquence, on peut associer l'intelligibilité de la parole à la capacité qu'a un auditeur à reconnaître et interpréter ces évènements acoustiques, qui est influencée par une multitude de facteurs personnels et environnementaux. BENESTY et al. [19] proposent une définition plus formelle qui caractérise l'intelligibilité de la parole comme étant le "taux de répétition correct d'un stimulus verbal par un ou des auditeurs pour un test d'intelligibilité donné" [1]. L'intelligibilité peut alors être mesurée, que ce soit par des tests subjectifs qui permettent une mesure de l'intelligibilité en conditions réelles, ou par des mesures objectives qui en proposent une estimation mathématique.

Dans ce chapitre, nous nous intéressons à la perception de la parole, aux différents facteurs qui influencent son intelligibilité et aux mesures disponibles permettant de la mesurer. Nous commencerons par introduire, section 1.1, d'importants aspects psychoacoustiques mis en jeu lors d'une écoute. Puis les principaux facteurs influençant l'intelligibilité de la parole seront présentés section 1.2. Enfin, les méthodes de mesures d'intelligibilité seront détaillées, d'abord les mesures subjectives dans la section 1.3, puis les mesures objectives dans la section 1.4.

1.1 Aspects psychoacoustiques

Pour comprendre les facteurs qui influencent la capacité à reconnaître et interpréter les évènements acoustiques qui composent la parole, il est important de comprendre les aspects psychoacoustiques qui sont mis en jeu lors de son écoute. Sans rentrer dans les détails du fonctionnement du système oreille-cerveau auditif qui sort de notre domaine d'étude, les aspects psychoacoustiques principaux qui conditionnent une écoute de signaux de parole sont abordés dans cette section.

1.1.1 Audibilité

Le système auditif permet d'interpréter des sons entre 20 Hz et 20 kHz environ mais dans cet intervalle la sensibilité auditive n'est pas uniforme. Le seuil d'audibilité d'un auditeur à une fréquence donnée correspond au niveau minimal pour qu'un son pur à cette fréquence produise une sensation auditive dans un environnement silencieux. Un exemple de courbe typique de seuils d'audibilité pour un normo-entendant est visible figure 1.1 (0 phone). Ces différences fréquentielles viennent d'un filtrage mécanique du système auditif qui engendre une amplification sélective, maximale dans les fréquences intermédiaires. Ces variations de sensibilité provoquent aussi une fluctuation de la perception du volume sonore en fonction du contenu fréquentiel du son. Ce volume perçu, ou sonie, est principalement déterminé par le niveau de pression acoustique, ou *sound pressure level* (SPL), mais il dépend également d'autres paramètres acoustiques, comme la structure spectrale, la durée du signal ou encore la présence d'un son masquant. Nous pouvons aussi observer figure 1.1, des lignes iso-soniques qui représentent les sons purs produisant la même sensation d'intensité en fonction de la fréquence dans le silence. L'unité utilisée pour une courbe iso-sonique est le phone correspondant au niveau sonore exprimé en dB SPL à 1 kHz.

Afin d'estimer la sonie de sons complexes, plusieurs pondérations en fréquences ont été proposées à appliquer aux mesures de pression acoustique brutes avec :

- la pondération A, qui est basée sur la courbe iso-sonique de 40 phones,
- la pondération B, qui est basée sur la courbe iso-sonique de 70 phones,
- la pondération C, qui est basée sur la courbe iso-sonique de 100 phones,
- la pondération Z, qui correspond à une absence de pondération, mais avec une limitation de la bande passante à l'intérieur du domaine audible.

1. citation originale : "*speech intelligibility is the proportion of speech items (e.g., syllables, words, or sentences) correctly repeated by (a) listener(s) for a given speech intelligibility test*"

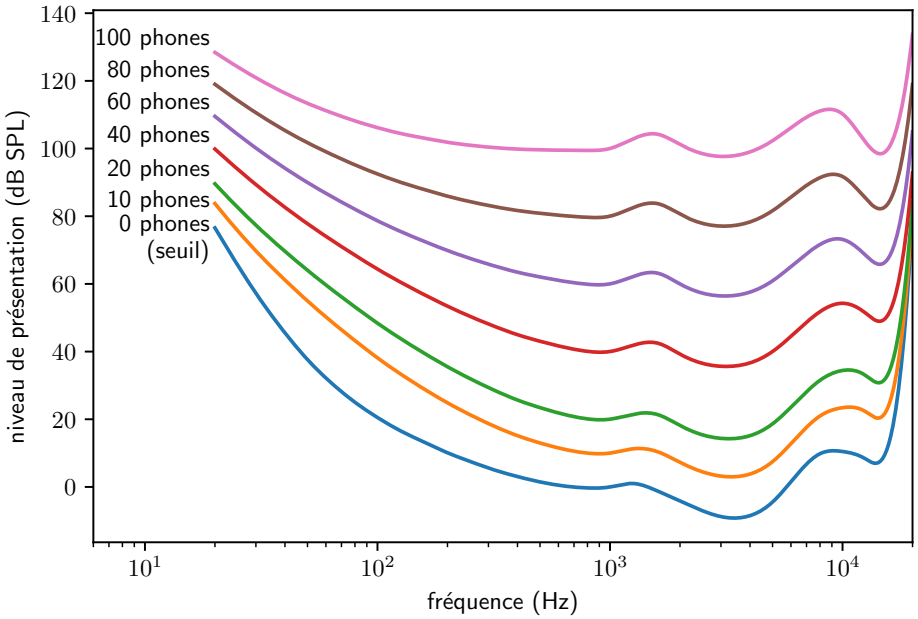


FIGURE 1.1 – Courbes iso-soniques (norme ISO 226-2003).

Les gains associés à ces pondérations sont visibles figure 1.2. Bien que le choix de la pondération utilisée pourrait être dépendant du niveau de présentation des signaux audio, la pondération A est souvent choisie par défaut. Elle est mandatée dans le standard international IEC 61672 pour être introduite dans les appareils de mesure du niveau sonore et elle est adoptée dans la majorité des applications mesurant la fatigue auditive ou les séquelles que peuvent provoquer des sons de différentes intensités, allant même jusqu'aux bruits d'avion. Il est important de noter que ces pondérations proposent seulement une estimation simplificatrice de la sonie réelle car de nombreux facteurs acoustiques l'influençant ne sont pas pris en compte dans le calcul.

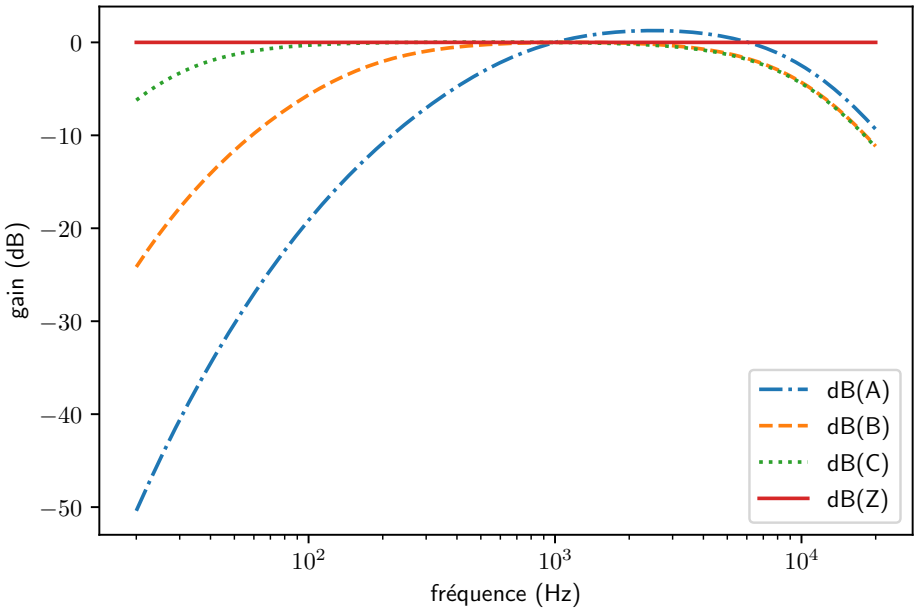


FIGURE 1.2 – Différentes pondérations.

1.1.2 Sélectivité fréquentielle et filtres auditifs

La sélectivité fréquentielle correspond à la faculté auditive de distinguer les composantes sinusoïdales d'un signal complexe et joue un rôle majeur dans la perception auditive que ce soit dans la perception du niveau sonore, du timbre ou de la hauteur d'un son. Cette sélectivité fréquentielle peut être modélisée en considérant le système auditif périphérique comme un banc de filtres dont les bandes passantes se recouvrent continûment : les filtres auditifs.

Caractérisation des filtres auditifs

Plusieurs méthodes de caractérisation de ces filtres à partir de mesures psychoacoustiques ont été proposées. Les filtres auditifs présentent une forme quasi-symétrique pour des niveaux d'excitation modérés, qui devient asymétrique à des niveaux plus élevés, avec un aplatissement du côté des basses fréquences. De plus, leur bande passante augmente avec leur fréquence centrale, ils sont donc plus sélectifs dans les basses fréquences où la capacité auditive à séparer deux composantes sera donc meilleure.

ZWICKER et al. ont étudié les variations de sonie avec la largeur de bande d'un signal en couplant des bruits à bande étroite avec des sondes tonales [222]. Les largeurs de bande des filtres auditifs ainsi estimées sont dénommées bandes critiques et, sur la base de ces résultats, une échelle de fréquence proportionnelle à la largeur de ces bandes a été proposée. L'unité de cette échelle est le Bark et elle divise la plage fréquentielle de l'audible en 24 bandes critiques dont les fréquences centrales et largeurs de bande sont indiquées dans le tableau 1.1.

Bark	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
centre (Hz)	50	150	250	350	455	570	700	845	1000	1175	1275	1600
borne (Hz)	0	100	200	300	400	510	630	770	920	1080	1270	1480	1720
largeur (Hz)	100	100	100	100	110	120	140	150	160	190	210	240

Bark	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
centre (Hz)	1860	2160	2510	2925	3425	4050	4850	5850	7050	8600	10750	13750
borne (Hz)	1720	2000	2320	2700	3150	3700	4400	5300	6400	7700	9500	12000	15500
largeur (Hz)	280	320	380	450	550	700	900	1100	1300	1800	2500	3500

TABLEAU 1.1 – Fréquences caractéristique de la décomposition en bandes critiques.

MOORE et al. ont ensuite révisé cette représentation en mesurant les largeurs de bande en couplant des bruits à échancrure avec des sondes tonales afin d'éviter tout phénomène de battement entre les bruits et les sondes [130]. L'échelle *Equivalent Rectangular Bandwidth* (ERB) correspond alors à la bande passante d'un filtre rectangulaire qui possède le même gain central que le filtre d'intérêt et laisse passer la même quantité d'énergie pour un bruit blanc. Voici un exemple d'équation décrivant les valeurs de l'échelle obtenue sur de jeunes auditeurs normo-entendants à des niveaux modérés, notée $\mathrm{ERB_N}$, en fonction des fréquences centrales F en Hz :

$$\mathrm{ERB_N}(\mathrm{F}) = 24,7\left(\frac{4,37}{1000}\mathrm{F}+1\right). \tag{1.1}$$

En plus de proposer une meilleure représentation des filtres auditifs, l'échelle ERB est définie analytiquement ce qui lui procure une évolution continue des fréquences caractéristiques se rapprochant ainsi du fonctionnement physiologique réel de l'oreille.

Pour simplifier les calculs, il est aussi courant d'utiliser une décomposition spectrale en bandes de tiers d'octaves qui se rapproche de l'analyse fréquentielle opérée par les filtres auditifs. La figure 1.3 présente une comparaison des largeurs de bandes en fonction de la fréquence centrale pour les différentes décompositions spectrales. Les bandes de tiers d'octaves présentent effectivement une bonne approximation de l'échelle ERB et plus particulièrement à partir de 100 Hz.

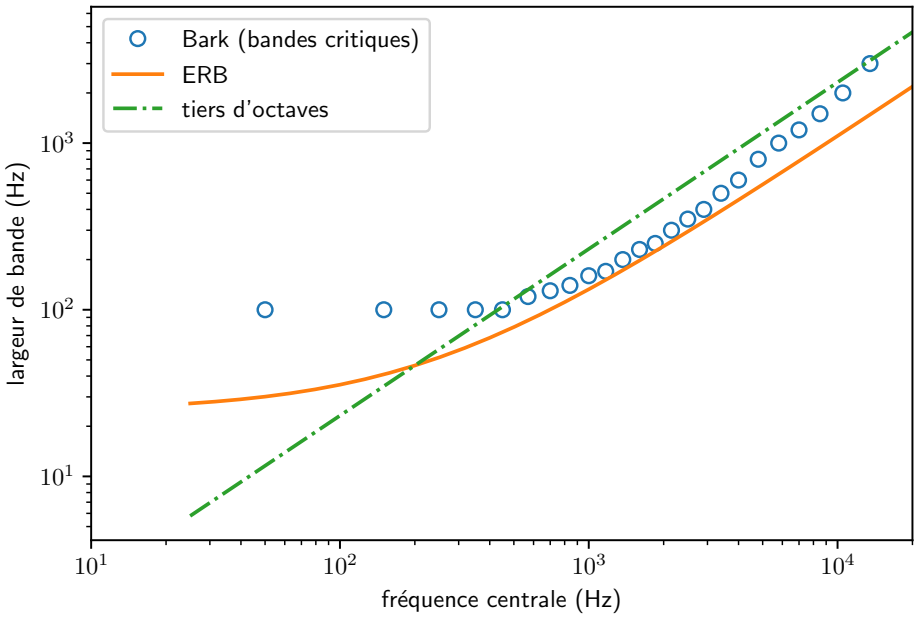


FIGURE 1.3 – Comparaison des largeurs de bandes en fonction de la fréquence centrale pour différentes décompositions spectrales : bandes critiques (Bark), ERB et bandes de tiers d'octaves.

Conséquences des filtres auditifs sur la perception de la parole

Le fonctionnement des filtres auditifs montre alors la capacité du système auditif à analyser de manière indépendante les composantes fréquentielles qui constituent les sons. Ainsi, l'information véhiculée dans les canaux fréquentiels de dizaines de filtres non recouvrants, présents sur la plage fréquentielle utile à la parole (entre 80 Hz et 8 kHz), permettent d'identifier les indices acoustiques des signaux de parole afin de les interpréter efficacement.

Le fonctionnement des filtres explique aussi les difficultés à détecter un son cible en présence d'un son concurrent de fréquence proche. On parle de masquage énergétique et cet effet est maximal lorsqu'une composante concurrente se situe dans la bande passante du filtre accordé sur la fréquence du son cible. Notons d'ailleurs que l'asymétrie des filtres auditifs engendre un étalement du masquage des basses vers les hautes fréquences, ainsi un son cible sera plus facilement masqué par un son de fréquence inférieure à la sienne.

Un auditeur atteint de déficience auditive liée à une détérioration de la cochlée présentera généralement des filtres auditifs plus larges que la normale entraînant alors une diminution de la sélectivité fréquentielle. Cela a pour conséquence directe de brouiller l'information dans les canaux fréquentiels voisins, complexifiant l'identification de certains indices acoustiques de la parole comme le timbre des voyelles. De plus, l'étalement de la bande passante des filtres auditifs augmente le risque de masquage énergétique et rend la présence de sons concurrents encore plus problématique.

1.2 Facteurs influençant l'intelligibilité de la parole

Maintenant que les aspects psychoacoustiques principalement responsables de la perception de la parole ont été introduits, intéressons nous aux différents facteurs conditionnant son intelligibilité. Il est important de noter que les études actuelles cherchant à identifier ces facteurs ne sont pas toujours conclusives et il arrive même qu'elles se contredisent. L'idée de cette section est donc d'introduire quelques variables qui semblent avoir un impact sur l'intelligibilité de la parole, avérées ou non, afin de bien maîtriser ces paramètres lors de son étude.

1.2.1 Facteurs liés au locuteur

Lorsqu'un auditeur écoute des stimuli verbaux, les premiers facteurs conditionnant l'intelligibilité de la parole sont les facteurs liés au locuteur. Ces facteurs peuvent être regroupés sur deux échelles : les caractéristiques générales et les caractéristiques individuelles.

Caractéristiques générales

Une majorité d'études trouvent que les femmes sont globalement plus intelligibles que les hommes [27, 54, 76, 219] mais les causes de ces différences ne sont pas encore clairement identifiées.

Une caractéristique générale spécifique qui pourrait expliquer ces différences est la fréquence fondamentale. Le fait que les femmes possèdent, en moyenne, un fondamental plus élevé est une première piste mais aucune corrélation significative n'a été relevée entre la valeur moyenne du fondamental et l'intelligibilité de la parole naturelle [22, 27, 76, 121]. Une deuxième piste porte sur la variance du fondamental, qui est aussi plus importante chez les femmes, pour laquelle une corrélation significative a cette fois été relevée [27]. Une plus grande dynamique de la fréquence fondamentale chez un locuteur pourrait alors entraîner une meilleure intelligibilité de la parole mais d'autres résultats nuancent ces observations et indiquent que d'autres phénomènes, comme ceux détaillés ci-après, seraient responsables des différences d'intelligibilité [121].

Le débit de parole est une autre piste générale qui pourrait expliquer les différences d'intelligibilité entre les genres. En effet, le débit est significativement plus lent chez les femmes [31] ce qui pourrait expliquer une meilleure intelligibilité. Et pourtant, la corrélation entre la durée des stimuli et l'intelligibilité de la parole naturelle n'est pas [27], ou peu [76], significative.

La maîtrise linguistique du locuteur est aussi une caractéristique générale conditionnant grandement l'intelligibilité de la parole. On comprend facilement qu'un locuteur avec une maîtrise imparfaite de la langue considérée sera moins intelligible qu'un natif, en revanche même une maîtrise approfondie de la langue chez les bilingues engendre des différences significatives de l'intelligibilité de la parole [25] dû à des interactions entre les structures des langues co-existantes.

Si le locuteur est atteint de déficience auditive [128] ou de trouble cognitif [119], l'intelligibilité de sa parole peut alors être altérée mais de façon très variable en fonction de la pathologie. Dans le cas d'une déficience auditive, le port d'une prothèse peut aussi influencer l'intelligibilité du locuteur [128]. Notons aussi que les stimuli verbaux venant d'une voix de synthèse sont généralement moins intelligibles que la parole naturelle [13]. Finalement, l'utilisation d'algorithmes d'analyse-modifications-synthèse est susceptible d'introduire des artefacts qui diminueront aussi l'intelligibilité des signaux originaux.

Caractéristiques individuelles

Les caractéristiques individuelles acoustiques et phonatoires du locuteur peuvent aussi conditionner l'intelligibilité de la parole. L'espace vocalique, qui sera rappelé chapitre 2, en est un exemple avec une corrélation significative entre sa taille et l'intelligibilité de la parole naturelle [22, 27]. Une analyse encore plus fine permet de mettre en évidence des facteurs liés à des prononciations, ou des affiliations de syllabes, spécifiques à certains locuteurs qui influencent l'intelligibilité localement [146, 27].

1.2.2 Facteurs liés à l'écoute

Au delà des facteurs liés au locuteur, de nombreux facteurs liés à l'écoute influencent l'intelligibilité de la parole. Les principaux facteurs concernent l'auditeur, le mode de restitution des stimuli et les caractéristiques de l'environnement d'écoute.

Auditeur

L'existence de déficience auditive [96] ou de troubles cognitif [119] influence grandement l'intelligibilité de la parole. La maîtrise linguistique de l'auditeur est aussi une caractéristique générale conditionnant la compréhension de la parole. Un auditeur non-natif avec une maîtrise imparfaite de la langue considérée comprendra difficilement les stimuli verbaux [20] mais il est intéressant de noter qu'il comprendra autant, voir mieux, un locuteur non-natif qu'un locuteur natif. Le vocabulaire utilisé peut aussi être un facteur d'incompréhension si la complexité du vocabulaire n'est pas adaptée à l'auditeur en fonction de sa catégorie socio-professionnelle par exemple [201].

Restitution

Le mode de restitution est aussi un facteur important pour l'intelligibilité de la parole. L'écoute par des haut-parleurs n'a pas la même intelligibilité que lors d'une écoute directe. Le type de haut-parleurs utilisé est aussi à prendre en compte : un casque présente une plus faible variabilité de l'intelligibilité mais des enceintes sont plus adaptées aux porteurs d'aide auditive [201]. Le volume de présentation, la qualité de restitution et les potentielles distorsion introduites par le matériel sont aussi autant de facteurs influençant l'intelligibilité de la parole pour l'auditeur.

Environnement d'écoute

Enfin, le facteur décisif influençant l'intelligibilité de la parole est l'environnement d'écoute et tout ce qui le caractérise avec la présence :

- de réverbération, lors d'une écoute dans un environnement réverbérant,
- de masquage énergétique, lors d'une écoute avec des signaux audio concurrents comme du bruit,
- de masquage informationnel, lors d'une écoute avec des signaux de parole concurrents,
- d'une charge cognitive supplémentaire, avec l'existence d'une tâche à résoudre autre que l'écoute des stimuli verbaux.

1.3 Mesures subjectives de l'intelligibilité de la parole

Si l'intelligibilité de la parole s'apparente à un taux de répétition correct, aussi appelé score, par des auditeurs, celle-ci ne peut être mesurée avec fiabilité qu'à partir de tests subjectifs. Au fil du temps, une multitude de tests d'intelligibilité ont été mis en place, principalement pour évaluer des déficiences auditives. Ces tests consistent à faire écouter des stimuli verbaux aux sujets qui doivent ensuite les répéter. Lors de la mise en place d'un test d'intelligibilité, il est primordial de prendre en compte de nombreux de paramètres, dont les facteurs introduits précédemment, et le but de cette section est de faire une synthèse de ces paramètres et de leur conséquences sur la mesure de l'intelligibilité de la parole dans le bruit.

1.3.1 Différents types de tests

En fonction de l'objectif du test, il peut être effectué dans le silence [180, 18, 120], ou en présence de bruit [150, 32, 97, 217]. D'après notre contexte, nous nous intéressons surtout aux tests en présence de bruit et SHARMA et al. [177] dressent une comparaison récente de ces principaux tests. La majorité des tests utilisent des phrases comme stimuli verbaux et ils sont appelés *Sentence Recognition in Noise* (SRN) *tests* dans la littérature anglophone, en opposition à l'utilisation de mots dans les *Word Recognition in Noise* (WRN) *tests*. Par exemple le test *Words In Noise* (WIN) [217] a choisi comme stimuli verbaux des mots mono-syllabiques car dans certains cas la répétition de phrases peut ne pas être adaptée, en présence de troubles de la mémoire par exemple. Dans les tests SRN, les phrases sont composées de mots clefs qui conditionneront le succès ou non d'une répétition.

Dans notre étude, afin de nous rapprocher d'une situation d'écoute réelle, nous nous concentrerons sur les tests SRN. Les informations détaillées dans la suite de cette section sont tout de même, en grande partie, applicables aux tests WRN.

1.3.2 Stimuli et score

Dans les tests SRN, il y a deux types de stimuli : les stimuli verbaux que le sujet doit répéter et les stimuli de bruit qui viennent perturber l'écoute. Le score correspond alors au taux de répétition des stimuli verbaux.

Stimuli verbaux

Les stimuli verbaux utilisés sont prononcés à partir d'un corpus de phrases par un locuteur unique afin de fixer les différents facteurs liés au locuteur influençant l'intelligibilité de la parole. Ce corpus doit être méticuleusement préparé pour que les phrases aient un vocabulaire adapté à la population de sujets ciblée et des mots clefs dont l'emplacement doit être cohérent d'une phrase sur l'autre. Pour que des résultats puissent être comparés, il faut que ces stimuli soient identiques d'un sujet à l'autre, c'est pourquoi il est important qu'ils soient pré-enregistrés.

Stimuli de bruit

Les paramètres spectro-temporels du bruit peuvent avoir des effets de masquage énergétique très diversifiés sur l'intelligibilité des signaux de parole, il est donc primordial de choisir un ou plusieurs stimuli de bruit parfaitement adaptés aux objectifs visés.

Les bruits stationnaires sont très utilisés car ils permettent un meilleur contrôle et une réplicabilité accrues des conditions d'écoute. Ils peuvent être issus d'enregistrements naturels ou synthétisés à partir d'un bruit blanc filtré afin de modeler le spectre selon les besoins. Lorsqu'il est synthétisé, le spectre visé doit refléter les situations d'écoute que l'on cherche à étudier, ainsi un type de bruit stationnaire largement utilisé est un bruit blanc de même enveloppe spectrale qu'un bruit de conversation, il est usuellement abrégé par l'acronyme anglais pour *Speech Shaped Noise* (SSN). Une bonne pratique consiste à générer le bruit SSN à partir du spectre moyen des stimuli verbaux proposant ainsi un RSB à peu près constant dans toutes les bandes de fréquence.

Les bruits non-stationnaires reflètent mieux des conditions d'écoute réelles mais ajoute une complexité importante aux tests d'intelligibilité. En effet, leur caractère variable rajoute un facteur difficilement maîtrisable qui peut grandement influencer les écoutes et ces types de bruit nécessiteront généralement plus de passages pour interpréter correctement les résultats. Les bruits fluctuants naturels les plus couramment utilisés sont les bruits de conversation, avec une ou plusieurs voix concurrentes. Il est aussi courant de synthétiser des bruits fluctuants en modulant l'amplitude d'un bruit stationnaire par une enveloppe donnée. Lorsque l'amplitude d'un bruit SSN est modulée par une enveloppe temporelle similaire à la parole, on le désigne par l'acronyme anglais pour *Speech Modulated Noise* (SMN).

De la même façon que pour les stimuli de parole, le stimulus de bruit doit être identique si l'on souhaite pouvoir comparer les résultats.

Score

Enfin le score peut se calculer de plusieurs façons différentes :

- un score à la phrase considère une répétition comme fausse si au moins un des mots clefs est mal répété, ce type de score est donc binaire,
- un score au mot introduit une nuance car il correspond au taux de mots clefs correctement répétés, cette approche est plus précise et c'est souvent cette solution qui est choisie,
- un score à la syllabe, encore plus précis car il correspond au taux de syllabes correctement répétées, il est plutôt utilisé au cours du développement des tests ou pour des langues aux écritures conjonctives.

1.3.3 Méthode de présentation

La méthode d'un test concerne la méthodologie utilisée pour présenter les stimuli verbaux à l'auditeur. On ne parle pas ici du moyen de restitution, qui est une variable importante abordée dans la section 1.3.4, mais plutôt du choix des niveaux auxquels sont présentés les stimuli verbaux. On différencie alors les méthodes fixes, dont les niveaux sont fixés à l'avance, des méthodes adaptatives, dont les niveaux évoluent en fonction des réponses du sujet.

Dans tous les cas il est aussi conseillé d'effectuer une étape préliminaire d'équilibrage des stimuli verbaux. L'équilibrage consiste à vérifier que les stimuli ont tous la même difficulté à être répétés dans le bruit considéré et à ajuster leurs niveaux si ce n'est pas le cas.

Méthodes fixes

Les méthodes fixes présentent les stimuli verbaux à des niveaux fixés à l'avance de sorte d'être en limite d'intelligibilité. Les scores moyens ainsi obtenus aux différents niveaux peuvent être comparés entre eux ou à une référence en fonction des besoins du test. Il est aussi intéressant d'estimer une courbe psychométrique à partir des scores obtenus aux différents niveaux permettant de visualiser l'évolution théorique de l'intelligibilité en fonction du niveau de présentation. Un exemple d'estimation de courbe psychométrique à partir des scores moyens obtenus à des RSB fixés est visible figure 1.4.

Les méthodes fixes ont l'avantage que les stimuli verbaux peuvent être générés en avance car les niveaux de présentation sont connus dès le début, ce qui facilite grandement la mise en place du test. C'est donc cette solution qui est souvent choisie pour sa praticité mais elle possède d'importants inconvénients. En effet, ces méthodes sont soumises à des effets de seuil, surtout si les niveaux n'ont pas été correctement choisis ou si il existe une grande variabilité au sein de la population. La précision de ces approches est aussi discutable car on ne maîtrise pas où l'on se trouve sur la courbe psychométrique. La mesure d'intelligibilité en des points où la pente est faible nécessitera beaucoup de sujets pour avoir une précision acceptable et pour pouvoir comparer des résultats entre eux.

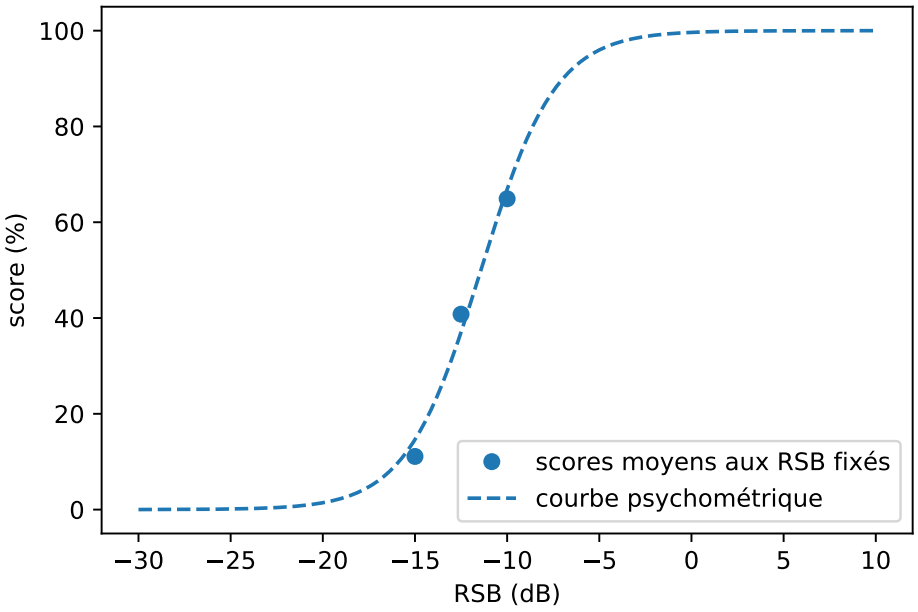


FIGURE 1.4 – Courbe psychométrique estimée à partir des scores moyens obtenus à des RSB fixés.

Méthodes adaptatives

Les méthodes adaptatives sont basées sur l'estimation d'un niveau pour lequel l'auditeur parvient à répéter X% des stimuli verbaux, X étant le taux de répétition, ou score, visé. L'estimation du niveau pour lequel un auditeur parvient à un score de 50% est appelé Seuil de Réception de la Parole (SRP). L'estimation du SRP est la plus utilisée dans la littérature car, comme nous pouvons le voir figure 1.4, il correspond au point d'inflexion des courbes psychométriques, la pente est donc maximale à ce niveau ce qui permet une estimation plus précise. Il arrive tout de même qu'on choisisse d'estimer le niveau pour lequel un auditeur parvient à un score de X% différent de 50, on appellera alors ce niveau le SRPX. Les procédures adaptatives d'estimation du SRP se déroulent de cette façon :

1. une liste composée d'au moins 10 stimuli verbaux est exploitée,
2. l'auditeur tente de répéter chaque stimulus verbal présenté,
3. le niveau de présentation de chaque stimulus verbal dépend de la répétition du stimulus précédent : si l'auditeur répète correctement, le stimulus suivant sera présenté moins fort, et vice-versa,

Il existe plusieurs façons de calculer la modification du niveau d'un stimulus verbal à l'autre. Les procédures adaptatives s'inspire en général de celle proposée par HAGERMAN et al. [73] dont le tableau 1.2 donne les changements de niveau à appliquer aux stimuli.

nombre de mots clefs répétés	0	1	2	3	4	5
changement de niveau (dB)	+4	+2	0	-2	-4	-6

TABLEAU 1.2 – Changements de niveau à appliquer aux stimuli verbaux en fonction du nombre de mots clefs correctement répété au stimulus précédent, pour des phrases composées de cinq mots clefs.

Il existe ensuite plusieurs façon d'estimer le SRP en fonction des résultats du tests :

- en prenant le dernier niveau de présentation i.e. en considérant que la procédure a convergé vers le SRP
- en moyennant les n derniers niveaux de présentations e.g. n=5 pour une liste de 10 stimuli et n=15 pour une liste de 20 stimuli
- en considérant chaque répétition comme une épreuve de Bernoulli dont la probabilité dépend du SRP qui peut donc être estimé en utilisant un estimateur du maximum de vraisemblance sur le processus de Bernoulli résultant (cette procédure permet d'obtenir une estimation du SRP dont l'erreur est inférieure à 1 dB [28])

Équilibrage de la difficulté

L'équilibrage de la difficulté des stimuli verbaux consiste à s'assurer que ces derniers ont la même difficulté à être répétés dans le bruit considéré. Ainsi l'équilibrage permet d'introduire des corrections de niveau de présentation : les stimuli verbaux plus compliqués seront présentés légèrement plus fort alors que les plus simples seront présentés légèrement moins fort. C'est une étape qui demande généralement plusieurs itérations de corrections sur des groupes de personnes différentes [150, 209]. Cependant, la méthode d'équilibrage proposée par Nielsen et al. [148] fait intervenir le jugement subjectif des sujets dans l'équilibrage ce qui permet d'obtenir de meilleurs résultats avec moins de sujets. L'équilibrage dépend grandement du bruit et devrait donc être effectué pour chaque bruit considéré.

1.3.4 Synthèse et autres variables importantes

Il y a aussi une multitude d'autres variables à prendre en compte lors de la mise en place et l'interprétation d'un test subjectif de mesure d'intelligibilité. Le tableau 1.3 dresse une synthèse des variables influençant les tests d'intelligibilité dans le bruit. Ce tableau est inspiré des travaux de THEUNISSEN et al. [201] et on remarque que, parmi les variables qui n'ont pas encore été abordées dans cette section, beaucoup font échos aux facteurs présentés section 1.2.

Variables			Influences
Stimuli	Phrases	Vocabulaire	La complexité du vocabulaire doit être adaptée à la population cible.
		Mots clefs	La position des mots clefs doit être cohérente d'une phrase à l'autre.
		Sens	Les phrases peuvent avoir du sens ou non : le choix dépend donc des objectifs du test.
	Locuteur		Le locuteur peut avoir une influence importante sur les résultats du test. Il est donc conseillé que les stimuli soient pré-enregistrés en utilisant le même locuteur pour faciliter la comparaison des résultats.
	Bruit		Le bruit conseillé par défaut est un bruit SSN mais ce choix dépend surtout des objectifs du test. En revanche, il est fortement recommandé d'utiliser toujours le même bruit pour faciliter la comparaison des résultats.
Présentation	Méthode		Méthode adaptative plus flexible (pas d'effets de seuil) et plus précise mais plus complexe à mettre en place.
	Restitution	Binauralité	Intelligibilité plus importante avec du binaural (séparation spatiale parole/bruit) : cela dépend donc des objectifs du test.
		Haut-parleurs	Le casque a une plus faible variabilité mais les enceintes sont plus adaptées aux porteurs d'aide auditive.
Auditeur	Déficience auditive		Toutes ces variables influencent les résultats des tests. Un contrôle de celles-ci est primordial pour une bonne interprétation des résultats.
	Age		
	Langue		
	Cognition		
Réponse	Score		Préférer au mot, ou à la syllabe pour les cas spéciaux.
	Canal	Oral	Les réponses données par voie orale sont sensibles à une mauvaise écoute de l'expérimentateur. De plus, si seul le score est relevé, les réponses entières ne sont pas archivées contraignant alors l'analyse des résultats.
		Écrit	Les réponses données par écrit sont sensible à une mauvaise lecture de l'expérimentateur et aux fautes de frappe du sujet. De plus, cela peut provoquer des pertes d'attention du sujet.

TABLEAU 1.3 – Synthèse des variables influençant les tests d'intelligibilité dans le bruit

1.4 Mesures objectives de l'intelligibilité de la parole

La mise en place de tests subjectifs requiert une mise en place rigoureuse et fait appel à de nombreux sujets. Des mesures objectives ont alors été créées afin d'estimer l'intelligibilité de la parole à l'aide de critères mathématiques. Les méthodes RSB consistent à calculer une estimation de l'intelligibilité de la parole en présence de masquage énergétique à partir du Rapport Signal sur Bruit (RSB) dans différents canaux fréquentiels. Elles s'opposent aux approches corrélationnelles qui se basent sur les différences spectro-temporelles entre le signal d'origine et le signal dégradé ne nécessitant donc pas de connaître explicitement la dégradation.

1.4.1 SII et extensions

Une mesure RSB de l'intelligibilité très populaire est le SII [7] qui est une révision de l'*Articulation Index* (AI) proposant d'incorporer des procédures spécifiques pour une meilleure précision dans différentes conditions d'utilisation. Le calcul exact du SII sera détaillé chapitre 5 mais il peut être résumé en trois étapes :

1. le calcul des niveaux effectifs moyens de parole et de perturbation dans différents canaux fréquentiels, généralement les bandes de tiers d'octaves ou les bandes critiques, la perturbation prenant en compte le bruit mais aussi la propagation du masquage énergétique vers les hautes fréquences et le seuil d'audibilité de l'auditeur,
2. le calcul de la contribution à l'audibilité de chaque bande à partir de son Rapport Signal sur Perturbation (RSP), la contribution est maximale pour un RSP de +15 dB et minimale pour un RSP de -15 dB, avec une prise en compte d'une potentielle distorsion introduite par des niveaux de parole trop élevés,
3. le calcul d'une somme pondérée des contributions, dont les poids sont déterminés par une Fonction d'Importance de Bande (FIB), donne alors le SII.

Il est important de préciser le terme "mesure d'intelligibilité" couramment utilisé pour le SII. Il serait en fait plus rigoureux d'y faire référence par "mesure d'audibilité" car le SII ne mesure pas directement l'intelligibilité. En effet, un SII de 50% ne signifie pas que l'on obtiendrait un score moyen de 50% lors d'un test d'intelligibilité. Au contraire, un signal de parole possédant un SII de 50% dans un environnement donné sera souvent complètement intelligible pour une population de normo-entendants. En fait, cela signifie que 50% des indices de la parole sont audibles et peuvent être interprétés dans cet environnement d'écoute. En revanche, le SII est corrélé au score d'intelligibilité et des fonctions de transfert empiriques peuvent être calculées, pour un ensemble donné de stimuli verbaux, afin d'obtenir la relation entre les deux. Elles dépendent du contenu syntaxique des stimuli, de la capacité d'expression des locuteurs et de la compétence d'écoute des auditeurs. Des exemples de fonctions de transfert obtenues avec des stimuli verbaux utilisés par trois ensembles de test sont représentées figure 1.5 [78]. Elles ont été obtenues à partir de tests d'écoute avec de nombreux auditeurs normo-entendants sur plusieurs centaines de conditions de masquage. À notre degré d'analyse, les concepts d'audibilité et d'intelligibilité seront confondus dans la suite de notre étude pour une meilleure clarté.

Le SII est une mesure objective très performante en présence de masquage énergétique stationnaire, ou suite à de simples filtrages linéaires, mais de nombreux facteurs ne sont pas pris en compte dans son calcul, ainsi de multiples extensions ont été proposées. Le *Short Term Articulation Index* (AI-ST) [163] propose de calculer le SII sur des fenêtres court-terme afin de prendre en compte les aspects temporels dans un bruit fluctuant. Le *Coherence Speech Intelligibility Index* (CSII) [89] propose de prendre en compte les distorsions harmoniques et les distorsions d'inter-modulation présentes dans les aides auditives en remplaçant les RSP par les rapports signal sur distorsion calculés à partir de la fonction de cohérence entre le signal d'entrée et celui traité par l'aide auditive. En travaillant directement sur le signal dégradé, le CSII n'est donc plus une approche RSB mais une approche corrélationnelle. Le *Hearing-Aid Speech Perception Index* (HASPI) n'est pas une extension du SII mais une mesure corrélationnelle adaptée aux aides auditives aussi basée sur la fonction de cohérence ajoutant l'utilisation d'un modèle du système périphérique auditif pour plus de précision [90].

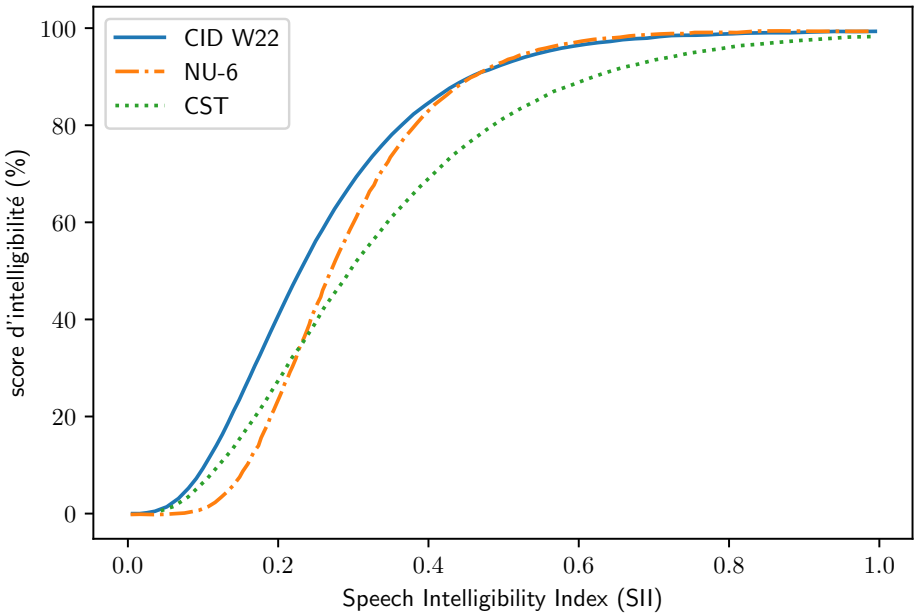


FIGURE 1.5 – Fonctions de transfert estimant la relation entre le SII et le score d'intelligibilité pour trois ensembles de stimuli verbaux.

1.4.2 STI et extensions

Le *Speech Transmission Index* (STI) [184] est une méthode corrélationnelle qui se base sur l'idée que la dégradation de l'intelligibilité introduite par masquage énergétique, ou par réverbération, peut être modélisée par une réduction de la modulation de l'enveloppe temporelle des signaux de parole. Le calcul du STI peut être résumé en quatre étapes :

1. le calcul d'une Fonction de Transfert de Modulation (FTM) pour différents canaux fréquentiels, généralement les bandes d'octaves, la FTM de chaque canal est calculée sur des fréquences de modulations connues pour leur importance vis à vis de l'intelligibilité de la parole (entre 0,63 et 12,7 Hz),
2. le calcul d'un RSB apparent sur chaque canal à partir de sa FTM et moyenné sur les fréquences de modulation,
3. le calcul de la contribution à la transmission de chaque bande à partir de son RSB apparent, ce calcul est identique à celui du SII avec une contribution maximale pour un RSB apparent de +15 dB et minimale pour un RSB apparent de -15 dB,
4. le calcul d'une somme pondérée des contributions, dont les poids sont déterminés par une FIB, donne alors le STI.

De nombreuses extensions ont été proposées, elles diffèrent de la méthode originale par la façon dont les FTM sont calculées et dont les contributions à la transmission sont calculés à partir de ces FTM [70]. Le STI, et ses extensions, permettent une estimation performante de l'intelligibilité de la parole et seront préférées au SII, et ses extensions, en présence de réverbération ou si le bruit masquant seul n'est pas accessible.

1.4.3 Autres méthodes

De nombreux travaux visent à proposer de nouvelles mesures objectives de l'intelligibilité de la parole, ou à améliorer celles déjà existantes. Parmi les travaux les plus marquants nous pouvons citer la proposition d'utiliser des FIB dynamiques rendant les mesures plus robustes en présence d'un bruit masquant fluctuant [118]. Aussi, le *Short-Time Objective Intelligibility* (STOI) qui est une méthode corrélationnelle simple mais dont le calcul porte

sur des segments court-terme des signaux de parole. Ou encore, le *Speech Intelligibility In Bits* (SIIB) qui est aussi une méthode corrélationnelle simple basée sur l'hypothèse que l'intelligibilité est conditionnée par la quantité d'information partagée, entre le signal original et le signal dégradé, en bits par seconde.

Enfin, comme le SII est une mesure d'audibilité souvent utilisée comme une mesure d'intelligibilité, il arrive que l'on utilise aussi d'autres mesures perceptives comme telle. Un exemple notable est le modèle *Glimpse Proportion* (GP), une méthode RSB destinée initialement à localiser les zones spectro-temporelles des signaux de parole qui sont moins affectées par le bruit que d'autres [40]. Bien que ce modèle n'ait pas été construit comme une mesure objective d'intelligibilité, sa sortie est fortement corrélée avec l'intelligibilité des signaux de parole dans un bruit masquant fluctuant. Ainsi de multiples travaux s'en servent comme mesure d'intelligibilité au point que les auteurs travaillent sur une proposition d'extension afin d'adapter spécifiquement le modèle à cette tâche [197].

Conclusion du chapitre 1

Dans ce chapitre, nous avons vu comment un auditeur perçoit la parole ainsi que les nombreux facteurs qui peuvent influencer son intelligibilité. Notre environnement d'étude, que sont les habitacles de voiture, sont soumis à une multitude de ces facteurs (charge cognitive, bruit, potentielle déficience auditive...), c'est pourquoi il est important de développer des méthodes d'amélioration de l'intelligibilité des signaux de parole bien adaptées à notre contexte.

Nous avons ensuite présenté comment l'intelligibilité pouvait être mesurée, d'abord par des tests subjectifs. Ces tests d'intelligibilité sont nécessaires afin d'obtenir une estimation réelle de l'intelligibilité, qui ne peut être obtenue qu'avec un protocole rigoureux de mise en place, de passation et d'analyse des résultats. Le choix du type de test et de la méthodologie doit impérativement dépendre du contexte d'écoute et de la population visée. Les recommandations introduites dans ce chapitre serviront donc de référence pour justifier les nombreux choix pour nos tests perceptifs qui seront conduits chapitre 7.

Enfin, nous avons détaillé comment l'intelligibilité pouvait être estimée par des mesures objectives, très pratiques à utiliser car elles ne nécessitent pas la mise en place de tests perceptifs coûteux. De nombreuses mesures existent et le choix de celle utilisée dans un contexte donné doit alors être justifié par le cadre d'utilisation. Dans notre contexte automobile, pour lequel le bruit est la source principale de détérioration de l'intelligibilité, les signaux de parole intactes ainsi que des estimations du bruit sont facilement accessibles en temps réel. C'est pourquoi le SII semble la mesure d'intelligibilité la plus adaptée à notre environnement d'étude. De plus, comme nous le verrons chapitre 7, les bruits prédominants sont quasi-stationnaires ne nécessitant donc pas l'utilisation d'extensions particulières du SII.

Chapitre 2

Production, analyse et modification numérique de la parole

Sommaire

Retour à la table des matières

Introduction du chapitre 2

Avant de s'intéresser à l'amélioration de l'intelligibilité, il est primordial d'introduire comment la parole est produite, ce qui la caractérise et les méthodes d'analyse à disposition. Ces méthodes d'analyse offrent généralement la possibilité de modifier les paramètres de la parole, puis de re-synthétiser un signal de parole modifié. Ces notions d'analyse-modifications-synthèse des signaux de parole seront exploitées de nombreuses fois au cours de ce manuscrit. Que ce soit pour comprendre les modifications naturelles, ou pour introduire des modifications numériques, visant à améliorer l'intelligibilité de la parole, il est donc crucial de les présenter dès maintenant.

Dans ce chapitre, nous nous intéressons donc à la production de la parole, aux principes des méthodes d'analyse-modifications-synthèse et aux différents éléments qui nous permettront de choisir un vocodeur adapté à notre étude. Les mécanismes de production de la parole, l'analyse court-terme associée et les caractéristiques spectrales résultantes, sont d'abord introduits section 2.1. Nous présenterons ensuite, dans la section 2.2, la base des approches fréquentielles et temporelles qui supportent l'intégralité des approches d'analyse-modifications-synthèse des signaux de parole. Enfin, dans la section 2.3, nous détaillerons les principaux vocodeurs, basés sur le modèle de production de parole, permettant de choisir sur lequel nous nous appuierons pour nos travaux.

2.1 Production et analyse de la parole

2.1.1 Mécanismes de production de la parole

La parole est produite à partir d'une excitation glottique, issue d'un flux d'air (souffle pulmonaire), interagissant avec les organes vocaux supérieurs, appelés cavités supra-glottiques. On distingue trois types d'excitation :

- une forme quasi-périodique, provenant de la modulation du flux d'air par les cordes (lèvres) vocales, elle sera responsable des sons voisés
- une forme aléatoire, provenant d'un écoulement tourbillonnaire du flux d'air, elle sera responsable des sons fricatifs
- une forme transitoire, provenant d'un blocage du flux d'air au niveau d'un organe articulateur suivi d'un relâchement soudain, elle sera responsable des sons occlusifs

Les cavités supra-glottiques, visibles figure 2.1, se composent des cavités pharyngale, buccale, nasale et labiale, et jouent le rôle de résonateur en affaiblissant ou renforçant certaines composantes spectrales de l'excitation glottique.

Lors de la production de la parole, ces différents éléments varient dans le temps c'est pourquoi les signaux de parole sont non-stationnaires par nature. La plupart des outils utilisés en traitement du signal se basent pourtant sur des hypothèses d'invariance temporelle du système et de la source ce qui pose problème quant à leur utilisation sur des signaux de parole.

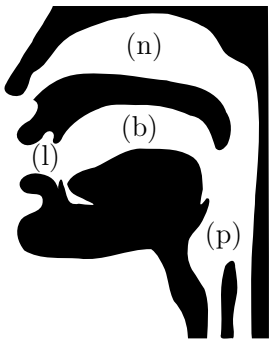
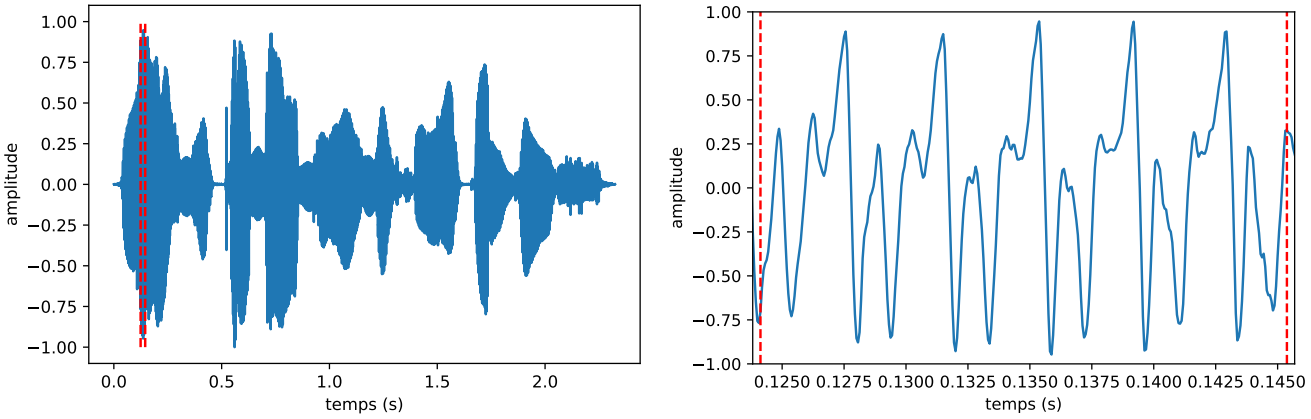


FIGURE 2.1 – Cavités supra-glottiques : (p) pharyngale, (b) buccale, (n) nasale et (l) labiale.

2.1.2 Analyse court-terme de la parole

Une caractéristique intéressante du système de production de la parole est que le flux d'air et les organes articulatoires varient lentement sur une échelle temporelle suffisamment courte. En prenant des segments de quelques dizaines de millisecondes, le système de production de la parole est considéré quasi-stationnaire, comme on peut le voir figure 2.2, et on peut alors utiliser la plupart des outils de traitement du signal stationnaire sur ces segments. Cette démarche d'analyse court-terme est à la base de la majorité des analyses et traitements numériques de la parole. Le signal est multiplié par des fenêtres d'analyse centrées autour d'instants d'analyse et c'est sur ces segments que l'on travaille.

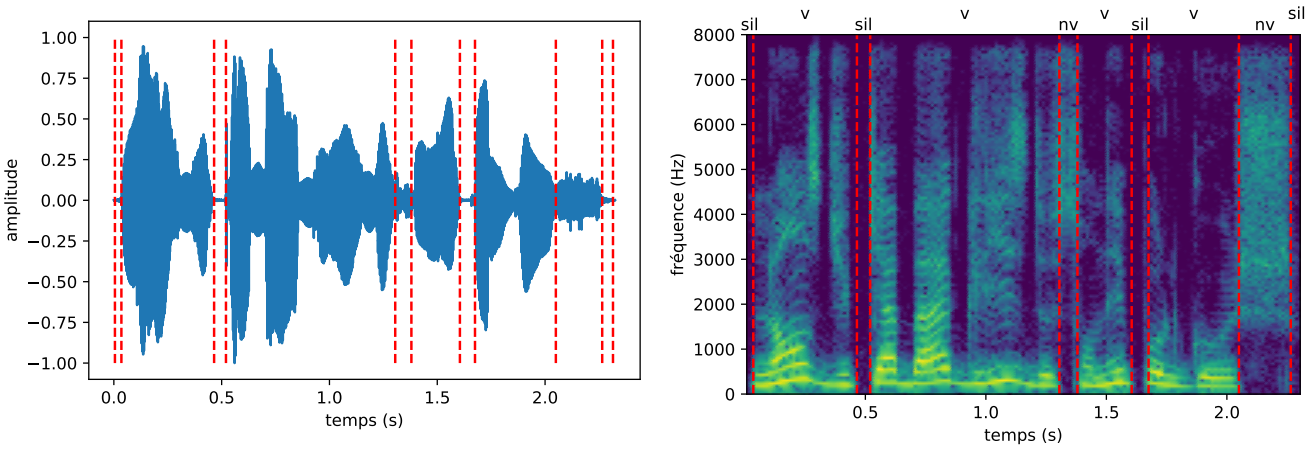
Les approches d'analyse-modifications-synthèse de la parole sont majoritairement basées sur la Transformée de Fourier à Court-Terme (TFCT) qui consiste à appliquer une transformée de Fourier discrète sur les segments d'analyse afin d'obtenir une représentation temps-fréquence du signal sur toute sa durée. Un exemple d'une TFCT est visible figure 2.3



(a) forme d'onde classique d'un signal de parole

(b) zoom sur le segment délimité par les pointillés

FIGURE 2.2 – Zoom sur une forme d'onde d'un signal de parole classique faisant apparaître le comportement quasi-stationnaire du système de production de parole sur une échelle temporelle suffisamment courte.



(a) forme d'onde classique d'un signal de parole

(b) TFCT associée et voisement des segments

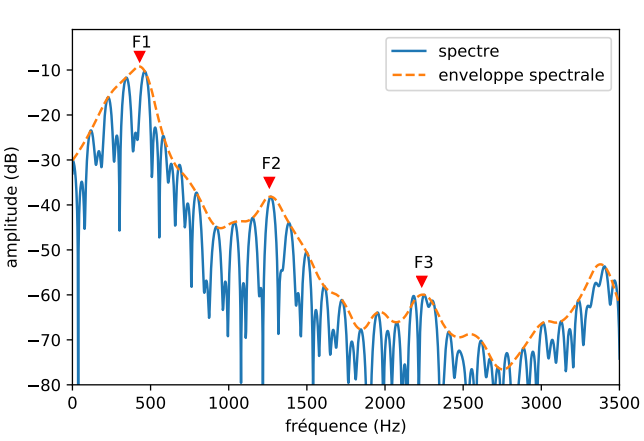
FIGURE 2.3 – Exemple d'une forme d'onde classique d'un signal de parole et de sa TFCT associée. La TFCT fait apparaître la structure harmonique des segments voisés (notés v) et la structure stochastique des segments non-voisés (notés nv). L'absence d'énergie correspond à des silences (notés sil)

2.1.3 Caractéristiques spectrales de la parole

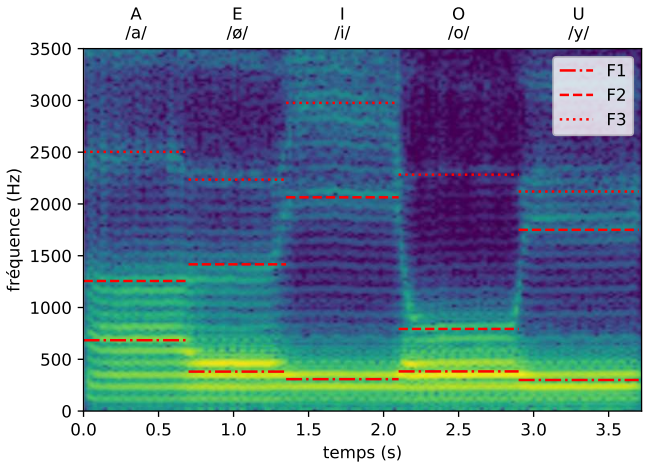
L'analyse spectrale court-terme permet de faire apparaître des caractéristiques de la parole très importantes pour son analyse et sa modification. L'exemple d'un spectre de la voyelle française E, ou /ø/ dans l'Alphabet Phonétique International (API), est visible figure 2.4a. On observe très bien l'interaction entre l'excitation glottique et les cavités supra-glottiques qui produit un spectre par échantillonnage de l'enveloppe des cavités à la fréquence fondamentale des vibrations des cordes vocales, environ 120 Hz sur la figure. La fréquence fondamentale, notée F0, et ses variations caractérisent l'intonation de la parole. De plus, on remarque des maxima d'énergie dans l'enveloppe spectrale, les formants, qui sont des caractéristiques acoustiques majeures dans la perception des phonèmes et en particulier des voyelles. Les fréquences typiques des trois premiers formants pour les voyelles de la langue française sont consultables tableau 2.1 [207] et leur visualisation sur une TFCT est proposée figure 2.4b. Les formants indiqués par les pointillés correspondent à ceux du tableau 2.1 des voyelles correspondantes et on remarque bien que les maxima d'énergie sont proches de ces valeurs. En reportant ces valeurs dans un graphique avec F2 en abscisse et F1 en ordonnée, on obtient l'espace vocalique, souvent représenté par un triangle ou un trapèze, visible figure 2.5. Dans cet espace vocalique, l'ouverture des voyelles se traduit par une augmentation de F1, de /i/ et /u/ vers /a/, et leur profondeur se caractérise par une diminution de F2, de /i/ vers /u/.

	voyelles (API)									
	fermées			mi-fermées			mi-ouvertes			ouverte
	i	y	u	e	ø	o	ɛ	œ	ɔ	a
F_1 (Hz)	308	300	315	365	381	383	530	517	531	684
F_2 (Hz)	2064	1750	764	1961	1417	793	1718	1391	998	1256
F_3 (Hz)	2976	2120	2027	2644	2235	2283	2558	2379	2399	2503

TABLEAU 2.1 – Fréquences typiques des trois premiers formants pour les voyelles de la langue française [207].



(a) spectre, enveloppe spectrale et formants de la voyelle ɛ



(b) TFCT de AEIOU et deux premiers formants associés

FIGURE 2.4 – (a) Exemple d'un spectre de la voyelle française E, ou/ø/dans l'API avec mise en évidence de l'enveloppe spectrale et des trois premiers formants. (b) TFCT d'un signal de parole enchaînant les voyelles françaises AEIOU, ou [aøioy] dans l'API, en maintenant un fondamental constant : les trois premiers formants typiques associés à chaque voyelle, dont les valeurs viennent du tableau 2.1, sont indiqués en pointillés sur la TFCT.

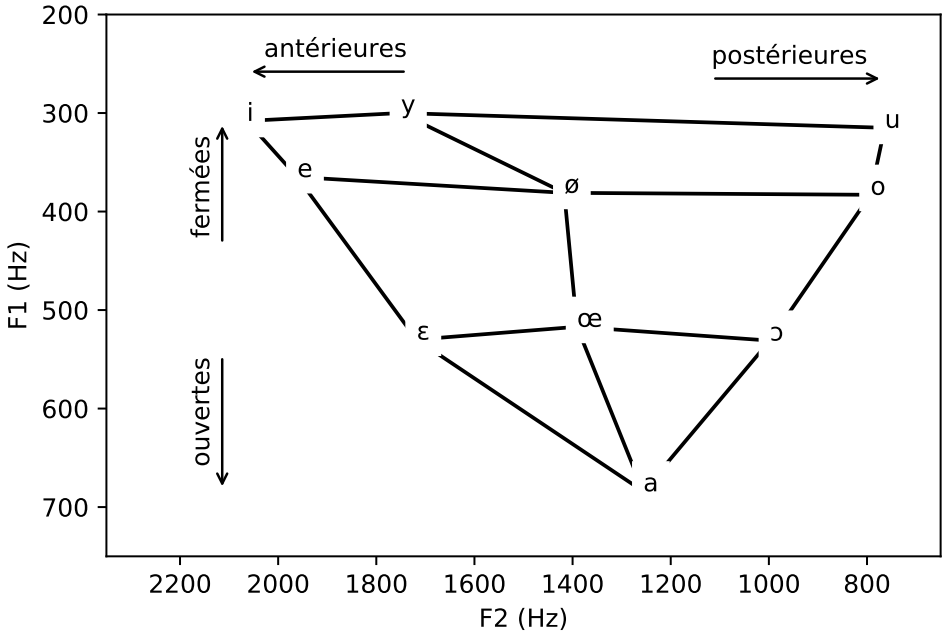


FIGURE 2.5 – Espace vocalique typique des voyelles de la langue française obtenu avec les valeurs du tableau 2.1.

2.2 Modifications des signaux de parole : approches fréquentielles et temporelles

Une fois le signal de parole analysé, il devient alors possible d'étudier mais aussi de modifier les segments d'analyse afin d'obtenir des segments de synthèse que l'on combine afin de générer le signal modifié. Les caractéristiques des fenêtres utilisées et l'emplacement des instants varient d'une approche à une autre en fonction des besoins de l'analyse et de la synthèse.

2.2.1 Approches fréquentielles

Le vocodeur de phase [57] introduit en 1966 par FLANAGAN et al. se base pleinement sur l'utilisation de la TFCT comme outil d'analyse-modifications-synthèse. L'analyse consiste à appliquer une TFCT classique comme introduit précédemment afin d'obtenir une représentation temps-fréquence du signal sur des instants d'analyse uniformément espacés. Les spectres court-terme peuvent alors être étudiés et modifiés donnant une nouvelle TFCT. En appliquant une transformée de Fourier inverse, on obtient des segments de synthèse qu'il convient alors d'additionner pour générer le nouveau signal de parole.

Une modification courante pour laquelle le vocodeur de phase est utilisé est la modification de l'échelle temporelle. En fixant des instants de synthèse différents des instants d'analyse, on peut faire varier le débit des signaux de parole tout en conservant le contenu spectral. Si les instants de synthèse sont différents des instants d'analyse, une discontinuité de la phase va alors apparaître d'une fenêtre à l'autre au sein d'une même composante fréquentielle. C'est pourquoi il est nécessaire d'effectuer un déroulement de phase, ou *phase unwrapping*, afin de rétablir cette "cohérence horizontale" des phases.

S'il est possible de faire varier l'échelle temporelle sans modifier le contenu spectral, il devrait être possible de faire l'inverse. En effet, il suffit d'appliquer des modifications temporelles puis de ré-échantillonner le signal pour compenser ces variations. Le débit de la parole sera donc inchangé et ce sera le contenu spectral qui aura subit les transformations. Par exemple, en prenant des instants de synthèse deux fois plus espacés et en sous-échantillonnant le signal synthétisé d'un facteur deux, on aura compressé tout le contenu spectral du signal de ce même facteur tout en conservant le rythme original de la parole. Notons que c'est bien tout le contenu spectral qui est compressé ce qui est problématique si l'on cherche seulement à modifier le fondamental.

Le vocodeur de phase est une des premières approches fréquentielles d'analyse-modifications-synthèse de la parole et les modifications de la parole qu'elle propose sont assez limitées. Pourtant, par sa simplicité elle permet de comprendre la base des approches fréquentielles et des notions importantes comme le déroulement de phase que l'on retrouvera dans de nombreuses approches. De multiples améliorations ont été proposées par la suite afin d'améliorer le vocodeur de phase permettant notamment un calcul plus performant de la phase [158, 104].

2.2.2 Approches temporelles

Une autre façon de procéder consiste à modifier le signal dans le domaine temporel. L'approche *Time Domain Pitch Synchronous Overlap and Add* (TD-PSOLA) est l'approche temporelle la plus populaire et fonctionne exclusivement dans le domaine temporel en travaillant sur des instants d'analyse et synthèse synchrones avec le fondamental [141]. La méthode consiste à générer un signal à partir d'une réorganisation du flux d'analyse court-terme afin d'obtenir les modifications prosodiques attendues. Le signal est alors synthétisé par une méthode d'addition-recouvrement des segments d'analyse aux instants de synthèse associés.

Pour des modifications de faible ampleur, les transformations appliquées par TD-PSOLA sont de très bonne qualité. Cependant pour des facteurs de modification trop importants, les nombreuses répétitions de segments non-voisés introduisent une corrélation à court-terme entre les segments ce qui provoque l'apparition d'un bruit tonal dérangeant.

De nombreuses méthode d'analyse-modifications-synthèse de la parole sont basées sur PSOLA, elles utilisent une analyse synchrone avec le fondamental et une synthèse par addition-recouvrement mais elles sont généralement combinées avec d'autres types d'analyse.

Frequency-Domain PSOLA (FD-PSOLA) [37] diffère de TD-PSOLA en travaillant sur une analyse fréquentielle des segments issus d'une TFCT. Il est alors possible de changer directement les caractéristiques spectrales des segments. Notons que les tailles des fenêtres d'analyse sont plus grandes, généralement quatre fois la période du fondamental, afin que les composantes harmoniques soient présentes dans la représentation fréquentielle.

Synchronous Overlap and Add (SOLA) [168] (resp. *Waveform Similarity Overlap and Add* (WSOLA) [212]), est une approche temporelle avec des instants d'analyse et synthèse non plus synchrones avec le fondamental mais choisis afin de favoriser une continuité naturelle au niveau des jointures de l'addition-recouvrement en maximisant une mesure de similarité basée sur le signal synthétisé (resp. sur le signal original). En proposant une procédure plus robuste pour les modifications temporelles, ces approches perdent la possibilité de manipuler le contenu spectral.

2.3 Principaux vocodeurs

Les méthodes présentées précédemment ont permis d'introduire les notions d'analyse court-terme, d'approche fréquentielle et d'approche temporelle. Cependant, afin de proposer des modifications mieux maîtrisées, les méthodes d'analyse-modifications-synthèse des signaux de parole se basent majoritairement sur les mécanismes de production de parole. En effet, lors d'une analyse court-terme des signaux de parole, les cavités supra-glottiques peuvent être modélisées par un filtre linéaire invariant. Le segment de parole étudié peut donc être modélisé comme l'interaction entre le filtre et une excitation glottique : on parle de modèle source-filtre de la parole. Dans cette section nous présentons succinctement les méthodes les plus utilisées dans le domaine qui permettront une compréhension des concepts utilisés dans le reste du document. Cependant, la liste n'est pas exhaustive et il existe de nombreuses autres approches cherchant à coupler celles déjà existantes ou proposant des concepts novateurs qui ne seront pas abordés ici.

2.3.1 Codage prédictif linéaire

Le codage prédictif linéaire, ou *Linear Predictive Coding* (LPC), de la parole a été introduit vers les années 60 [81, 9]. On parle de prédiction linéaire car on se base sur l'hypothèse qu'il est possible de prédire le signal à un instant donné à partir d'une combinaison linéaire des échantillons précédents. L'analyse LPC s'effectue à

des instants d'analyse uniformément répartis et son objectif principal est de calculer les coefficients du filtre qui minimisent le résidu qui représente l'excitation glottique. La re-synthèse des signaux de parole s'effectue classiquement par addition-recouvrement, les segments de synthèse étant obtenus en excitant le filtre avec le résidu.

La représentation source-filtre du codage LPC permet la modification de nombreux paramètres liés à la production de parole. Le premier type de modification concerne l'excitation glottique modélisée par le résidu. On choisira souvent des méthodes de transformation temporelle de type PSOLA pour modifier la hauteur ou le débit de la parole en agissant directement sur le résidu ; une interpolation du filtre sur les nouveaux instants de synthèse est alors nécessaire. Cette méthode s'appelle LP-PSOLA [141].

Concernant les modifications spectrales, tout type de modification peut être apporté au filtre mais il est aussi possible d'exploiter une caractéristique très intéressante du codage LPC qui est la modélisation du filtre par un type "tout pôle". Il devient alors aisé de calculer les pôles du filtre qui correspondent aux formants que l'on peut alors amplifier ou déplacer librement. On observera souvent une transformation mathématique des coefficients LPC en coefficients *Line Spectral Pairs* (LSP) qui permettent une meilleure robustesse à la quantification mais aussi permettant une manipulation plus naturelle des formants.

L'utilisation du codage LPC comme outil de modification de la parole est intéressant pour sa simplicité de compréhension et d'implémentation mais des modifications trop importantes de la source ou du filtre entraînent une synthèse de mauvaise qualité. En effet, le codage LPC a été développé dans le but de réduire la quantité d'information afin de pouvoir transmettre des signaux de parole à des débits intéressants. La simplicité du modèle atteint alors ses limites lorsqu'on lui impose des modifications de paramètres trop importantes.

2.3.2 Modèle sinusoïdal par somme de sinusoïdes

Le modèle sinusoïdal de la parole a été introduit par McAulay et al. en 1986 [122]. Il propose de modéliser l'excitation glottique comme une somme de composantes sinusoïdales aux caractéristiques variables. Les cavités supra-glottiques étant modélisées par un filtre linéaire, le signal de parole peut alors s'exprimer lui aussi comme une somme de composantes sinusoïdales. L'analyse consiste alors à estimer de façon robuste les fréquences, amplitudes et phases des composantes sinusoïdales à partir d'une TFCT. Sur chaque fenêtre d'analyse, des candidats sont extraits à partir des maxima d'amplitude spectral et leurs amplitudes et phases correspondantes sont estimées directement à partir du spectre. Les différents candidats entre les fenêtres adjacentes sont alors associés, cela se base sur les notions de "naissance", "continuité" et "mort" des pistes sinusoïdales.

Les paramètres des pistes sinusoïdales peuvent être modifiés avant la re-synthèse afin d'apporter des transformations au signal de parole. Il est aussi possible de décomposer les paramètres en composantes associées à l'excitation et au filtre à partir d'une déconvolution homomorphique. Il devient alors possible de modifier le fondamental et le filtre indépendamment.

Le fait qu'il n'y a pas de différenciation concernant le type d'excitation glottique responsable des différentes composantes sinusoïdales rend cette approche très robuste. Par contre, cela lui procure aussi un manque de souplesse dès lors que l'on cherche à modifier le signal de parole de manière plus fine. De plus, représenter la partie aléatoire sous forme de sinusoïde est très coûteux.

2.3.3 Modèle sinusoïdal "harmonique + bruit"

Pour palier ce problème, Laroche et al. introduisent en 1993 le modèle sinusoïdal "harmonique + bruit" [105], usuellement abrégé par l'acronyme anglais pour *Harmonic plus Noise Model* (HNM), qui suppose que les signaux de parole peuvent être décomposés en une partie déterministe et une partie stochastique. La partie déterministe représente la structure quasi-périodique du signal de parole et est donc modélisée par une somme harmonique de composantes sinusoïdales. La partie stochastique représente la structure aléatoire et transitoire du signal de parole. Des approches de ce type avaient déjà été proposées par le passé avec le modèle *Multi-Band Excitation* (MBE) par exemple [71] mais le modèle HNM s'est imposé comme la référence en introduisant de nombreuses simplifications dans la méthode qui la rende bien plus accessible et facile à implémenter. Quelques

exemples non-exhaustifs de ces simplifications sont une analyse synchrone au fondamental retirant alors la nécessité d'une correction de phase, ou encore une modélisation intelligente de la partie stochastique avec un modèle auto-régressif pour l'aspect fréquentiel, et une enveloppe énergétique pour le comportement temporel. Pour les segments voisés, une fréquence de voisement est estimée qui sépare le spectre en deux parties : la partie basse comporte la partie déterministe quasi-périodique et la partie haute comporte la partie stochastique. L'estimation des paramètres de la partie déterministe s'effectue en minimisant un critère des moindres carrés dans le domaine temporel, contrairement au modèle par somme de sinusoïdes où elle s'effectue dans le domaine fréquentiel ; cela permet de travailler avec des fenêtres d'analyse plus courtes ce qui améliore la modélisation des transitoires. Les coefficients du filtre "tout pôle" de la partie stochastique sont estimés par une approche classique corrélationnelle comme pour le codage LPC. Pour la re-synthèse, les composantes déterministes et stochastiques sont générées séparément puis elles sont additionnées.

Les paramètres des composantes sinusoïdales peuvent être modifiés avant la re-synthèse afin d'apporter des transformations au signal de parole. Comme pour le modèle par somme de sinusoïdes, il est aussi possible de décomposer les paramètres en composantes associées à l'excitation et au filtre pour pouvoir modifier le fondamental et le filtre indépendamment. Pour cela on estime une enveloppe spectrale continue à partir des valeurs discrètes connues au niveau des harmoniques [33] et de nombreux travaux proposent des estimations de plus en plus robustes [214, 14, 133].

En couplant une approche harmonique, ne nécessitant plus la mise en place de pistes sinusoïdales complexes, et un traitement à part de la partie stochastique, permettant une synthèse dénuée de bruit tonal, le modèle HNM est généralement le modèle sinusoïdal utilisé pour les modifications prosodiques classiques de la parole. La plupart des études sur le modèle HNM se concentrent sur la partie harmonique et négligent la partie stochastique dont la modélisation est pourtant primordiale lorsque des modifications liés à cette partie sont introduites e.g. modifications temporelles et effort vocal. Des travaux, comme ceux proposés par D'ALESSANDRO et al., procurent alors une représentation plus fidèle de la partie stochastique permettant des modifications prosodiques complexes tout en conservant une très bonne qualité de synthèse [165, 218].

2.3.4 STRAIGHT

L'outil d'analyse-modifications-synthèse *Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum* (STRAIGHT) [91] développé par KAWAHARA depuis 1997 qui, contrairement aux approches précédentes, ne cherche pas à réduire la quantité d'information, a pour objectif principal de permettre des modifications importantes des paramètres de la parole tout en conservant une re-synthèse de haute qualité. Le diagramme de flux d'information du vocodeur STRAIGHT est observable figure 2.6.

Analyse

L'estimation du fondamental est une étape importante, dans tout modèle d'analyse-modifications-synthèse de la parole, et se fait généralement par une approche temporelle par auto-corrélation, ou fréquentielle en cherchant les maxima d'amplitude d'une TFCT. Ces méthodes se basent sur des hypothèses de stabilité et de périodicité locales du signal. En réalité, il ne l'est pas parfaitement, la méthode d'estimation proposée dans STRAIGHT [93] est de supposer que le signal a une structure quasi-harmonique :

$$x(t) = \sum_{l=1}^{\mathrm{L}(t)} \mathrm{A}_l(t) \cos\left(l \int_0^t (\omega_0(\tau) + \omega_l(\tau)) d\tau + \phi_l \right), \tag{2.1}$$

avec $\omega_0(t)$ la fréquence instantanée du fondamental, $\mathrm{A}_l(t)$ l'amplitude variable de la $l^{\text{ème}}$ composante harmonique et $\omega_l(t)$ la perturbation lentement variable qui lui est associée. La méthode d'estimation proposée exploite alors plutôt la phase de la TFCT en estimant le fondamental à partir d'une analyse du spectre de fréquence instantanée par une méthode de points fixes comme introduit par CHARPENTIER [36], puis reprise par ABE et al. [2, 3], et améliorée par KAWAHARA et al. [92]. Si sur certains segments d'analyse aucune valeur de fondamental ne se

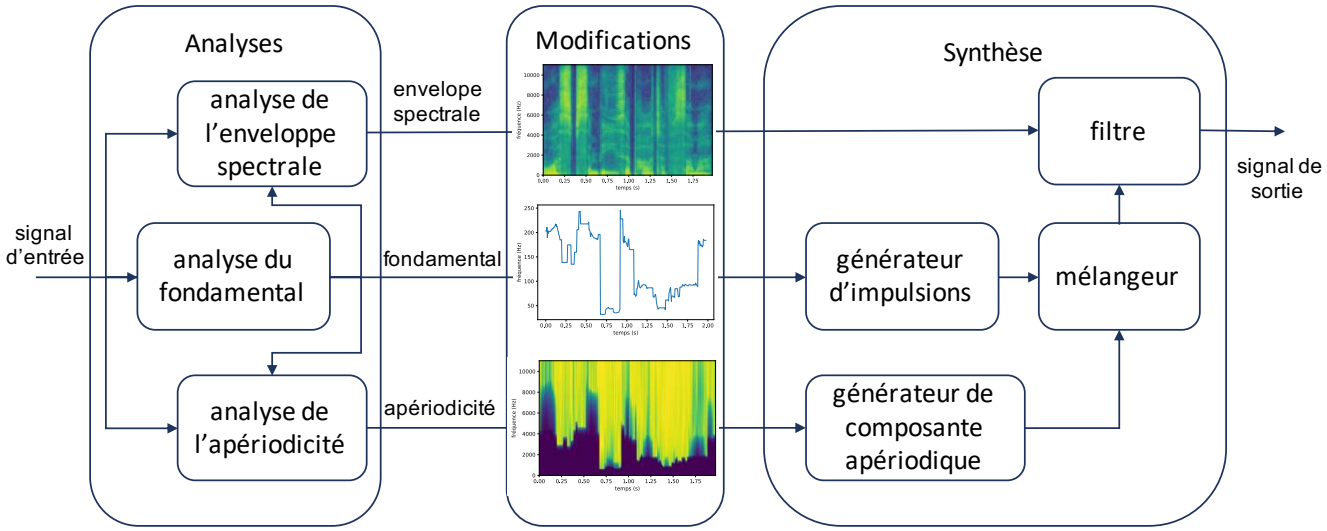


FIGURE 2.6 – Diagramme de flux d'information dans le vocodeur STRAIGHT.

distingue, ils sont considérés comme non voisés. Sur les segments voisés, les harmoniques sont aussi estimés par une approche similaire et les composantes obtenues servent à raffiner l'estimation du fondamental. Les instants et la taille des fenêtres d'analyse, sur lesquels vont être estimés les autres paramètres, sont alors synchrones au fondamental pour les segments voisés, et constants pour les segments non-voisés.

Pour l'estimation de l'enveloppe spectrale $S(\omega, t)$, on suppose que les amplitudes $A_l(t)$ des composantes harmoniques en sont des échantillons et on cherche à reconstruire la surface continue, sans interférences temporelles, ni fréquentielles, liées à la quasi-périodicité du signal. L'utilisation d'une fenêtre temporelle compensatoire permet d'annuler les interférences temporelles. Un lissage par fonctions splines permet quant à lui d'annuler les interférences fréquentielles tout en étant robuste aux erreurs d'estimation du fondamental.

Finalement, les perturbations lentement variables des harmoniques présentées équation 2.1 vont introduire des composantes supplémentaires sur les fréquences non-harmoniques. L'énergie présente sur ces fréquences, normalisée par l'énergie totale, sert alors comme mesure de l'apériodicité $P_{AP}(w, t)$ du signal. Plus précisément, l'apériodicité à un instant t est calculée comme le rapport de l'enveloppe spectrale inférieure $S_i(\omega, t)$, obtenue en reliant les vallées de $S(\omega, t)$, normalisée par l'enveloppe spectrale supérieur $S_s(\omega, t)$, obtenue en reliant les sommets de $S(\omega, t)$. On a alors :

$$P_{AP}(w, t) = \frac{\int W(\lambda - \omega) \mid S(\lambda, t) \mid^2 \frac{|S_i(\lambda,t)|^2}{|S_s(\lambda,t)|^2} d\lambda}{\int W(\lambda - \omega) \mid S(\lambda, t) \mid^2 d\lambda} \tag{2.2}$$

avec $W(\omega)$ une fenêtre de pondération représentant la forme des filtres auditifs.

Synthèse

Les paramètres de la source (trajectoire du fondamental, composantes apériodiques) et du filtre (enveloppe spectrale) sont alors utilisés pour synthétiser le signal de sortie. La première étape est la génération de la source qui pourrait se faire classiquement en utilisant des trains d'impulsions pour les segments voisés et du bruit blanc pour les segments non-voisés. Un problème inhérent de ces approches est l'apparition d'un bourdonnement qui dégradent la qualité de la synthèse. En se basant sur la littérature traitant des effets de la phase sur la perception du timbre des signaux de parole, STRAIGHT propose une génération de la source par contrôle de la structure temporelle fine en se basant sur des filtres passe-haut permettant de manipuler le retard de groupe. Les détails de calcul sont présentés dans l'article de référence de la méthode STRAIGHT [91]. Les segments de synthèse sont obtenus par convolution de la source et de filtres à minimum de phase calculés à partir de l'enveloppe spectrale. Le signal de parole est alors synthétisé par addition-recouvrement des segments de synthèse.

Modifications

Les paramètres extraits sont représentés sous la forme de vecteurs (fondamental) ou matrices (enveloppe spectrale, apériodicité) de nombres réels. Il est donc aisé de modifier librement ces paramètres afin d'appliquer les transformations prosodiques attendues.

Discussion

STRAIGHT est une approche d'analyse-modifications-synthèse très performante. Les paramètres peuvent être modifiés indépendemment ce qui procure une bonne flexibilité et le signal synthétisé conserve une sonorité naturelle même pour d'importantes modifications. Cependant, le processus demande de nombreuses opérations et nécessite une connaissance du signal dans son entièreté avant de pouvoir le traiter. Tandem-STRAIGHT [94], est une version de l'outil permettant d'obtenir des paramètres proches de ceux extraits par Legacy-STRAIGHT (appellation désormais donnée à la première version pour les distinguer) mais avec largement moins d'opérations et donc un temps de traitement beaucoup plus court.

Cette réduction de la complexité s'accompagne d'une baisse de la qualité du signal synthétisé [134] et une utilisation en temps réel n'est toujours pas envisageable. Une extension pour une telle utilisation a été proposée [16] mais les simplifications additionnelles engendrent une baisse supplémentaire de la qualité du signal synthétisé.

WORLD

L'outil d'analyse-modifications-synthèse WORLD [139], proposé par MORISE et al., est une approche récente qui se place dans la continuité de STRAIGHT en visant une très bonne qualité de synthèse même pour des modifications importantes des paramètres vocaux, avec une utilisation possible en temps réel. Pour cela, WORLD se base sur la même approche que STRAIGHT avec une estimation des mêmes trois flux de paramètres mais avec des algorithmes modernes à savoir DIO [137] pour le fondamental, CheapTrick [133] pour l'enveloppe spectrale et PLATINUM [136] pour l'apériodicité. En l'absence de modification, la première version de WORLD génère des signaux de qualité légèrement supérieure à Legacy-STRAIGHT pour un temps de calcul bien inférieur.

Il est aussi important de noter que WORLD est mis à jour de manière récurrente, les dernières modifications en date sont une estimation du fondamental avec l'algorithme Harvest [135] et une estimation de l'apériodicité avec l'algorithme D4C [134]. En 2018, MORISE et al. montrent qu'en l'absence de modification la version la plus récente de WORLD fournit une synthèse de qualité significativement supérieure à Legacy-STRAIGHT. En revanche, ils précisent bien que les prochaines étapes de validation consisteront à les comparer pour des tâches de modification de parole ou de synthèse paramétrique.

Conclusion du chapitre 2

Nous avons vu dans ce chapitre qu'il existe de nombreuses approches d'analyse-modifications-synthèse de la parole. Afin de proposer des modifications naturelles de la parole, leur analyse est basée sur les mécanismes physiologiques de production de la parole et la majorité des approches cherchent à modéliser ces mécanismes et à estimer les paramètres qui leur sont associés.

Un critère important qui a été introduit dans ce chapitre est aussi le rapport entre quantité d'information extraite et qualité du signal synthétisé. En effet, certaines approches cherchent à réduire la quantité d'information nécessaire afin de pouvoir stocker, ou transmettre, efficacement les signaux de parole. D'autres cherchent plutôt à offrir la possibilité de produire d'importantes modifications avec pour seul contrainte le naturel et la qualité des signaux synthétisés. Naturellement, les premières auront de moins bons résultats que les deuxièmes car la présence de contraintes supplémentaires entraînent des compromis qui impactent nécessairement les performances du système en terme de qualité de synthèse.

Pour nos travaux de transformation paramétrique de la parole, qui seront présentés Partie 3, le vocodeur WORLD semble le plus adapté car il offre une grande flexibilité de transformation tout en conservant une qualité de synthèse théoriquement importante. Cependant, l'absence d'études sur la qualité de synthèse de WORLD, suite à d'importantes modifications, nous orientera plutôt vers le vocodeur Legacy-STRAIGHT dont les performances ont déjà été validées à de nombreuses reprises. Cela permettra aussi de faciliter la comparaison avec les travaux connexes, majoritairement basés sur ce vocodeur. En revanche, les mises à jours permanentes de WORLD, qui s'accompagnent de performances de plus en plus prometteuses, méritent d'être surveillées et une transition prochaine vers ce vocodeur est très largement envisagée. De plus, ces deux vocodeurs exploitent exactement le même flux d'information ce qui rend le passage, de l'un à l'autre, quasi-transparent.

Chapitre 3

Améliorations naturelles de l'intelligibilité de la parole dans le bruit

Sommaire

Retour à la table des matières

Introduction du chapitre 3

Comme nous l'introduisions Chapitre 1, de nombreux facteurs peuvent être responsables d'une dégradation de l'intelligibilité de la parole. Lors d'une conversation directe, et en fonction du contexte d'écoute, des stratégies sont alors mises en place par le locuteur afin d'améliorer l'intelligibilité de sa parole pour l'auditeur. Les modifications naturelles ainsi introduites dépendent du phénomène responsable de la dégradation de l'intelligibilité. Les stratégies utilisées par le locuteur varieront alors si l'environnement d'écoute est bruyant ou réverbérant, si l'auditeur est un malentendant, une personne âgée ou encore un enfant. En renforcement de la parole, les modifications numériques mises en place, visant à améliorer l'intelligibilité de la parole, s'inspirent principalement de ces modifications naturelles.

Les signaux diffusés dans les habitacles automobiles ont généralement déjà été rehaussés. De plus, l'utilisation de matériaux acoustiques absorbants, devenue habituelle chez les constructeurs automobiles, permet de minimiser grandement le temps de réverbération des signaux audio et de rendre ce facteur quasi-négligeable du point de vue de l'intelligibilité. La présence de bruit dans l'environnement d'écoute est alors le facteur principal qui limite l'intelligibilité des signaux de parole. Ce sont donc les stratégies naturelles d'adaptation de la production de parole, permettant une amélioration de l'intelligibilité dans le bruit, qui nous intéressent pour le développement de méthodes numériques de renforcement de la parole en contexte automobile.

Dans ce chapitre, nous nous intéressons donc à deux de ces stratégies naturelles d'adaptation de la production de parole qui permettent une amélioration de l'intelligibilité dans le bruit à savoir la parole Lombard et la parole claire. Ces deux modes de parole, ainsi que des bases de données d'étude associées, seront présentés section 3.1. Puis, une analyse détaillée de l'intérêt, et de la contribution, de plusieurs aspects spectro-temporels de ces stratégies, vis-à-vis de l'amélioration de l'intelligibilité dans le bruit, est proposée section 3.2.

3.1 Améliorations naturelles de l'intelligibilité de la parole

Lors d'une conversation naturelle, le locuteur modifie sa production de parole pour mieux se faire comprendre par son interlocuteur. Les modifications naturelles ainsi introduites dépendent du phénomène responsable de la dégradation de l'intelligibilité. De plus, les stratégies possèdent une grande variabilité qui rendent complexe l'étude des modifications introduites. En effet, le genre, la langue, la catégorie socio-professionnelle ou même l'humeur sont une liste non exhaustive de facteurs interpersonnels qui influent sur les stratégies mises en place.

3.1.1 Parole Lombard

La parole Lombard tient son nom de l'oto-rhino-laryngologiste français LOMBARD qui la traite pour la première fois en 1911 [113]. Elle correspond au mode de production de parole lorsqu'une personne tente de se faire comprendre en présence de bruit ambiant.

Afin d'étudier cette stratégie de production de parole, de nombreuses bases de données Lombard ont été mises en place. Elles consistent à enregistrer un locuteur effectuant une tâche vocale dans un environnement bruyant. À l'ensemble des paramètres interpersonnels, introduits précédemment, s'ajoute l'influence de nombreux paramètres qui vont influencer les stratégies observées, à savoir :

- le type de bruit utilisé et son volume, ces paramètres influence la stratégie Lombard employée par les locuteurs,
- le mode de présentation du bruit, un casque audio fermé influence la stratégie Lombard mais des hauts parleurs complexifie l'enregistrement de la parole Lombard sans le bruit, un casque audio ouvert est donc ce qui est préféré,
- la tâche vocale, elle consiste majoritairement en une tâche de lecture simple mais il arrive qu'on enregistre de la parole spontanée, ou que le locuteur doivent interagir avec une tierce personne, ce qui complexifie l'analyse des données mais influence grandement la stratégie Lombard,

- le matériel linguistique, il doit être choisi avec précaution car la stratégie Lombard ne sera pas la même e.g. sur des listes de mots, des listes de chiffres, des listes de phrases qui ont du sens, ou non.

Une liste quasi-exhaustive de 40 bases de données Lombard existantes en 2007 est dressée par GARNIER [59], on s'aperçoit alors que les paramètres introduits précédemment varient grandement d'une base à une autre, ainsi que les résultats d'analyse de celles-ci.

Quelques bases de données Lombard en contexte bruité automobile très fournies existent mais elles sont difficilement exploitables. Par exemple, la base de données en libre accès AVICAR [107], pour laquelle une centaine de locuteurs anglais ont été enregistrés dans de multiples conditions (voiture à l'arrêt ou à différentes vitesses, avec les fenêtres ouvertes ou fermées). Les enregistrements ont été effectués par une rangée de 8 microphones placée devant le locuteur dans le véhicule, cependant l'utilisation d'algorithmes de séparation de source dégrade fortement le signal de parole isolé et limite son exploitation pour des analyses qui nécessitent un signal de parole net. Ou encore, la base de donné SpeechDatCAR [132], pour laquelle plusieurs centaines de locuteurs, pour plusieurs langues européennes, ont été enregistrés dans des conditions encore plus diversifiées prenant même en compte les conditions de la chaussée ou l'utilisation du système son du véhicule. Les enregistrements ont, cette fois, été effectués à travers un micro directif, placé proche du locuteur, permettant d'avoir des signaux de parole quasi-intacts. En revanche, la base de données est payante et le coût, proportionnel à sa qualité, n'était pas envisageable pour notre étude.

Pour notre étude, c'est donc sur une base de données Lombard récente en libre accès, Lombard-GRID (2018) [6], que nous nous sommes orientés. Cette base de données est composée d'enregistrements, audio et vidéo, de 54 sujets normo-entendants (30 femmes et 24 hommes). Chaque locuteur prononce 50 phrases sans sens, dont la syntaxe est tirée de la base de données GRID [42], une fois dans le silence et une fois dans un bruit SSN. Ces phrases de six mots anglais, sont générées aléatoirement à partir d'une liste finie de mots clefs, qui sont visibles dans le tableau 3.1. Les choix des mots sont fait à partir d'une distribution uniforme et chaque mot apparaît plusieurs fois pour chaque sujet. Voici quelques avantages de la base de données :

- enregistrement de la parole normale et la parole Lombard pour chaque phrase,
- parole Lombard enregistrée sans bruit, grâce au bruit diffusé via un casque audio ouvert,
- un grand nombre d'enregistrements (54 sujets, 2700 phrases, 16200 mots).

Et voici quelques inconvénients :

- phrases sans sens donc l'effet Lombard produit peut être impacté,
- mots anglais, bien que le français soit phonétiquement proche de l'anglais, l'effet Lombard dépend de la langue,
- phrases produites dans un bruit SSN différent des bruits automobiles.

Ainsi, l'analyse et la mise en place d'algorithmes sur la base de données Lombard GRID est une bonne première étape et, en fonction des résultats obtenus dans notre contexte automobile, il sera possible de dimensionner nos besoins en terme de données afin d'acquérir une base plus adaptée, voir seulement une sous partie de celle-ci.

TABLEAU 3.1 – Mots utilisés dans les phrases GRID

commande	couleur	préposition	lettre	numéro	adverbe
bin	blue	at	A-Z	1-9, zero	again
lay	green	by	sans W		now
place	red	in			please
set	white	with			soon

3.1.2 Parole claire

La parole claire correspond au mode de production de parole lorsqu'il est demandé à un locuteur de parler clairement pour se faire comprendre dans une situation d'écoute défavorable comme, par exemple, un canal de

communication perturbé, un auditeur non-natif ou un auditeur atteint de déficience auditive. Bien qu'il ne soit pas dirigé vers une écoute en environnement bruyant, on remarque que la parole claire engendre tout de même une amélioration de l'intelligibilité significative en présence de bruit [153].

De même que pour la parle Lombard, de nombreuses bases de données de parole claire existent. Bien que l'absence de bruit facilite grandement les enregistrements, la parole claire n'est pas clairement définie et les stratégies employées par le locuteur dépendront grandement de la tâche qui leur a été communiquée, en plus de l'ensemble des paramètres interpersonnels introduits précédemment.

Pour notre étude, nous nous sommes orientés sur la base de données LUCID [15]. Cette base de données est composée d'enregistrement audio de 40 sujets normo-entendants (20 femmes et 20 hommes). Chaque locuteur prononce 144 phrases sensées, décrivant des situations de vie en anglais, une fois normalement comme s'il s'adressait à un ami et une fois clairement comme s'il s'adressait à une personne atteint de déficience auditive. Voici deux exemples de phrases prononcées : "*The old lady ate the peach.*" et "*The young children loved the beach.*".

3.1.3 Stratégies naturelles

Dans cette section, nous proposons de faire une synthèse des modifications introduites par ces deux stratégies de production de parole. Cette synthèse est basée sur les travaux de COOKE et al. qui recensent l'état actuel des connaissances sur les modifications de la parole induites par de nombreux contextes d'écoute spécifiques [43]. Certaines modifications seront illustrées par des analyses que nous avons menés sur les bases de données Lombard-GRID et LUCID introduites précédemment.

Améliorations de l'audibilité

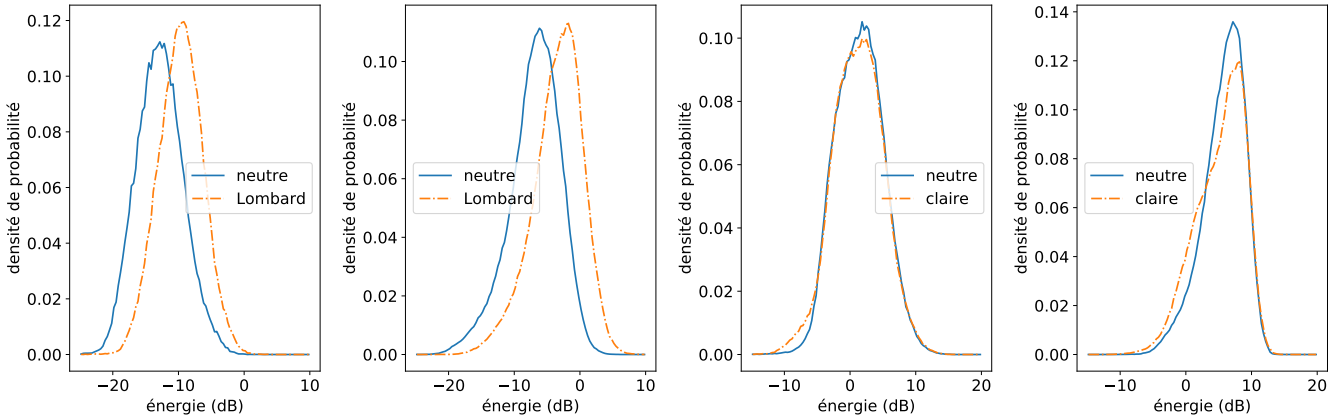
En présence de masquage énergétique, la modification de la parole la plus évidente est l'augmentation de son audibilité et plusieurs stratégies sont employées dans l'effet Lombard, à savoir :

- une augmentation de l'intensité globale [216, 103, 52, 188, 23, 83, 34], ce qui permet une émergence du contenu spectral de la parole au-dessus du bruit masquant,
- une élévation du fondamental [188, 21] et un aplatissement de la pente spectrale [156, 182, 188, 127, 83, 200, 34, 157], ce qui pourrait être une conséquence physiologique de l'augmentation de l'intensité globale [203, 111, 190], ou une stratégie de concentration du contenu spectral dans les médiums où l'oreille est la plus sensible,
- renforcement des sons voisés [52, 182, 188, 127, 83, 34, 85, 24, 62], l'augmentation de l'intensité des voyelles est plus importante que pour les consonnes et l'augmentation de l'intensité des consonnes sonantes est plus importante que pour les consonnes obstruantes,
- une ré-allocation spectro-temporelle, dépend grandement du type de bruit et il en existe deux approches :
  - une amplification sélective de la parole dans des régions du spectre où le bruit est très présent pour empêcher le masquage [127, 84], plutôt adapté aux bruits large bande,
  - un déplacement de la parole dans des régions du spectre [59, 115, 116, 62], ou du temps [44, 11], où le bruit est moins important, plutôt adapté aux bruits à bande étroite.

La parole claire n'étant pas produite dans un environnement bruyant, l'amélioration de l'audibilité ne vise pas à faire émerger la parole par dessus un potentiel bruit masquant. On remarque tout de même souvent une augmentation de l'intensité globale toujours rattachée à une élévation du fondamental [26] et un aplatissement de la pente spectrale [154]. En revanche, on remarque un comportement inverse du renforcement des sons en fonction du voisement. En effet, c'est plutôt un équilibrage de l'audibilité qui est observé avec une augmentation de l'intensité des consonnes [154, 26] et plus particulièrement des consonnes obstruantes [38]. Finalement, on note aussi un rehaussement des formants [154, 98].

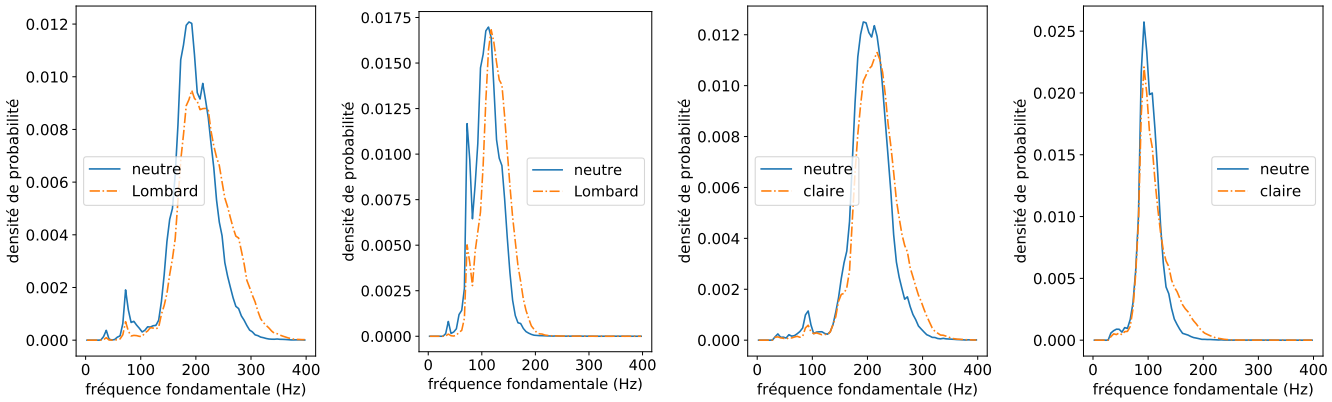
Les distributions de l'énergie en fonction du voisement sur les signaux neutre/Lombard et neutre/claire des deux bases de données sont visibles figure 3.1. On remarque effectivement une augmentation du niveau global

pour la parole Lombard, légèrement plus importante pour les sons voisés. Pour la parole claire, on n'observe pas d'augmentation du niveau global dans la base de données LUCID, par contre on relève une légère diminution de l'énergie des sons voisés diminuant l'écart avec l'énergie des sons non-voisés. Les distributions de la fréquence fondamentale en fonction du genre sur les signaux neutre/Lombard et neutre/claire des deux bases de données sont visibles figure 3.2. On remarque effectivement une augmentation du fondamental pour les deux modes de parole. Enfin, une comparaison des spectres long-terme neutre/Lombard et neutre/claire des deux bases de données sont visibles figure 3.3. À travers les spectres relatifs, on remarque la tendance, des deux modes de parole, de concentrer l'énergie de la parole dans les médiums et plus particulièrement pour la parole Lombard.



(a) GRID non-voisés (b) GRID voisés (c) LUCID non-voisés (d) LUCID voisés

FIGURE 3.1 – Distributions de l'énergie en fonction du voisement sur les signaux neutre/Lombard de la base de données Lombard-GRID et sur les signaux neutre/claire de la base de données LUCID.



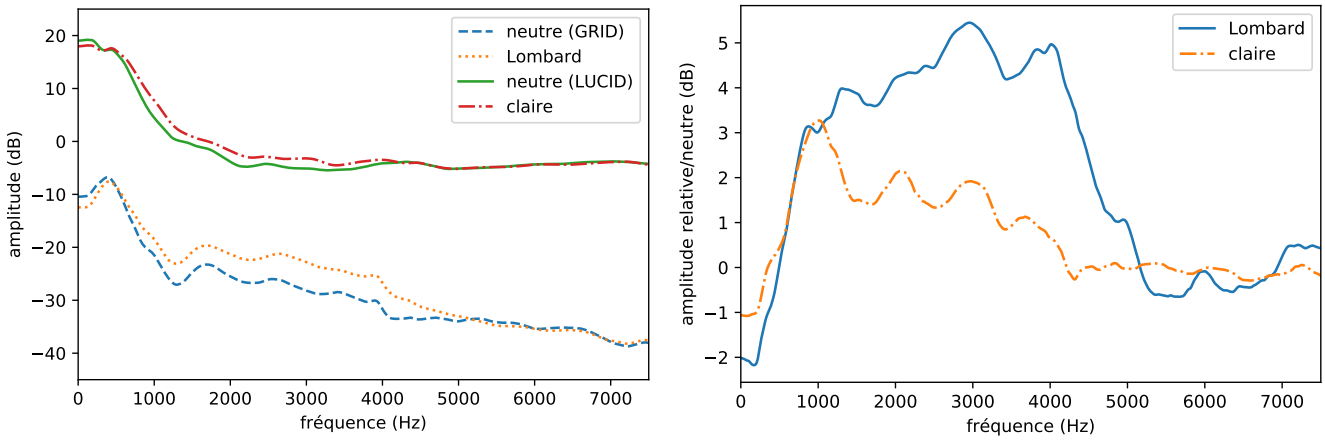
(a) GRID femmes (b) GRID hommes (c) LUCID femmes (d) LUCID hommes

FIGURE 3.2 – Distributions de la fréquence fondamentale en fonction du genre sur les signaux neutre/Lombard de la base de données Lombard-GRID et sur les signaux neutre/claire de la base de données LUCID.

Améliorations de la séparation

Améliorer la séparation entre son discours et le bruit est aussi une tendance qui a été remarquée lors d'études sur l'effet Lombard. Dans un bruit intense, augmenter l'amplitude dynamique des caractéristiques acoustiques permet une meilleure discrimination et les stratégies recensées qui permettent d'y parvenir sont les suivantes :

- amplification des modulations basse fréquence de l'enveloppe temporelle [62],
- amplification de la modulation du fondamental (vibrato) [24, 59, 59].



(a) spectres long-terme

(b) spectres relatifs

FIGURE 3.3 – Comparaison des spectres long-terme neutre/Lombard de la base de données Lombard-GRID et neutre/claire de la base de données LUCID. Le calibrage des bases de données n'étant pas connu, le niveau de référence utilisé pour chaque base de données est arbitraire.

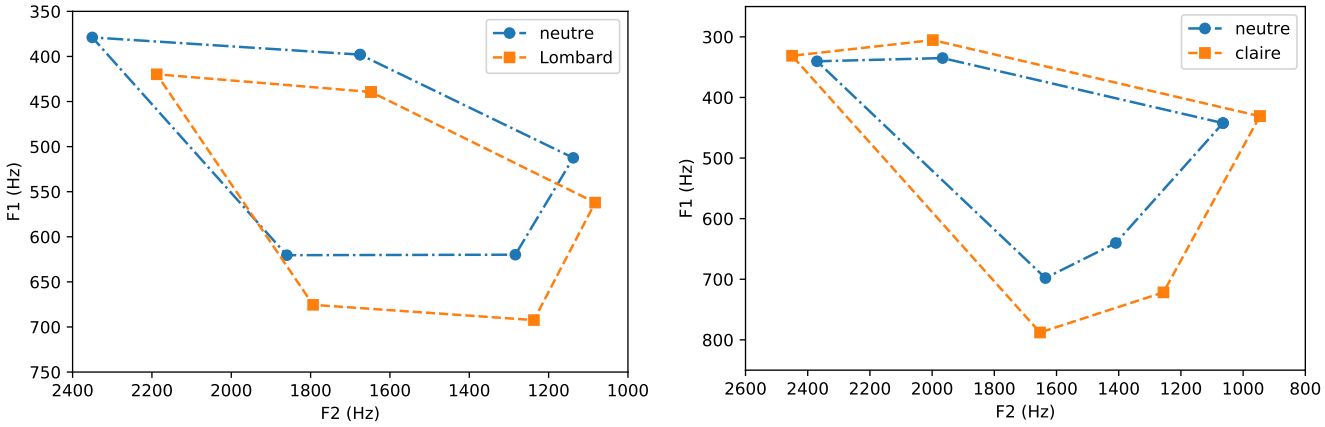
Par contre, dans un bruit de conversation avec peu de locuteurs aux caractéristiques acoustiques d'identité vocale proches du locuteur, donc avec un très fort masquage informationnel, on pourrait croire que la parole Lombard cherche à modifier ses caractéristiques acoustiques afin de se démarquer de la, ou les, parole(s) concurrente(s). Cependant, les études sur l'effet Lombard n'indiquent pas de telles modifications [115, 44].

Bien que la parole claire ne soit pas produite en présence d'un signal concurrent explicite, on retrouve tout de même une amplification des modulations basse fréquence de l'enveloppe temporelle dans ce mode de parole [98].

Renforcement des informations linguistiques

Le renforcement des informations linguistiques est un aspect qui est très présent dans les deux modes de parole avec :

- une augmentation du délai d'établissement du voisement, ce qui permet un meilleur contraste entre les différentes consonnes occlusives aussi bien pour la parole Lombard [75] que pour la parole claire [38, 154],
- une modification de la prosodie avec :
  - l'insertion de pauses entre certaines syllabes plus importantes pour la parole claire [154] que pour la parole Lombard [63],
  - augmentation de la durée des syllabes, et modification de la trajectoire du fondamental, à la frontière entre les mots et les phrases pour bien les démarquer [61, 63].



(a) GRID

(b) LUCID

FIGURE 3.4 – Comparaison des espaces vocaliques neutre/Lombard de la base de données Lombard-GRID et neutre/claire de la base de données LUCID.

Concernant les voyelles, une hyper-articulation est aussi observée dans les deux modes de parole. Pour la parole Lombard, on assiste à une translation de l'espace vocalique, introduit chapitre 2, vers les hautes fréquences, avec une élévation systématique du premier formant F1 et du second formant F2 [188, 23, 60]. Comme pour le fondamental F0, mais pas pour le F2, l'élévation de F1 est fortement corrélée au forçage vocal [111], ce qui pourrait donc être une conséquence physiologique plus qu'une stratégie d'élévation du premier formant. Pour la parole claire, c'est un élargissement de l'espace vocalique qui est observé avec une élévation du F2 pour les voyelles antérieures et un abaissement pour les voyelles postérieures. En revanche, bien qu'on observe généralement une élévation du F1 pour les voyelles ouvertes et un abaissement pour les voyelles fermées [38, 154, 98], il arrive que le F1 soit aussi élevé pour les voyelles fermées résultant en une translation sur l'échelle du premier formant [55]. Cette différence peut s'expliquer par la présence plus ou moins intense de forçage vocal dans la parole claire étudiée qui peut alors présenter des comportements variables.

Une comparaison des espaces vocaliques neutre/Lombard et neutre/claire des deux bases de données sont visibles figure 3.4. Pour la parole Lombard, on remarque effectivement une élévation des F1 et F2 résultant en une translation de l'espace vocalique. Pour la parole claire, c'est le comportement classique qui est observé avec un élargissement de l'espace vocalique. Cette dernière observation n'est pas surprenante car l'absence d'augmentation du niveau global, vu sur la figure 3.1, souligne la faible présence de forçage vocal dans la base de données LUCID.

Réduire l'effort cognitif

D'autres stratégies vont plutôt chercher à réduire l'effort cognitif de l'auditeur pour faciliter la compréhension de la parole avec :

- un débit de parole ralenti pour la parole claire [154, 26], que l'on retrouve moins marqué mais toujours observable pour la parole Lombard [156, 83, 85] bien que ce ne soit pas toujours le cas [59],
- la mise en avant des mots importants et des informations nouvelles [61, 59, 152], en augmentant le contraste du fondamental et de l'intensité par rapport aux mots-outils,
- une augmentation de l'amplitude des paramètres articulatoires pour la parole Lombard lorsque l'auditeur voit le locuteur [56], mais pas toujours observé [64].

3.2 Intérêt des modifications Lombard et parole claire dans le bruit

3.2.1 Intérêt des modifications Lombard dans le bruit

COOKE et al. proposent d'analyser indépendamment la contribution à l'amélioration de l'intelligibilité de plusieurs modifications introduites par l'effet Lombard [117, 46]. Le principe de l'étude est d'apposer indépendamment certaines modifications locales, ou globales, de signaux de parole Lombard sur des signaux de parole neutre, et vice-versa, afin d'étudier leur contribution à l'amélioration de l'intelligibilité. Les tests subjectifs mis en place suivent une méthode de présentation fixe correspondant à un score d'intelligibilité moyen d'environ 50% pour la parole neutre.

Élévation du fondamental

Dans un premier temps, les auteurs proposent d'étudier la contribution de l'élévation de la valeur moyenne du fondamental à l'amélioration de l'intelligibilité dans un bruit stationnaire [117]. Une manipulation est effectuée consistant à rehausser la valeur moyenne du fondamental du signal neutre à celle des signaux Lombard correspondants avec le vocodeur STRAIGHT. Les résultats des test subjectifs dans un bruit SSN indiquent que l'augmentation de la fréquence fondamentale ne contribue pas à une amélioration significative de l'intelligibilité. Bien que le bruit SSN soit un bruit majoritairement basse fréquence, l'élévation du fondamental pourrait ne pas être suffisante pour que la migration de l'énergie spectrale vers les hautes fréquences apporte un gain d'intelligibilité significatif. Dans un bruit de conversation, il est aussi possible que l'élévation du fondamental favorise la distinction entre les locuteurs, améliorant ainsi l'intelligibilité.

Aplatissement de la pente spectrale

Dans un deuxième temps, les auteurs proposent d'étudier la contribution de l'aplatissement de la pente spectrale à l'amélioration de l'intelligibilité dans le bruit [117]. Pour chaque paire de signaux, un filtre est conçu pour que le spectre du signal neutre, au fondamental rehaussé, filtré par celui-ci colle au spectre du signal Lombard. Il est alors possible d'appliquer ce filtre aux signaux neutres originaux pour introduire l'aplatissement de la pente spectrale de la parole Lombard correspondante. Contrairement à l'augmentation de la fréquence fondamentale, les résultats des test subjectifs dans un bruit SSN indiquent que l'aplatissement de la pente spectrale procure un gain d'intelligibilité significatif d'environ 20 point de pourcentage (p.p.). Ce gain d'intelligibilité notable est pourtant significativement moins important que celui des paroles Lombard naturelles qui est autour de 30 p.p.. Ainsi l'aplatissement global de la pente spectrale est un facteur important de l'amélioration de l'intelligibilité dans la parole Lombard mais d'autres modifications prosodiques pourraient également y contribuer.

Modification du débit de parole

Enfin, ils proposent d'étudier la contribution des modifications de débit de la parole Lombard à l'amélioration de l'intelligibilité dans le bruit [46]. Pour chaque paire de signaux, la première approche proposée consiste à modifier uniformément l'échelle temporelle afin de faire correspondre leur durée. La deuxième approche proposée consiste à les aligner localement en combinant de la déformation temporelle dynamique et des techniques temporelles de modification de débit en utilisant le logiciel *VocALign* [191] basé sur TD-PSOLA. Ainsi, il est possible d'apposer le débit local des signaux d'un mode, neutre ou Lombard, sur les signaux de l'autre mode. Les résultats des tests subjectifs dans un bruit SSN ne montrent pas de changement significatif de l'intelligibilité, que ce soit avec l'approche locale ou l'approche uniforme. Cela pourrait s'expliquer par l'amplitude des adaptations du débit de la parole Lombard qui n'est pas suffisamment importante pour générer un gain d'intelligibilité significatif.

3.2.2 Intérêt des modifications de la parole claire dans le bruit

Bien qu'il ne soit pas dirigé vers une écoute en environnement bruyant, on remarque que la parole claire engendre tout de même une amélioration de l'intelligibilité significative en présence de bruit [153]. Nous avons d'ailleurs remarqué que de nombreuses tendances de la parole Lombard se retrouvaient dans la parole claire. Dans une série d'articles intitulés *Speaking Clearly for the Hard of Hearing*, plusieurs auteurs ont cherché à étudier l'apport de différentes modification de la parole observées dans la parole claire sur l'intelligibilité.

Modifications de la pente spectrale et des modulations basse fréquence de l'enveloppe temporelle

KRAUSE et al. étudient l'influence de deux modifications naturelles observées dans la parole claire. Premièrement, l'aplatissement de la pente spectrale qu'ils obtiennent en amplifiant l'énergie spectrale au niveau des deuxièmes et troisièmes formants par l'utilisation d'une TFCT et de son inverse. Deuxièmement, l'amplification des modulations basse fréquence de l'enveloppe temporelle qu'ils obtiennent en appliquant un banc de filtres puis en amplifiant les composantes basse fréquence.

Au cours d'évaluations subjectives dans un bruit SSN, ils obtiennent une amélioration significative de l'intelligibilité pour l'aplatissement de la pente spectrale, avec un gain de 14 p.p.. En revanche, lorsque la modulation basse fréquence de l'enveloppe temporelle est appliquée, c'est une diminution significative de l'intelligibilité qui est observée. Comme pour la parole Lombard, l'aplatissement de la pente spectrale semble donc responsable d'une grande partie de l'amélioration de l'intelligibilité de la parole claire mais d'autres modifications prosodiques sont responsables du gain manquant [99].

Modification du débit de parole

On pourrait imaginer que le ralentissement important du débit de parole, et l'insertion de nombreuses pauses dans la parole claire contribuent fortement au gain d'intelligibilité de ce mode de parole dans le bruit. Plusieurs études portant sur l'influence des modifications temporelles de la parole claire n'ont pourtant pas obtenu ces conclusions.

PICHENY et al. proposent une approche uniforme en faisant correspondre la durée des stimuli de parole claire sur les stimuli de parole neutre correspondants, et vice-versa [155]. Dans les deux cas, une baisse de l'intelligibilité des signaux a été relevée pour des auditeurs atteints de déficience auditive. Puis, UCHANSKI et al. proposent une approche locale en insérant dans la parole neutre les pauses identifiées dans la parole claire et, inversement, en retirant ces pauses des stimuli de parole claire [208]. Dans les deux cas, une baisse de l'intelligibilité des signaux a été relevée pour des auditeurs atteints de déficience auditive et pour des auditeurs normo-entendants dans du bruit SSN. Enfin, KRAUSE et al. ont demandé à des locuteurs de parler clairement sous une contrainte de temps afin de conserver un débit comparable. Après quelques entraînements, la parole claire "rapide" résultante obtient des gains d'intelligibilité comparable à la parole claire "lente" pour des normo-entendants dans du bruit SSN [100].

Ainsi, si les modifications du débit de parole de la parole claire contribuent à l'amélioration de l'intelligibilité dans le bruit, cela provient de modifications plus complexes qu'un ralentissement uniforme ou que la simple insertion de pauses. De plus, l'obtention d'une parole claire "rapide" aussi performante qu'une parole claire "lente", avec de l'entraînement, atteste de l'importance des modifications spectrales pour le gain d'intelligibilité.

Modifications de l'espace vocalique

À notre connaissance, aucune étude spécifique concernant l'influence sur l'intelligibilité dans le bruit des modifications de l'espace vocalique par la parole claire, ou par la parole Lombard, n'a été menée. FERGUSON et al. ont mené une analyse de régression sur les résultats d'un test d'intelligibilité de voyelles dans un bruit de conversation (avec plusieurs voix concurrentes) [55]. Les résultats indiquent que les modifications de l'espace vocalique jouent un rôle significatif dans le gain d'intelligibilité de la parole claire dans le bruit.

Conclusion du chapitre 3

Dans ce chapitre, nous avons détaillé certaines stratégies naturelles mises en place afin d'améliorer l'intelligibilité de la parole. Nous nous sommes concentrés sur les deux modes de parole les plus étudiés, générant des gains d'intelligibilité significatifs dans divers environnements d'écoute bruités : la parole claire et la parole Lombard.

Les modifications introduites par ces deux modes de parole présentent une variabilité interpersonnelle très importante. Cependant, certaines modifications se démarquent par leur récurrence et le gain d'intelligibilité qu'elles semblent procurer, parfois validé par des études spécifiques. Pour la parole Lombard, c'est l'augmentation du fondamental, l'aplatissement de la pente spectrale, et l'augmentation du délai d'établissement du voisement, qui sont principalement considérés. Et pourtant, le changement de la pente spectrale est la seule modification pour laquelle un gain d'intelligibilité significatif a été relevé. Pour la parole claire, on retrouve aussi un aplatissement de la pente spectrale qui semblerait majoritairement responsable de l'amélioration de l'intelligibilité dans le bruit. Les modifications du débit de parole et les modifications de l'espace vocalique induisant une hyper-articulation sont les deux principaux candidats qui pourraient expliquer le gain restant de la parole claire. Des études plus approfondies sur ces deux types de modifications sont alors nécessaires pour confirmer ces suppositions.

Dans la suite de notre étude, nous proposerons des pistes d'adaptations et d'améliorations des méthodes de renforcement de la parole sous le spectre de notre contexte bruité automobile. Nous verrons que la grande majorité des modifications numériques, utilisées en renforcement de la parole dans le bruit, s'inspirent des stratégies naturelles introduites dans ce chapitre. Que ce soit les méthodes directes, qui seront abordées Partie 2, et plus particulièrement les méthodes paramétriques, qui seront abordées Partie 3.

Deuxième partie

Renforcement direct de la parole dans le bruit par maximisation exacte d’un critère d’intelligibilité sous contrainte énergétique pondérée

Chapitre 4

Approches actuelles de renforcement direct de la parole dans le bruit et suggestion d'une nouvelle contrainte énergétique

Sommaire

Retour à la table des matières

Introduction du chapitre 4

Le renforcement direct de la parole consiste à des traitements de signaux audio afin d'améliorer leur intelligibilité dans un environnement qui la dégrade. Il s'oppose au renforcement paramétrique de la parole qui se base sur l'utilisation de modèles d'analyse-modifications-synthèse afin d'extraire des paramètres spécifiques à la parole et de les modifier. En renforcement direct, les informations utilisées qui conditionnent le traitement peuvent cependant provenir d'outils d'analyse de la parole mais n'exploitent pas la re-synthèse.

L'utilisation de matériaux acoustiques absorbants, devenue habituelle chez les constructeurs automobiles, permet de minimiser grandement le temps de réverbération des signaux audio et de rendre ce facteur quasi-négligeable du point de vue de l'intelligibilité. La présence de bruit est alors le facteur principalement responsable de la dégradation de l'intelligibilité et ce sont donc les approches de renforcement de la parole dans le bruit qui nous intéressent. Dans l'étude de ces approches, il est quasi-systématique d'introduire une contrainte énergétique or notre environnement d'étude automobile est un exemple de cadre applicatif où la contrainte énergétique classique peut se révéler inadaptée. En effet, la conservation de l'énergie moyenne des signaux de parole ne prend pas en compte la sensibilité auditive de l'auditeur.

Dans ce chapitre, nous nous intéressons donc aux approches actuelles de renforcement direct de la parole dans le bruit et à leur limites concernant l'absence de prise en compte de la sensibilité auditive dans des cadres applicatifs tels que les habitacles automobiles. Dans la section 4.1, après avoir introduit l'intérêt et les limites de la contrainte énergétique classique basée sur la conservation de l'énergie globale des signaux de parole, nous proposons une nouvelle contrainte énergétique pondérée permettant de prendre en compte la sensibilité auditive. Ensuite, nous présentons les différentes approches actuelles de renforcement direct de la parole dans le bruit sous contrainte énergétique constante, avec un regard critique sur leur adaptation à la nouvelle contrainte proposée : nous introduirons d'abord, dans la section 4.2, les traitements qui sont indépendants du bruit dans lequel les signaux sont diffusés puis, dans la section 4.3, ceux qui le prennent en compte par une boucle de retour.

4.1 Contraintes énergétiques et proposition d'adaptation

Dans cette section, nous présenterons les raisons pour lesquelles une contrainte énergétique est généralement imposée en renforcement de la parole, puis, après avoir exposé les limites des contraintes actuelles, nous proposons une nouvelle contrainte énergétique pouvant être plus adaptée dans certains cadres applicatifs.

4.1.1 Contraintes énergétiques actuelles

Contraintes matérielles

Le matériel utilisé pour produire le signal acoustique de parole présente des contraintes techniques qu'il peut être important de prendre en compte lors du développement de méthodes de renforcement de la parole. En particulier pour des systèmes avec de petits haut-parleurs, des traitements trop brusques peuvent aisément provoquer de la saturation, ou une dégradation des actionneurs. Pour prévenir ces phénomènes, les algorithmes s'imposent des traitements sous contrainte énergétique. La contrainte énergétique la plus répandue en renforcement de la parole est de conserver l'énergie moyenne des signaux. Cependant, des contraintes plus adaptées peuvent être nécessaires en portant sur l'énergie instantanée par exemple, ou encore sur une limitation de l'énergie sur différentes bandes de fréquence.

Contraintes liées à l'auditeur

Au delà des contraintes matérielles, les contraintes liées à l'auditeur sont d'autant plus importantes du fait qu'une redistribution énergétique mal contrôlée peut dégrader le confort d'écoute, engendrer des douleurs, voir même provoquer des séquelles. La contrainte énergétique la plus utilisée reste de conserver l'énergie moyenne des signaux mais d'autres contraintes existent portant sur l'énergie instantanée, ou encore sur une limitation de l'énergie dans des bandes de fréquence sélectionnées, par exemple.

Contrainte comparative

Une dernière raison pour contraindre l'énergie des signaux est de permettre une comparaison des algorithmes indépendante du niveau de présentation. L'intelligibilité de la parole étant très fortement corrélée au niveau de présentation, ce facteur est systématiquement supprimé en normalisant l'énergie des signaux. Comme nous l'avons vu section 1.3.3, l'intelligibilité d'un signal est d'ailleurs généralement estimée comme le niveau nécessaire pour qu'un pourcentage donné des stimuli vocaux soit répété, avec le SRP. La contrainte énergétique utilisée consiste alors toujours à normaliser l'énergie moyenne des signaux de parole.

4.1.2 Contrainte perceptive nouvelle

Comme nous le verrons dans ce chapitre, la majorité des approches de renforcement de la parole sont basées sur de la redistribution spectro-temporelle et, indirectement, sur la concentration de l'énergie spectrale du signal dans des zones où l'oreille est plus sensible. Cela a pour conséquence d'augmenter l'intensité perçue du signal de parole sans modifier son énergie moyenne. Ces approches sont performantes et souvent bien adaptées pour des applications soumises à des contraintes matérielles. En revanche, pour certaines applications où l'auditeur maîtrise le volume de présentation du signal de parole qu'il règle à un niveau confortable d'écoute, une augmentation de l'intensité perçue le poussera à baisser le volume. On peut alors imaginer une succession d'augmentations du niveau perçu par l'algorithme, suivies de diminutions du niveau global par l'auditeur qui cherche à maintenir son niveau d'écoute confortable. Dans ces conditions, le traitement sera perturbé et le gain d'intelligibilité s'en retrouvera très fortement impacté.

Nous proposons donc d'introduire une contrainte énergétique comparative nouvelle consistant à conserver non plus l'énergie moyenne du signal traité mais une estimation de son niveau perçu. Cela consiste à maintenir l'énergie calculée avec la pondération dB(A), présentée section 1.1, qui permet alors de prendre en compte la sensibilité auditive. Étudier les performances des algorithmes de renforcement de la parole sous ce nouveau prisme nous semble alors très intéressant et permettrait de justifier l'intérêt des différents algorithmes futurs ou existants pour des applications soumises à cette problématique. La figure 4.1 rappelle la courbe de pondération de l'échelle perceptive en dB(A) que nous proposons d'utiliser.

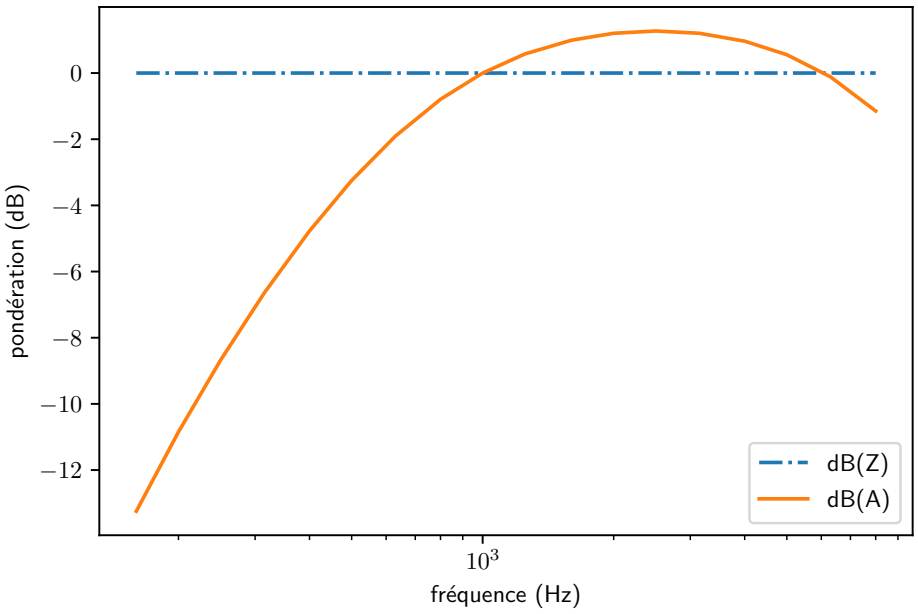


FIGURE 4.1 – Courbe de pondération en dB(A), avec comparaison à la courbe de pondération classique en dB(Z).

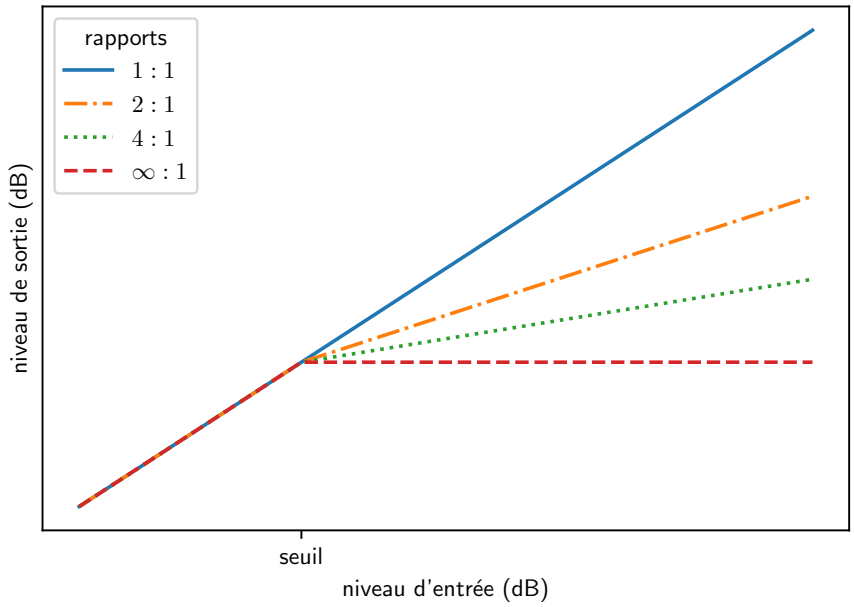


FIGURE 4.2 – Lois d'entrée-sortie d'une compression dynamique pour différents rapports de compression.

4.2 Traitements sans prise en compte du bruit

Les systèmes de renforcement de la parole dans un environnement bruyant ne sont pas forcément équipés de capteurs permettant d'avoir une estimation en temps réel du bruit. Ainsi, certains traitements se basent sur des connaissances générales de l'intelligibilité de la parole ou sur les caractéristiques des environnements bruyants. Dans cette section nous proposons de présenter ces différentes approches génériques.

4.2.1 Compression dynamique

La compression dynamique est historiquement la première approche qui a été utilisée en renforcement de la parole. Elle consiste à diminuer l'écart de niveau entre les sons de faible intensité, principalement les consonnes, et les sons de forte intensité, principalement les voyelles. Un compresseur est un amplificateur dont le gain varie selon la valeur du signal à son entrée et il est caractérisé par quatre principaux paramètres :

- le seuil de compression, amplitude à partir de laquelle le compresseur réduit le niveau du signal audio,
- le rapport de compression, facteur d'atténuation des échantillons qui dépassent le seuil de compression,
- le temps d'attaque, durée nécessaire pour que le compresseur s'active lorsque le seuil de compression est dépassé, un temps long permet de conserver des attaques plus naturelles mais accentue le risque d'atteindre des niveaux excessifs,
- le temps de retour, durée nécessaire pour que la compression se relâche lorsque le niveau repasse sous le seuil de compression.

Écrêtement

Ces méthodes ont d'abord été introduites en 1947 pour des applications militaires en appliquant un simple écrêtement [102], ou *peak clipping*. L'écrêtement est un cas particulier de la compression avec un rapport de compression de $(\infty : 1)$ couplé avec un temps d'attaque et un temps de retour très courts, c'est à dire qu'on introduit une saturation et le signal traité ne dépasse donc jamais son seuil de compression. Cette première approche était surtout utilisée pour les avantages qu'elle procure lors d'une communication par modulation d'amplitude

car en diminuant la dynamique du signal, les sons faibles sont plus robustes aux bruits de transmission. Cependant, il a été montré que même sans prendre en compte le mode de communication, appliquer un écrêtement améliore l'intelligibilité des signaux de parole [101, 202] et ainsi l'augmentation du rapport de niveau entre les consonnes et voyelles devint une piste importante en renforcement de la parole.

Limitation

Le problème majeur de l'écrêtement est qu'il introduit d'importantes distorsions dans le signal de parole qui ont tendance à se situer dans la zone où se situe les seconds formants dont l'importance pour l'intelligibilité est avérée.Pour réduire l'apparition de ces distorsions, il a été proposé d'utiliser un limiteur [101, 147] qui est un compresseur avec un rapport de compression important, généralement supérieur à (10 : 1), et un temps d'attaque souvent court, cela ressemble donc à un écrêteur mais son temps de retour étant supérieur il permet de conserver l'allure générale du signal traité, d'éviter les effets de saturation et donc de minimiser l'apparition de distorsions.

Compression multibandes

La compression multibandes consiste à appliquer une compression, avec des réglages spécifiques, sur plusieurs bandes de fréquence prédéfinies. Les prothèses auditives se sont vite vues équipées de ce type de traitement permettant des réglages fins de la compression en fonction de la déficience auditive de l'utilisateur. Cependant, de nombreuses études sur le sujet se contredisent quant à l'efficacité de la compression multibandes dans les prothèses auditives. Une synthèse détaillée de l'influence de nombreux facteurs sur l'efficacité de la compression dynamique dans les prothèses auditives a été dressée en 2010 par KATES [88]. Les différences de résultats entre les études s'expliquent par des conditions expérimentales trop variées et l'utilisation de la compression multibandes dans les prothèses auditives nécessite un réglage adaptatif, e.g. à la déficience de l'utilisateur, au volume du signal et au volume du bruit, pour que le traitement soit efficace. Pour une application destinée aux normo-entendants, les tentatives d'utiliser la compression multibandes n'est pour l'instant pas parvenu à améliorer l'intelligibilité des signaux de parole, au contraire cela semble plutôt la détériorer [77].

Compression spécifique

La loi entrée/sortie peut prendre des formes plus complexes définies par une courbe intitulée *Input-Output Envelope Characteristic* (IOEC). Cette courbe est généralement continue, définie par morceaux avec un facteur de compression spécifique associé à chaque morceau. On peut observer figure 4.3 un exemple de courbe IOEC proposée par ZORILA et al. [221]. En proposant une plus grande flexibilité qu'une compression classique, l'utilisation d'une compression spécifique basée sur une courbe IOEC permet d'adapter la compression dynamique à différents besoins et procure un gain d'intelligibilité notable, il est donc commun d'ajouter un tel module en bout de chaîne d'un traitement visant à améliorer l'intelligibilité de la parole en environnement bruyant [167].

Ré-allocation temporelle spécifique à la parole

La compression dynamique est par définition un amplificateur dont le gain varie en fonction du niveau instantané du signal. Certaines approches s'exemptent de ces restrictions et proposent une ré-allocation temporelle de l'énergie sur la base d'une analyse du signal.

Une idée largement exploitée est, par exemple, de diminuer le rapport de niveau entre les segments voisés et les segments non voisés [181, 79]. Ces approches procurent des gains d'intelligibilité significatifs mais l'absence de comparaisons directes avec les méthodes classiques de compression dynamique ne permet pas de quantifier l'intérêt potentiel. D'autant plus que le voisement étant fortement corrélé à l'énergie, nous pouvons nous attendre à des traitements relativement proches.

En revanche, l'idée d'une compression non plus seulement basée sur le niveau du signal, mais sur une analyse préalable le segmentant en zones à rehausser ou non, est à l'origine de nouvelles propositions de traitement.

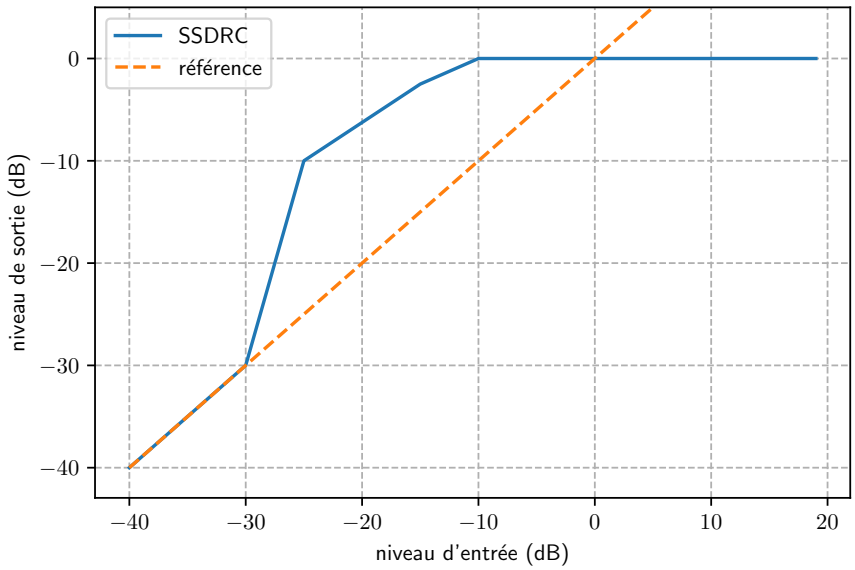


FIGURE 4.3 – Loi d'entrée-sortie de la compression dynamique proposée par ZORILA et al. dans l'algorithme SSDRC.

Les transitoires ayant un rôle majeur dans l'intelligibilité de la parole, d'autres approches cherchent alors à rehausser les transitoires plutôt que des zones au contenu fréquentiel stationnaire [220, 198, 161]. Des améliorations significatives d'intelligibilité ont été mesurées pour ces méthodes avec des gains particulièrement intéressants dans des conditions d'écoute sévèrement bruitées.

Sensibilité de la compression dynamique à la nouvelle contrainte perceptive

Les approches par compression dynamique visent alors à rendre plus audibles les segments de faible intensité, dont la contribution à l'intelligibilité est importante, au détriment des segments à forte énergie. Nous savons que les sons non-voisés, susceptibles d'être amplifiés, ont une densité spectrale plus importante dans les zones sensibles de l'oreille que les sons voisés, susceptibles d'être diminués. Nous pouvons observer, sur la figure 4.4, des spectres long-terme en fonction du voisement obtenu sur une base de données de parole arbitraire de 10 locuteurs (5 hommes, 5 femmes). Même si les sons non-voisés sont aussi moins représentés, on s'attend à ce que l'utilisation de compression dynamique augmente alors l'intensité perçue.

Pour avoir une meilleure idée de l'ordre de grandeur de cette augmentation, nous avons appliqué une importante compression dynamique via un écrêtement avec un seuil de compression égal à 30% de l'énergie maximale des signaux de la base de données. Après avoir appliqué l'écrêtement sur chaque signal et normalisé au niveau moyen original de 65 dB(Z), nous remarquons une augmentation du niveau perçu de moins d'1 dB(A). Ce résultat, ainsi que ceux des analyses suivantes, sont consultables dans le tableau 4.1.

Traitement	Aucun	Compression dynamique	Pré-accentuation	Amélioration du SII
Niveau perçu	60,9 dB(A)	61,8 dB(A)	62,1 dB(A)	63,4 dB(A)

TABLEAU 4.1 – Évolutions du niveau perçu en dB(A), après l'application d'approches simplifiées de renforcement direct de la parole, puis normalisé au niveau d'origine à 65 dB(Z).

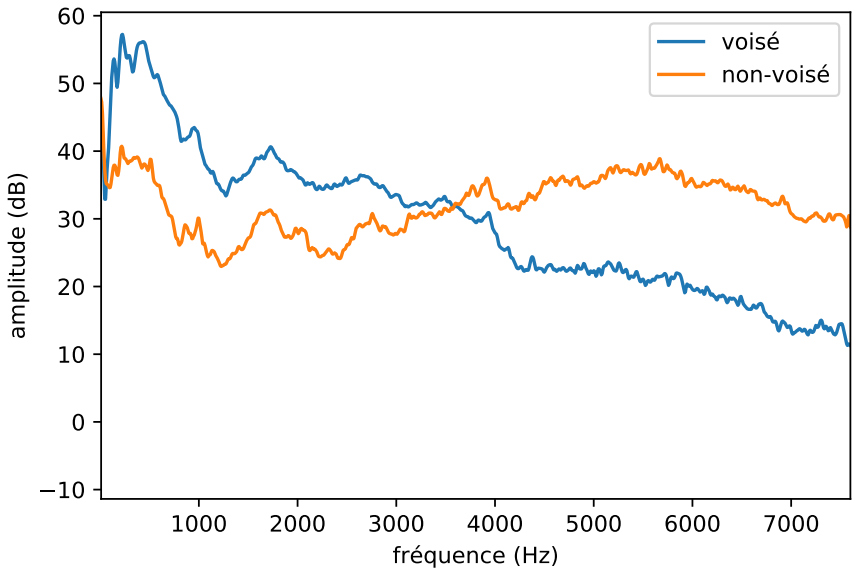


FIGURE 4.4 – Comparaison entre le spectre long-terme des segments voisés et non-voisés de signaux de parole.

4.2.2 Méthodes de filtrage : rehaussement des formants

Une autre façon de procéder est d'utiliser des techniques de filtrage permettant des adaptations fréquentielles du signal de parole. Le filtrage peut être stationnaire en cherchant à manipuler des fréquences dont la corrélation avec l'intelligibilité est connue, ou il peut être variable en étant basé sur une analyse du signal qui conditionne alors le filtrage à appliquer à différents instants.

Filtrage stationnaire

L'importance du second formant vis-à-vis de l'intelligibilité de la parole étant avérée, un filtre passe-haut, dit de pré-accentuation, est alors souvent appliqué au signal en renforcement de la parole [202, 147, 35, 79] car cela permet de rehausser le deuxième formant par rapport au premier formant et donc d'améliorer significativement l'intelligibilité des signaux de parole dans le bruit. Plus récemment HALL et al. comparent l'utilisation d'un filtre de pré-accentuation du premier ordre avec un filtre de "normalisation formantique" consistant à appliquer l'inverse du spectre d'amplitude moyen des formants d'un locuteur de référence [74]. Les deux approches obtiennent des améliorations d'intelligibilité similaires dans un bruit de conversation (avec plusieurs voix concurrentes) mais la qualité de la normalisation formantique est préférée par les sujets.

Filtrage variable

En suivant l'évolution des formants, il est aussi possible d'utiliser un filtrage qui s'adapte en fonction de ceux-ci. C'est ce que propose JOKINEN et al. en suivant l'évolution des formants afin d'utiliser un filtrage rehaussant les seconds formants de manière adaptative [82]. Aucune différence significative n'a cependant été observée entre l'utilisation d'un filtrage adaptatif et un filtrage fixe, la raison évoquée pourrait être liée à la simplicité d'estimation des formants résultant en un traitement moyen qui n'apporte rien de plus qu'un traitement fixe bien paramétré. Une manipulation plus précise des formants sera introduite dans le chapitre 8 concernant les méthodes de renforcement paramétrique de la parole.

Dans leur algorithme SSDRC [221], ZORILA et al. proposent de rehausser tous les formants en utilisant un filtre adaptatif calculé à partir de l'enveloppe spectrale à laquelle on soustrait la pente spectrale [159]. Le gain d'intelligibilité de cet algorithme est très significatif et parfois même plus important que le Lombard naturel [45], mais le rehaussement des formants étant appliqué en combinaison avec un filtre de pré-accentuation et une

compression dynamique, il est difficile de quantifier l'apport du module de filtrage adaptatif. SSDRC se place actuellement comme la référence des algorithmes de renforcement directe de la parole sans prise en compte du bruit et ses modules sont souvent utilisés tels quels dans d'autres travaux [69, 167].

Sensibilité des filtrages à la nouvelle contrainte perceptive

L'utilisation de filtrage visant à rehausser les formants, et plus particulièrement les seconds formants qui sont situés dans des zones sensibles de l'oreille, risque encore d'augmenter l'intensité perçue du signal de parole. Pour avoir une meilleure idée de l'ordre de grandeur de cette augmentation, nous avons étudié l'effet d'un simple filtre de pré-accentuation comportant un gain de 6dB/octave à partir de 1,1kHz. Après avoir appliqué le filtrage sur chaque signal, issu de la même base de données que celle utilisée dans la section précédente, et normalisé au niveau moyen original de 65 dB(Z), nous remarquons une augmentation du niveau perçu d'environ 1 dB(A). Ce résultat, ainsi que ceux des autres analyses, est consultable dans le tableau 4.1.

4.3 Traitements avec prise en compte du bruit

Un des principaux facteurs responsables de la baisse d'intelligibilité dans du bruit est le masquage énergétique. Les méthodes introduites dans la section précédente vont généralement chercher à faire émerger des éléments du signal de parole importants vis-à-vis de l'intelligibilité, e.g. les transitoires et certains formants. Sans informations précises sur le bruit dans lequel le signal est diffusé, il est difficile de savoir si l'amplitude des modifications est suffisante pour surmonter le masquage. Si au contraire une boucle de retour permet d'avoir une estimation du bruit moyen ou instantané, ces traitements peuvent être adaptés afin de paramétrer efficacement les modifications introduites sur le signal de parole.

4.3.1 Compression dynamique et filtrage adaptatifs au bruit

Des travaux proposent d'appliquer les méthodes de ré-allocation spectro-temporelle de l'énergie en prenant en compte l'estimation du bruit.

Étude de différentes stratégies de compression dynamique adaptative au bruit

TANG et al. proposent une étude complète de différentes stratégies de ré-allocation spectro-temporelle, sans augmentation de l'énergie totale, appuyée par des mesures objectives et subjectives d'intelligibilité [194, 196]. Dans cette étude les stratégies cherchent toutes à manipuler différents aspects du RSB du signal. La liste des stratégies et leurs principes sont les suivants :

- SegSNR décompose le signal en segments de 50 ms avec un taux de recouvrement de 50% et fixe leur RSB local à la valeur du RSB global,
- ChanSNR utilise un banc de filtres Gammatone et fixe le RSB des 55 canaux à la valeur du RSB global,
- LocalSNR combine les méthodes SegSNR et ChanSNR en fixant le RSB local de chaque canal fréquentiel à la valeur du RSB global,
- SelectBoost amplifie de 20 dB le RSB local des canaux fréquentiels entre 1800 et 7500 Hz lorsque celui-ci est inférieur à 5 dB.

Deux types de bruits ont été utilisés pour les tests d'intelligibilité à savoir un bruit stationnaire (SSN), et un bruit fluctuant (SMN). Pour différents niveaux de présentation des stimuli, les trois premières méthodes peinent à améliorer l'intelligibilité en particulier pour le bruit SMN pour lequel elles auront plutôt tendance à la dégrader. La méthode SelectBoost, au contraire, procure un net gain d'intelligibilité dans toutes les situations. Les conclusions de l'étude suggèrent que chercher à maintenir un certain RSB au niveau spectral et/ou temporel ne serait pas bénéfique vis-à-vis de l'intelligibilité. En effet, cela provoque un étalement de l'énergie permettant au bruit de dégrader encore plus de zones spectro-temporelles qui étaient épargnées jusque là, en particulier pour de

faibles RSB. En opposition, un traitement plus spécifique redistribuant l'énergie de manière plus éclatée mais localisée sur des aspects du signal importants pour l'intelligibilité serait bien plus adapté comme observé avec SelectBoost.

Filtrage : rehaussement des formants adaptatif au bruit

Le rehaussement formantique introduit précédemment peut aussi être adaptatif au bruit, c'est ce que propose BROUCKXON et al. en introduisant un filtrage aux gains variables permettant d'assurer un certain RSB pour les trois premiers formants [30]. Des tests perceptifs montrent un gain d'intelligibilité du même ordre de grandeur que pour les approches non-adaptatives au bruit. Une étude commune serait tout de même nécessaire pour tirer des conclusions sur l'intérêt du traitement adaptatif.

Sensibilité à la nouvelle contrainte perceptive

L'utilisation de la compression dynamique, ou du filtrage visant à rehausser les formants, dans le cadre de la nouvelle contrainte perceptive et les potentielles augmentations de l'intensité perçue associées ont été traitées dans la section 4.2. On peut s'attendre à un ordre de grandeur équivalent, inférieur à 1 dB(A), pour des traitements similaires adaptatifs au bruit.

4.3.2 Maximisation d'un critère d'intelligibilité

Principe

Les critères d'intelligibilité de la parole ont été introduits dans la section 1.4, une approche appréciée en renforcement direct de la parole consiste à manipuler les signaux afin de maximiser un de ces critères. En effet, les modifications proposées jusqu'ici sont basées sur des observations liées à l'intelligibilité de la parole et il peut être intéressant d'utiliser des mesures objectives d'intelligibilité afin de paramétrer finement ces modifications. C'est ce que propose de nombreuses approches qui conditionnent les modifications à appliquer en s'appuyant sur des mesures objectives de l'intelligibilité de la parole déjà existantes comme l'AI [160] ou le SII [170, 193, 183, 167], mais aussi des mesures moins populaires [195] voir mises en place pour l'occasion [192]. Le fait d'utiliser un critère complexe prenant en compte de nombreux facteurs perceptifs permet de raffiner, mais aussi de justifier, le paramétrage parfois arbitraire de certaines approches. Il est toutefois important de noter que la richesse des mesures objectives d'intelligibilité engendre une complexité mathématique importante dans leur calcul, c'est pourquoi ce type d'approche cherche généralement à résoudre le problème d'optimisation de manière approchée en travaillant sur des approximations des mesures.

Sensibilité de la maximisation d'un critère d'intelligibilité à la nouvelle contrainte perceptive

Le SII, qui est une mesure de référence concernant l'intelligibilité de signaux de parole dans le bruit et sur laquelle de nombreux travaux de maximisation ont déjà eu lieu, consiste à mesurer l'audibilité des indices de la parole contribuant à son intelligibilité dans différents canaux fréquentiels. Une description détaillée de ce critère sera effectuée chapitre 5, cependant comme nous l'avons déjà vu section 1.4.1, une FIB associe des coefficients de pondération à chaque canal en fonction de leur importance. Ainsi, la maximisation de ce critère poussera à concentrer l'énergie du signal de parole dans ces bandes fréquentielles d'importance qui, encore une fois, se situent dans des zones sensibles de l'oreille. On peut alors s'attendre à une augmentation du niveau perçu, d'autant plus importante du fait que cette approche vise directement à augmenter l'audibilité du signal de parole.

Pour avoir une meilleure idée de l'ordre de grandeur de cette augmentation, nous avons étudié l'influence d'une égalisation fréquentielle avec des gains proportionnels aux coefficients d'une FIB du SII. Après avoir appliqué l'égalisation sur chaque signal, de la même base de données utilisée dans les sections précédentes, et normalisé au niveau moyen original de 65 dB(Z), nous remarquons une augmentation du niveau perçu de plus de 2 dB(A). Ce résultat, ainsi que ceux des analyses précédentes, sont consultables dans le tableau 4.1.

Conclusion du chapitre 4

Dans ce chapitre, nous avons mis en évidence une limitation de la contrainte énergétique classique, utilisée en renforcement de la parole, consistant à maintenir constante l'énergie moyenne des signaux de parole, pour certaines applications où l'auditeur a accès au niveau de présentation. Après avoir proposé une nouvelle contrainte énergétique basée sur une échelle perceptive, nous avons présenté les principales approches actuelles de renforcement direct de la parole dans le bruit.

En analysant succinctement l'impact que pourrait avoir cette nouvelle contrainte perceptive sur leur performances vis-à-vis de l'amélioration de l'intelligibilité, les approches basées sur la maximisation d'un critère d'intelligibilité semblent les plus sensibles à cette nouvelle contrainte. De plus, les méthodes de résolutions actuelles du problème de maximisation de certaines mesures d'intelligibilité, faisant face à une complexité importante, se basent systématiquement sur des approximations pour simplifier la procédure.

Dans la suite de cette partie, nous proposons donc d'étudier l'influence de la nouvelle contrainte perceptive sur une méthode de maximisation d'un critère d'intelligibilité. Pour cela nous travaillerons sur le SII qui est une mesure de référence concernant l'intelligibilité de signaux de parole dans le bruit et sur laquelle de nombreux travaux de maximisation ont déjà eu lieu. Pour s'assurer de l'exploitation maximale du potentiel de cette méthode, nous proposerons aussi un protocole de maximisation exacte du critère qui, jusqu'à maintenant, était toujours basé sur une approximation de celui-ci.

Chapitre 5

Proposition de maximisation exacte d'un critère d'intelligibilité sous contrainte énergétique pondérée

Sommaire

Retour à la table des matières

Introduction du chapitre 5

Comme nous l'avons vu dans le chapitre précédent, la maximisation d'un critère d'intelligibilité est une approche populaire en renforcement direct de la parole dans le bruit. Des mesures objectives reconnues, basées sur des études perceptives poussées de l'intelligibilité de la parole, servent alors à justifier et affiner le paramétrage de redistributions spectro-temporelles des signaux afin de les rendre plus intelligibles dans un environnement bruyant. La mesure la plus utilisée à cet effet est le SII [170, 193, 183, 167] et le gain d'intelligibilité procuré par sa maximisation est très significatif. Cependant, nous pensons que deux aspects de ces travaux méritent un approfondissement.

Premièrement, pour proposer une modélisation précise de l'intelligibilité des signaux de parole dans le bruit, le calcul du SII est complexe et présente de nombreuses non-linéarités. C'est pourquoi toutes les tentatives actuelles de maximisation de ce critère sont basées sur des approximations de ce dernier. En fonction des paramètres d'étude comme le type de bruit, ou le RSB, les résultats par approximation peuvent être plus ou moins proches de la solution réelle. Ainsi, nous proposons de poser formellement le problème d'optimisation et de le résoudre de façon exacte afin d'étudier les effets des différentes simplifications sur la maximisation du SII.

Deuxièmement, comme introduit chapitre 4, l'utilisation d'une contrainte énergétique nouvelle basée sur une échelle perceptive peut être nécessaire pour certaines applications, notamment lorsque l'auditeur a accès au niveau de présentation des stimuli. En effet, comme nous le verrons dans cette section, la maximisation du SII s'accompagne d'une concentration de l'énergie spectrale dans les zones de sensibilité auditive ce qui augmente inéluctablement le niveau perçu des signaux de parole traités. Il est donc légitime de se demander comment la maximisation de cette mesure se comporte lorsque la contrainte énergétique ne porte plus sur une conservation de l'énergie totale du signal mais sur une conservation de l'énergie perçue.

Dans la section 5.1, nous présenterons le SII, de son principe à son calcul mathématique détaillé. Cela nous permettra ensuite d'expliquer le protocole d'optimisation exacte du critère sous contrainte énergétique pondérée dans la section 5.2. Puis, nous présenterons, dans la section 5.3, les procédures actuelles d'optimisation du SII par approximation, ainsi que les adaptations nécessaires à leur utilisation sous la nouvelle contrainte énergétique. Enfin, dans la section 5.4, nous présenterons et analyserons les résultats d'optimisation obtenus sur trois bruits classiques sous l'ancienne, et la nouvelle, contrainte énergétique, nous étudierons le comportement des approximations et nous verrons comment exploiter les résultats afin de traiter efficacement les signaux de parole.

5.1 Présentation du critère : SII

Les mesures objectives de l'intelligibilité de la parole sont basées sur de solides connaissances empiriques et de nombreuses hypothèses. L'hypothèse principale du SII est que la parole est composée de canaux fréquentiels qui sont porteurs d'informations indépendantes. La norme *American National Standards Institute* (ANSI)/*American Standards Association* (ASA) S3.5 [7] prévoit le calcul du critère pour différentes décompositions en i^{max} canaux de fréquences centrales F_i :

- en bandes d'octaves, avec $i^{max} = 6$,
- en bandes de tiers d'octaves, avec $i^{max} = 18$,
- en bandes critiques, avec $i^{max} = 21$.

Les décompositions en bandes de tiers d'octaves et bandes critiques sont très proches et plus fines qu'en bandes d'octaves. Les bandes de tiers d'octaves sont les plus utilisées dans les études similaires, c'est donc cette décomposition que nous avons choisie dans notre étude et les fréquences caractéristiques utilisées sont notées dans le tableau 5.1. Cependant, les raisonnements qui vont suivre sont tout à fait applicables aux autres décompositions.

Le SII est calculé à partir des niveaux du spectre équivalent de parole E_i, et des niveaux du spectre équivalent de bruit N_i, dans chaque bande et en décibels (dB). Ces niveaux s'obtiennent en intégrant le périodogramme de chaque signal sur leurs canaux respectifs et en normalisant par la largeur de bande associée b_i. Les niveaux des spectres équivalents servent alors au calcul des coefficients d'audibilité et de distorsion introduits ci-après.

F_i (Hz)		160	200	250	315	400	500	630	800	1000	
bornes (Hz)	141	178	224	282	355	447	562	708	891	1122	
b_i (Hz)		37	46	58	73	92	115	146	183	231	
H_i (dB)		-13,2	-10,8	-8,7	-6,6	-4,8	-3,2	-1,9	-0,7	0,0	
F_i (Hz)		1250	1600	2000	2500	3150	4000	5000	6300	8000	
bornes (Hz)	1122	1413	1778	2239	2818	3548	4467	5623	7079	8913	
b_i (Hz)		291	365	461	579	730	919	1156	1456	1834	
H_i (dB)		0,6	1,0	1,2	1,3	1,2	1,0	0,6	-0,1	-1,1	

TABLEAU 5.1 – Fréquences caractéristiques de la décomposition en bandes de tiers d'octaves avec les fréquences centrales F_i, les bornes correspondantes et les largeurs de bande b_i. Les H_i correspondent aux pondérations physiologique A.

5.1.1 Coefficients d'audibilité

Les coefficients d'audibilité représentent la proportion du spectre audible au-dessus des diverses perturbations qui impactent l'intelligibilité. Ils nécessitent un calcul préalable des niveaux D_i du spectre équivalent de perturbation donné par l'équation suivante :

$$D_i = max(T_i, Z_i)\,, \tag{5.1}$$

avec T_i les seuils d'audibilité de l'auditeur et Z_i les niveaux du spectre équivalent de masquage. Les Z_i prennent en compte l'étalement du masquage et se calculent en appliquant l'équation suivante, fournie par la norme ANSI/ASA S3.5 [7] :

$$Z_i(\beta_i) = 10\cdot\log(10^{N_i/10} + \sum_{j<i} 10^{(\beta_j+3{,}32\cdot\alpha_j\cdot\log(0{,}89\cdot F_i/F_j))/10})\,, \tag{5.2}$$

les coefficients β_i permettent de sélectionner la grandeur qui, du bruit ou de la parole elle-même, est responsable du masquage dans chaque canal à savoir :

$$\beta_i = \max(E_i - 24\,\text{dB}, N_i) \tag{5.3}$$

et les coefficients α_i conditionnent l'étalement du masquage et se calculent de la façon suivante :

$$\alpha_i = -80\,\text{dB} + 0{,}6\cdot(\beta_i + 10\cdot\log(F_i) - 6{,}353\,\text{dB})\,. \tag{5.4}$$

Notons que les seuils d'audibilité d'un auditeur normo-entendant dans un environnement bruyant sont souvent inférieurs aux niveaux de masquage i.e. $T_i \leq Z_i$ ainsi le spectre équivalent de perturbation est généralement confondu avec le spectre équivalent de masquage mais pas nécessairement.

On remarque que, par l'intermédiaire des coefficients β_i, les Z_i peuvent dépendre des E_i ce qui compliquera grandement la résolution du problème d'optimisation. Nous posons alors l'hypothèse suivante :

Hypothèse 1 *Les niveaux du spectre équivalent de bruit sont supérieurs aux niveaux du spectre équivalent d'auto-masquage :*

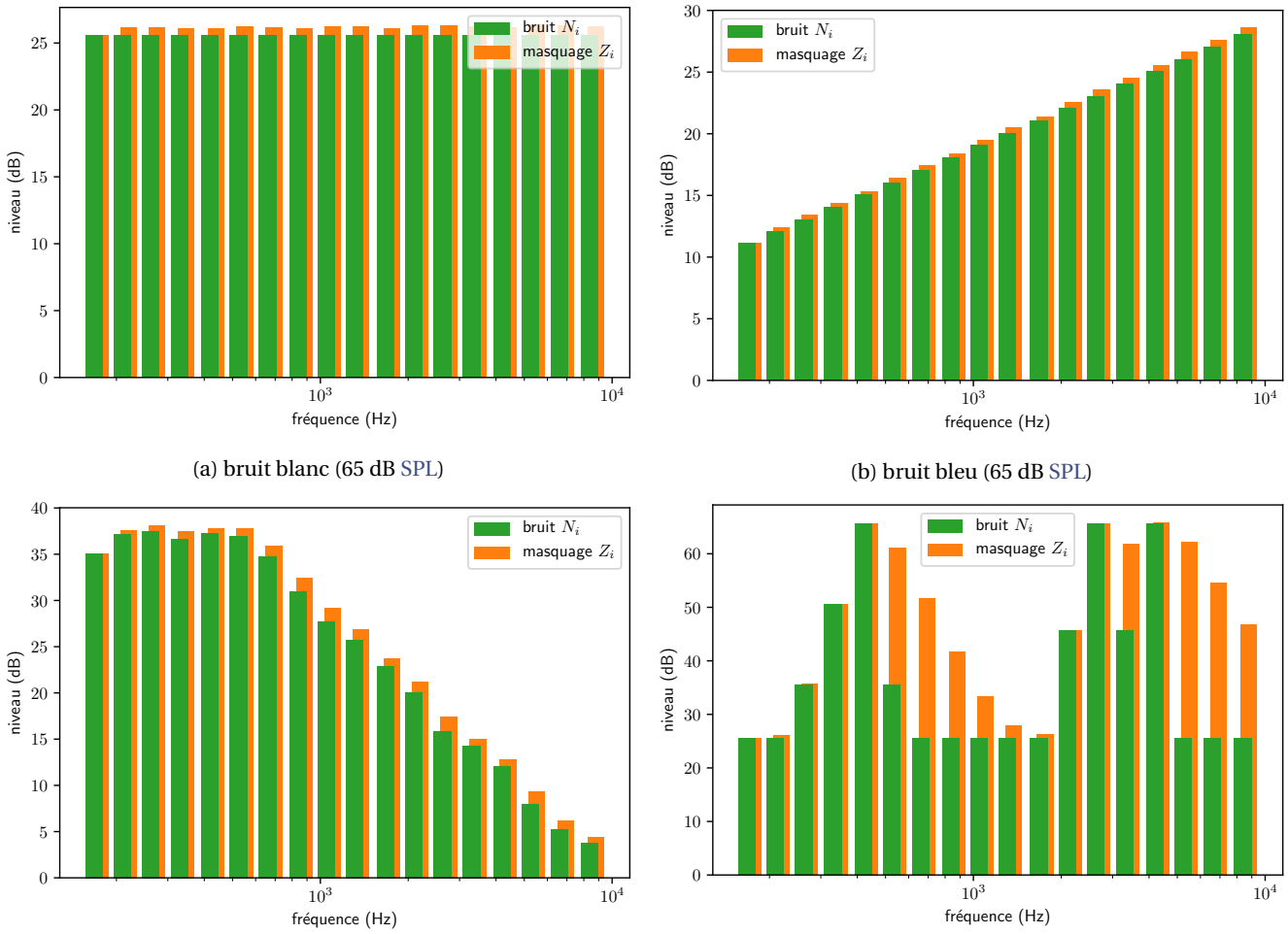
$$N_i \geq E_i - 24\,\text{dB}\,. \tag{5.5}$$

Si cette hypothèse est vérifiée, d'après l'équation 5.3 on a $\beta_i = N_i$, ainsi les Z_i se calculent uniquement à partir de l'ensemble $\{N_j\}_{j\leq i}$ et ne dépendent que du bruit. Dans notre étude, on considérera cette hypothèse vérifiée et ainsi, les niveaux D_i du spectre équivalent de perturbation sont fixés pour un auditeur et un bruit donnés. Nous reviendrons plus tard sur les conditions de validation de cette hypothèse.

Des exemples de spectres équivalents de bruits et leurs spectres équivalents de masquage associés sont visibles sur la figure 5.1. Les quatre bruits proposés sont les suivants :

- un bruit blanc, dont les niveaux du spectre équivalent sont constants,
- un bruit bleu, dont les niveaux du spectre équivalent augmentent de 3 dB/octave,
- un bruit de type SSN, dont la densité spectrale correspond à celle d'un signal de parole typique proposé par la norme ANSI/ASA S3.5,
- un bruit synthétique, dont la densité spectrale a été créée manuellement afin d'observer plus clairement l'étalement du masquage.

Les trois premiers bruits classiques seront intéressants à étudier car, en plus d'être utilisés dans la majorité des tests d'intelligibilité, ils possèdent des densités spectrales complémentaires.



(a) bruit blanc (65 dB SPL)

(b) bruit bleu (65 dB SPL)

(c) bruit SSN (65 dB SPL)

(d) bruit synthétique (niveau arbitraire)

FIGURE 5.1 – Spectres équivalents et spectres de masquage correspondants pour trois bruits classiques (blanc, bleu et SSN) et pour un quatrième bruit synthétique afin d'observer plus clairement l'étalement du masquage.

Les coefficients d'audibilité A_i sont alors calculés comme indiqué par l'équation suivante :

$$A_i(E_i, D_i) = \min(\max(\frac{E_i - (D_i - 15\,\text{dB})}{30\,\text{dB}}, 0), 1)\,. \tag{5.6}$$

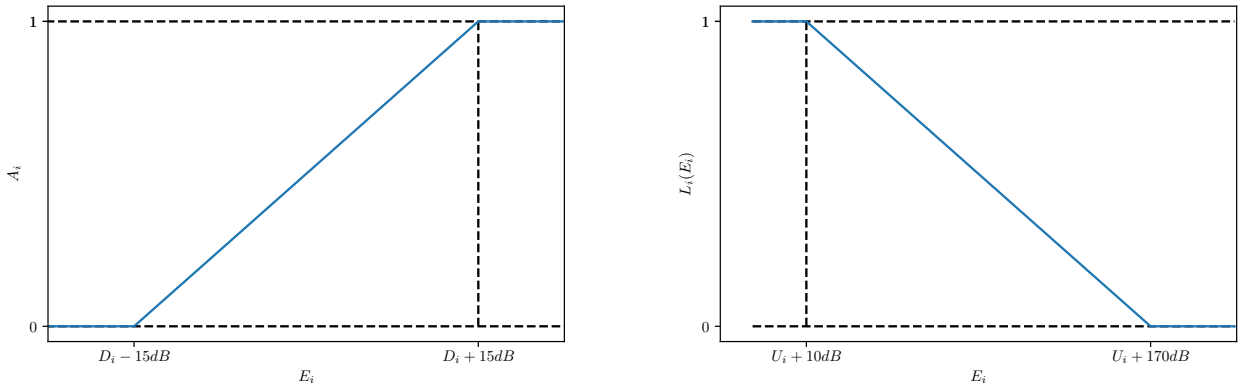
Le tracé des coefficients d'audibilité en fonction des niveaux E_i du spectre équivalent est visible sur la figure 5.2a. Une bande ne participe donc pas à l'intelligibilité d'un signal lorsque son RSP est inférieur à −15 dB car on considère que le masquage la rend inaudible. Sa participation en terme d'audibilité augmente ensuite linéairement avec le RSP entre −15 dB et 15 dB. Au delà de 15 dB, on considère que la bande est suffisamment audible et son augmentation ne contribue pas plus à l'intelligibilité du signal.

5.1.2 Coefficients de distorsion

Les coefficients de distorsion L_i prennent en compte la distorsion introduite lorsque les niveaux par bande s'éloignent trop des niveaux U_i d'un spectre équivalent de parole de référence fournit dans la norme ANSI/ASA S3.5 [7] et visible sur la figure 5.3a. Ils sont calculés comme indiqué par l'équation suivante :

$$L_i(E_i) = \min(1 - \frac{E_i - (U_i + 10\,\text{dB})}{160\,\text{dB}}, 1)\,. \tag{5.7}$$

Le tracé des coefficients de distorsion en fonction des niveaux E_i du spectre équivalent est visible sur la figure 5.2b. Pour des niveaux supérieurs de plus de 10 dB par rapport aux niveaux U_i de référence, on considère qu'une bande introduit de la distorsion qui détériore l'intelligibilité du signal. Cette réduction est linéaire jusqu'à 170 dB de plus que la référence. Ces niveaux ne seront, bien entendu, jamais atteints car bien trop nocifs pour l'audition.



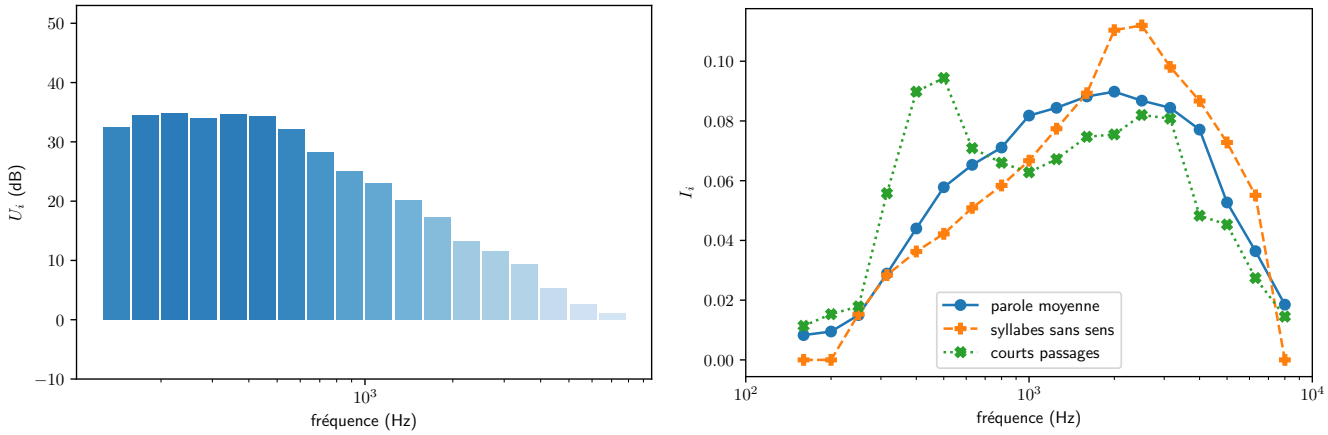
(a) coefficients d'audibilité

(b) coefficients de distorsion

FIGURE 5.2 – Évolutions des coefficients d'audibilités A_i, et de distorsion L_i, en fonction des niveaux E_i du spectre équivalent de parole pour un bruit donné et donc pour des niveaux D_i du spectre équivalent de perturbation fixés

5.1.3 Fonction d'importance de bande

Toutes les bandes ne contiennent pas la même quantité d'information relative à la parole, elles n'ont donc pas la même importance vis-à-vis de l'intelligibilité. Ainsi, une FIB, dont les coefficients sont notés I_i, est appliquée pour pondérer chaque bande. Plusieurs FIB sont mises à disposition dans la norme ANSI/ASA S3.5 [7] en fonction des stimuli verbaux utilisé. Ces fonctions ont été ajustées à partir de tests subjectifs d'intelligibilité dont le principe a été présenté chapitre 1. Quelques FIB sont présentées sur la figure 5.3b et on remarque qu'elles sont globalement proches des courbes de sensibilité auditive. Dans la suite de notre étude nous travaillerons exclusivement avec la FIB intitulée "parole moyenne" qui est la plus utilisée dans les travaux existants mais ce choix arbitraire n'influencera pas le raisonnement.



(a) spectre équivalent de parole de référence

(b) exemples de FIB

FIGURE 5.3 – Niveaux U_i du spectre équivalent de parole de référence et quelques FIB, fournis par la norme ANSI/ASA S3.5.

5.1.4 Formule du SII

La formule du SII correspond alors à une somme pondérée de ces différents facteurs dans chaque bande :

$$\mathrm{SII}(\{E_i\},\{D_i\}) = \sum_{i=1}^{i^{max}} I_i \cdot A_i(E_i,D_i) \cdot L_i(E_i) = \sum_{i=1}^{i^{max}} f_i(E_i,D_i)\,. \tag{5.8}$$

$$\text{avec} \quad f_i(E_i,D_i) = I_i \cdot A_i(E_i,D_i) \cdot L_i(E_i)\,. \tag{5.9}$$

La figure 5.4 montre l'allure des fonctions $f_i(\cdot,D_i)$ pour une bande dans trois situations :

- $U_i + 10\,\mathrm{dB} \le D_i - 15\,\mathrm{dB} \iff U_i \le D_i - 25\,\mathrm{dB}$, présence de bruit très importante,
- $D_i - 15\,\mathrm{dB} < U_i + 10\,\mathrm{dB} < D_i + 15\,\mathrm{dB} \iff D_i - 25\,\mathrm{dB} < U_i < D_i + 5\,\mathrm{dB}$, présence de bruit modérée,
- $U_i + 10\,\mathrm{dB} \ge D_i + 15\,\mathrm{dB} \iff U_i \ge D_i + 5\,\mathrm{dB}$, faible présence de bruit.

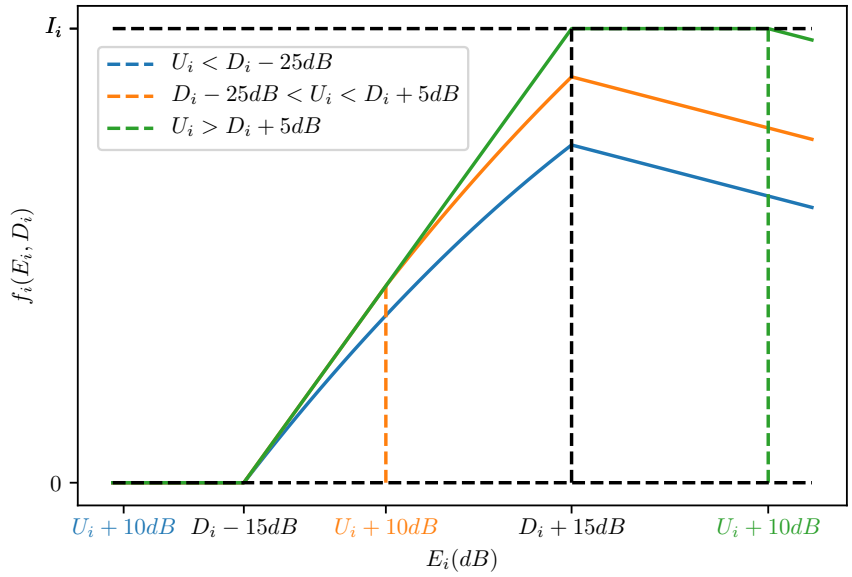


FIGURE 5.4 – Évolution des fonctions f_i en fonction des niveaux E_i pour trois situations distinctes.

5.2 Protocole d'optimisation exacte sous contrainte énergétique pondérée

Dans cette sous-section, nous posons le problème d'optimisation du SII avec une contrainte énergétique originale basée sur une pondération donnée.

5.2.1 Définition du problème

L'objectif de l'optimisation est de trouver les niveaux optimaux E_i^{opt} qui maximisent le SII pour un bruit donné, sans augmenter l'énergie d'un signal. Afin de prendre en compte la sensibilité auditive de l'auditeur, nous proposons d'utiliser une échelle de puissance adaptée, la pondération (A). Le choix de cette échelle a été justifié chapitre 4, cependant le raisonnement mathématique qui suit ne dépend pas de la valeur des coefficients de pondération, l'optimisation pourrait alors se faire avec n'importe quelle autre échelle choisie. En notant H_i les coefficient (en dB) d'une pondération (X), le niveau S^{X} se calcule de la manière suivante :

$$\mathrm{S}^{\mathrm{X}} = \sum_i \mathrm{b}_i \cdot 10^{(\mathrm{E}_i+\mathrm{H}_i)/10} = \sum_i g_i(\mathrm{E}_i)\,, \tag{5.10}$$

avec

$$g_i(\mathrm{E}_i) = \mathrm{b}_i \cdot 10^{(\mathrm{E}_i+\mathrm{H}_i)/10}\,, \tag{5.11}$$

et le niveau en dB(X) se calcule naturellement par :

$$\mathrm{S}^{d\mathrm{BX}} = 10 \cdot \log(\mathrm{S}^{\mathrm{X}})\,. \tag{5.12}$$

On remarque qu'en prenant $\mathrm{H}_i = 0, \forall i$, nous retrouvons la contrainte énergétique classique avec une pondération (Z). Les coefficients de la pondération (A) qui seront utilisés plus tard dans l'étude sont visibles dans le tableau 5.1. En notant alors $\mathrm{S}^{\mathrm{X}}_{ref}$ le niveau de référence, on peut donc formuler le problème d'optimisation de la façon suivante :

$$\{\mathrm{E}_i^{opt}\} = \underset{\{\mathrm{E}_i\}}{\arg\max} \sum_i f_i(\mathrm{E}_i, \mathrm{D}_i)\,, \tag{5.13}$$

soumis à

$$\sum_i g_i(\mathrm{E}_i) = \mathrm{S}^{\mathrm{X}}_{ref}\,, \tag{5.14}$$

On obtient alors un problème d'allocation de ressource classique, aussi appelé problème du sac à dos.

5.2.2 Nouvelle procédure d'optimisation exacte

Les fonctions de contrainte g_i étant des fonctions exponentielles, elles sont dérivables et convexes sur tout $\mathbb{R}$. L'évolution des fonctions de coût $f_i(\cdot, \mathrm{D}_i)$, est visible sur la figure 5.4 dans différents cas de figure. En notant :

$$\begin{cases} \mathrm{D}_i^- = \mathrm{D}_i - 15\,\mathrm{dB} \\ \mathrm{D}_i^+ = \mathrm{D}_i + 15\,\mathrm{dB} \\ \mathrm{D}_i^u = \mathrm{U}_i + 10\,\mathrm{dB} \end{cases} \tag{5.15}$$

on peut noter que ces fonctions sont :

- constantes et minimales sur $]-\infty, \mathrm{D}_i^-]$, donc soit $\mathrm{E}_i^{opt} > \mathrm{D}_i^-$, soit la bande est vide,
- décroissantes sur $[\mathrm{D}_i^+, +\infty[$ donc $\mathrm{E}_i^{opt} \leq \mathrm{D}_i^+$,
- continues et concaves sur $[\mathrm{D}_i^-, \mathrm{D}_i^+]$,
- dérivable sur $[\mathrm{D}_i^-, \mathrm{D}_i^+]$, sauf en D_i^u si et seulement si $\mathrm{D}_i^- < \mathrm{D}_i^u < \mathrm{D}_i^+$.

Notons $\Omega_2 = \{i, \mathrm{D}_i^- < \mathrm{D}_i^u < \mathrm{D}_i^+\}$ l'ensemble des bandes qui possèdent deux intervalles de recherche où f_i est concave et dérivable : $[\mathrm{D}_i^-, \mathrm{D}_i^u]$ et $[\mathrm{D}_i^u, \mathrm{D}_i^+]$. Les bandes restantes forment l'ensemble Ω_1 et possèdent seulement un intervalle de recherche où f_i est concave et dérivable : $[\mathrm{D}_i^-, \mathrm{D}_i^+]$. Pour chaque bande, soit E_i^{opt} appartient à un de ses intervalles de recherche, soit la bande i doit être vide, on notera alors $\mathrm{E}_i^{opt} = -\infty$. Nous avons donc $3^{|\Omega_2|} \cdot 2^{|\Omega_1|}$ sous-problèmes que l'on peut résoudre par la méthode des multiplicateurs de Lagrange [29] puis sélectionner la meilleure solution.

Chaque sous-problème est résolu de la façon suivante. Notons Ω_{deact} l'ensemble des bandes vides, les bandes restantes forment l'ensemble $\Omega_{act} = (\Omega_2 \cup \Omega_1) \setminus \Omega_{deact}$ et leurs intervalles de recherche sont notés $[l_i, u_i]$ avec $l_i \in \{\mathrm{D}_i^-, \mathrm{D}_i^u\}$ et $u_i \in \{\mathrm{D}_i^u, \mathrm{D}_i^+\}$. Notons λ le multiplicateur de Lagrange pour l'équation 5.14, v_i pour $\mathrm{E}_i \geq l_i$, et w_i pour $\mathrm{E}_i \leq u_i$. Les conditions de Karush-Kuhn-Tucker (KKT) pour chaque sous-problème s'écrivent de la façon suivante :

$$\sum_{i \in \Omega_{act}} g_i(\mathrm{E}_i) = \mathrm{S}_{ref}^{\mathrm{X}}, \tag{5.16}$$

$$\forall i \in \Omega_{act}, \quad l_i \leq \mathrm{E}_i \leq u_i, \tag{5.17}$$

$$-f_i' + \lambda \cdot g_i' - v_i + w_i = 0, \tag{5.18}$$

$$v_i \cdot (l_i - \mathrm{E}_i) = 0, \tag{5.19}$$

$$w_i \cdot (\mathrm{E}_i - u_i) = 0, \tag{5.20}$$

$$v_i \geq 0, \tag{5.21}$$

$$w_i \geq 0. \tag{5.22}$$

Pour $i \in \Omega_{act}$ notons $\overline{\mathrm{E}}_i(\lambda)$ la solution de $-f_i' + \lambda \cdot g_i' = 0$ i.e. :

$$\overline{\mathrm{E}}_i(\lambda) = \begin{cases} 10 \cdot \log\left(\frac{\mathrm{I}_i}{3 \cdot \ln 10 \cdot \lambda \cdot b_i}\right) - \mathrm{H}_i & \text{si } \mathrm{D}_i^u \geq u_i, \\ \frac{160 + \mathrm{D}_i^u + \mathrm{D}_i^-}{2} - \frac{10}{\ln 10} \mathrm{W}\left(\frac{24 \cdot \lambda \cdot b_i}{\mathrm{I}_i / \ln^2(10)} 10^{(2\mathrm{H}_i + 160 + \mathrm{D}_i^u + \mathrm{D}_i^-)/20}\right) & \text{si } \mathrm{D}_i^u \leq l_i. \end{cases} \tag{5.23}$$

Notons l'utilisation de la fonction de Lambert W lorsque $\mathrm{D}_i^u \leq l_i$. Les équations suivantes satisfont toutes les conditions de KKT sauf la (5.16) :

$$\mathrm{E}_i^{opt}(\lambda) = \begin{cases} l_i & \text{si } \overline{\mathrm{E}}_i(\lambda) \leq l_i, \\ \overline{\mathrm{E}}_i(\lambda) & \text{si } l_i < \overline{\mathrm{E}}_i(\lambda) < l_i, \\ u_i & \text{si } \overline{\mathrm{E}}_i(\lambda) \geq u_i, \end{cases} \tag{5.24}$$

$$v_i(\lambda) = \begin{cases} -f_i'(l_i) + \lambda \cdot g_i'(l_i) & \text{si } \overline{\mathrm{E}}_i(\lambda) \leq l_i, \\ 0 & \text{si } \overline{\mathrm{E}}_i(\lambda) > l_i, \end{cases} \tag{5.25}$$

$$w_i(\lambda) = \begin{cases} 0 & \text{si } \overline{\mathrm{E}}_i(\lambda) < u_i, \\ f_i'(u_i) - \lambda \cdot g_i'(u_i) & \text{si } \overline{\mathrm{E}}_i(\lambda) \geq u_i. \end{cases} \tag{5.26}$$

Le λ optimal peut alors être identifié par itération en évaluant $\overline{\mathrm{E}}_i(\lambda)$ (équation 5.23) de manière à ce que la contrainte énergétique soit satisfaite i.e. $\sum_{i \in \Omega_{act}} g_i(\mathrm{E}_i^{opt}(\lambda)) = \mathrm{S}_{ref}^{\mathrm{X}}$.

Le nombre de sous-problèmes à résoudre est bien trop important pour être effectué en temps réel. Par contre, l'optimisation ne dépendant que des niveaux du spectre équivalent de perturbation D_i et du niveau de référence de la parole, il est donc possible de calculer les spectres équivalents optimaux en avance si les conditions d'écoute sont connues.

5.2.3 Vérification de l'hypothèse d'auto-masquage

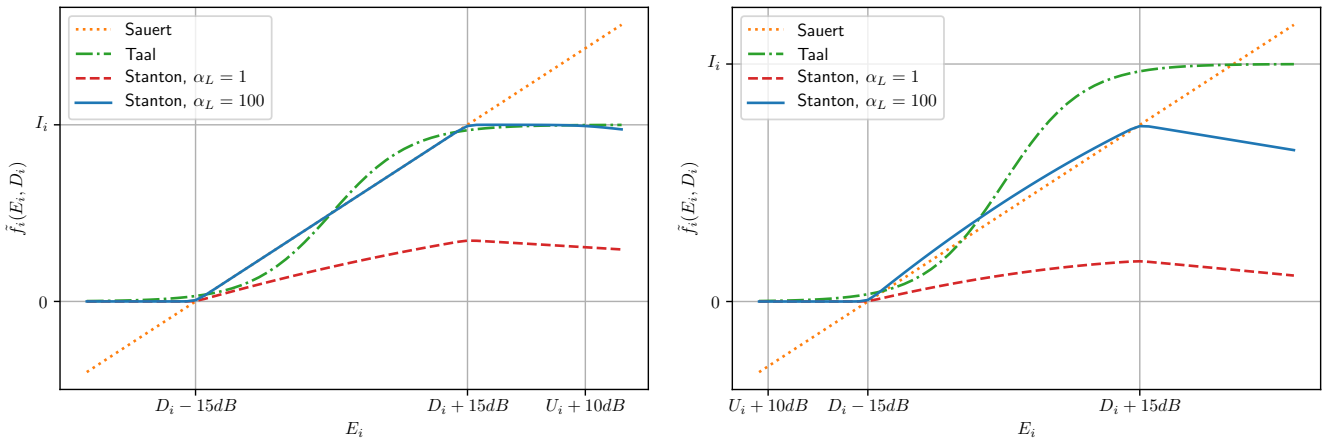
Finalement, revenons sur l'hypothèse 1 supposant qu'il n'y a pas de phénomène d'auto-masquage de la parole. En effet, toute cette procédure d'optimisation se base sur une conséquence de cette hypothèse qui est que les niveaux du spectre équivalent de perturbation D_i ne dépendent pas des E_i. Nous savons que le domaine de recherche optimal porte nécessairement sur les $E_i \leq D_i + 15\,\text{dB}$, or si l'hypothèse 1 n'est pas vérifiée, on a $E_i > N_i + 24\,\text{dB}$ et donc $D_i \geq N_i + 9\,\text{dB}$. Ainsi, la seule raison pour laquelle l'hypothèse 1 ne serait pas vérifiée, durant la procédure d'optimisation, est si un niveau du spectre équivalent de masquage du bruit dans une bande dépasse le niveau du spectre équivalent de bruit de 9 dB dans cette bande. Cela se produit seulement lorsque le bruit est présent dans une bande et bien moins dans la bande qui suit.

Un exemple de bruit qui serait concerné est le bruit synthétique représenté figure 5.1 : on voit que dans certaines bandes les niveaux de masquages sont largement supérieurs aux niveaux du bruit, il faudrait donc un niveau de parole très élevé pour maximiser la contribution de ces bandes ce qui les ferait violer l'hypothèse 1 en introduisant de l'auto-masquage et annulerait alors l'exactitude de notre procédure d'optimisation. En revanche, on remarque que ce n'est pas le cas pour les autres bruits pour lesquels les niveaux du spectre de masquage dominent légèrement les niveaux du bruit mais de bien moins que 9 dB. Il conviendra donc, pour être certain d'obtenir un résultat optimal lors de la résolution du problème d'optimisation, de vérifier la nouvelle hypothèse :

Hypothèse 2 *Les niveaux du spectre équivalent de masquage du bruit ne dépassent pas les niveau du spectre équivalent de bruit de plus de 9 dB dans chaque bande i.e.* $Z_i(\beta_i = N_i) \leq N_i + 9\,\text{dB}, \forall i$.

5.3 Présentation et adaptation des procédures d'optimisation par approximation

Toutes les procédures d'optimisation proposées dans la littérature actuelle consistent à faire une approximation des fonctions de coût f_i introduites équation 5.9 et observables dans différentes situations figure 5.4. Les tracés des approximations $\hat{f}_i$ que nous allons introduire dans cette section sont consultables figure 5.5. Notons que tous les travaux présentés se basent sur une contrainte énergétique classique en dB(Z), cependant dans leur description nous introduisons des pondérations H_i qui permettent d'étendre les méthodes à tout type de contrainte pondérée. Cela nous permettra alors de comparer les résultats sous différentes pondérations, mais il est tout à fait possible de retrouver les descriptions originales des méthodes en fixant $H_i = 0, \forall i$.



(a) $U_i + 10dB > D_i^+$ (b) $U_i + 10dB < D_i^-$

FIGURE 5.5 – Approximations des fonctions de coût $\hat{f}_i$ proposées par différents auteurs dans deux cas de figure différents. L'approximation "Stanton, $\alpha_L = 100$" étant visuellement indiscernable des fonctions de coût f_i d'origine, c'est cette approximation qui fait office de référence visuelle.

5.3.1 Approximation linéaire : SAUERT et al.

SAUERT et al. ont d'abord proposé une approximation linéaire de ces fonctions [170] en supprimant les effets de seuil des coefficients d'audibilité et en remplaçant les coefficients de distorsion par un simple facteur. Les fonctions de coût f_i ont donc pour approximation :

$$\hat{f}_i(\mathrm{E}_i, \mathrm{D}_i) = \gamma_i \cdot \hat{\mathrm{A}}_i(\mathrm{E}_i, \mathrm{D}_i)\,, \tag{5.27}$$

avec $\hat{\mathrm{A}}_i(\mathrm{E}_i, \mathrm{D}_i)$ les coefficients d'audibilité sans limitations à savoir :

$$\hat{\mathrm{A}}_i(\mathrm{E}_i, \mathrm{D}_i) = \frac{\mathrm{E}_i - \mathrm{D}_i^-}{30\,\mathrm{dB}}\,. \tag{5.28}$$

et γ_i les combinaisons entre les coefficients de la FIB et les coefficients de distorsions qui sont maintenant fixés :

$$\gamma_i = \mathrm{I}_i \cdot \mathrm{L}_i(\mathrm{D}_i^+) \tag{5.29}$$

Les fonctions étant maintenant linéaires, et donc concaves, sur tout l'espace, le problème d'optimisation sous contrainte énergétique (X) se résout par la méthode des multiplicateurs de Lagrange. Tout le procédé introduit section 5.2 pour la résolution des sous-problèmes est valable pour cette approximation en supprimant les bornes de recherche et en remplaçant les fonctions de coût f_i par leurs approximations, on obtient alors :

$$\mathrm{E}_i^{opt} = 10 \cdot \log\left(\frac{\gamma_i}{3 \cdot \ln 10 \cdot \lambda \cdot b_i}\right) - \mathrm{H}_i\,, \tag{5.30}$$

avec :

$$\sum_i g_i(\mathrm{E}_i(\lambda)) = \mathrm{S}^{\mathrm{X}}_{ref} \iff \sum_i \frac{\gamma_i}{3 \cdot \ln 10 \cdot \lambda} = \mathrm{S}^{\mathrm{X}}_{ref} \iff \lambda = \frac{1}{3 \cdot \mathrm{S}^{\mathrm{X}}_{ref} \cdot \ln 10} \sum_i \gamma_i\,, \tag{5.31}$$

on trouve alors :

$$\mathrm{E}_i^{opt} = 10 \cdot \log\left(\frac{\gamma_i \cdot \mathrm{S}^{\mathrm{X}}_{ref}}{b_i \cdot \sum_i \gamma_i}\right) - \mathrm{H}_i \tag{5.32}$$

Cette optimisation n'étant pas bornée, il peut arriver que des niveaux E_i^{opt} soient supérieurs à D_i^+ ce qui n'apporte rien de plus à l'intelligibilité du point de vue du SII. Les auteurs proposent donc de saturer à D_i^+ les bandes bandes de l'ensemble $\Omega_{sat} = \{i, \mathrm{E}_i^{opt} > \mathrm{D}_i^+\}$ et de relancer une procédure d'optimisation sur l'ensemble $\Omega_{res} = \{i, \mathrm{E}_i^{opt} \leq \mathrm{D}_i^+\}$ des autres bandes avec l'énergie restante à savoir :

$$\mathrm{E}_i^{opt} = \begin{cases} \mathrm{D}_i^+ & \text{si } i \in \Omega_{sat}\,, \\ 10 \cdot \log\left(\frac{\gamma_i \cdot \mathrm{S}^{\mathrm{X}}_{res}}{b_i \cdot \sum_{i \in \Omega_{res}} \gamma_i}\right) - \mathrm{H}_i & \text{si } i \in \Omega_{res}\,. \end{cases} \tag{5.33}$$

avec :

$$\mathrm{S}^{\mathrm{X}}_{res} = \mathrm{S}^{\mathrm{X}}_{ref} - \sum_{i \in \Omega_{sat}} g_i(\mathrm{D}_i^+) \tag{5.34}$$

Cette opération est alors répétée autant de fois qu'il existe des $\mathrm{E}_i^{opt} > \mathrm{D}_i^+$ à l'issue de celle-ci.

En revanche, aucune proposition n'est faite sur le fait que, si les niveaux E_i^{opt} sont inférieurs à D_i^-, ils n'apportent rien non plus à l'intelligibilité du point de vue du SII et qu'il serait alors intéressant de ré-allouer cette énergie gaspillée ailleurs. Au contraire, les fonctions de coût f_i ont été approximées par une fonction linéaire même en dessous de D_i^-, ce qui force alors l'optimisation à investir de l'énergie dans ces bandes pour ne pas trop pénaliser le coût total.

5.3.2 Approximation non-linéaire concave : TAAL et al.

TAAL et al. ont ensuite proposé une approximation non-linéaire des fonctions de coût f_i [193]. Dans leur méthode, les coefficients de distorsion ne sont pas pris en compte, ce sont donc les coefficients d'audibilité qui sont approchés par une approximation non-linéaire. Les fonctions de coût $f_i(\mathrm{E}_i, \mathrm{D}_i)$ ont alors pour approximation :

$$\hat{f}_i(\mathrm{E}_i, \mathrm{D}_i) = \mathrm{I}_i \cdot \hat{\mathrm{A}}_i(\mathrm{E}_i, \mathrm{D}_i)\,, \tag{5.35}$$

avec $\hat{\mathrm{A}}_i(\mathrm{E}_i, \mathrm{D}_i)$ les coefficients d'audibilité calculés par approximation à savoir :

$$\hat{\mathrm{A}}_i(\mathrm{E}_i, \mathrm{D}_i) = \frac{10^{(\mathrm{E}_i - \mathrm{D}_i^-)/10}}{1 + 10^{(\mathrm{E}_i - \mathrm{D}_i^-)/10}}\,. \tag{5.36}$$

Comme nous pouvons le voir figure 5.5, les approximations des fonctions de coût $\hat{f}_i$ proposées ne sont pas concaves en fonction des E_i. Cependant, en changeant de variable pour les niveaux bruts $e_i = 10^{\mathrm{E}_i/10}$, elles s'expriment alors :

$$\hat{f}_i(e_i, d_i) = \gamma_i \cdot \frac{e_i / d_i}{1 + e_i / d_i}\,. \tag{5.37}$$

avec $d_i = 10^{\mathrm{D}_i/10}$ et ces fonctions sont concaves en fonction de e_i sur $\mathbb{R}_+$. Le problème d'optimisation sous contrainte en fonction des e_i se formule alors de la façon suivante :

$$\{e_i^{opt}\} = \underset{\{e_i\}}{\arg\max} \sum_i \hat{f}_i(e_i, d_i)\,, \tag{5.38}$$

soumis à :

$$\sum_i b_i \cdot h_i \cdot e_i = \mathrm{S}^{\mathrm{X}}_{ref}\,, \tag{5.39}$$

$$\forall i, e_i \geq 0\,, \tag{5.40}$$

avec $h_i = 10^{\mathrm{H}_i/10}$. Les auteurs résolvent alors ce problème directement par la méthode des multiplicateurs de Lagrange. Tout le procédé introduit section 5.2 pour la résolution des sous-problèmes est valable pour cette approximation en supprimant la borne supérieure, en fixant la borne inférieur $l_i = 0$ et en changeant l'expression des f_i, on obtient alors :

$$e_i^{opt} = \max(0, \overline{e}_i(\lambda))\,, \tag{5.41}$$

avec :

$$\overline{e}_i(\lambda) = (\sqrt{\frac{\mathrm{I}_i \cdot d_i}{\lambda \cdot b_i \cdot h_i}} - d_i) \tag{5.42}$$

$$\sum_i b_i \cdot h_i \cdot e_i(\lambda) = \mathrm{S}^{\mathrm{X}}_{ref} \tag{5.43}$$

$$\sum_i b_i \cdot h_i \cdot e_i(\lambda) = \mathrm{S}^{\mathrm{X}}_{ref} \iff \frac{1}{\sqrt{\lambda}} = \frac{\mathrm{S}^{\mathrm{X}}_{ref} + \sum\limits_{i, \overline{e}_i(\lambda))>0} b_i \cdot h_i \cdot d_i}{\sum\limits_{i, \overline{e}_i(\lambda)>0} \sqrt{\mathrm{I}_i \cdot b_i \cdot h_i \cdot d_i}}\,, \tag{5.44}$$

Le λ optimal peut alors être identifié par itération en évaluant $\overline{e}_i(\lambda)$ (équation 5.42) de manière à ce que la contrainte énergétique soit satisfaite i.e. $\sum\limits_i b_i \cdot h_i \cdot e_i(\lambda) = \mathrm{S}^{\mathrm{X}}_{ref}$.

Notons que les auteurs de cette méthode ne prennent pas en compte les effets de masquages ($\mathrm{Z}_i = \mathrm{N}_i$) et ils supposent que les niveaux du bruit sont supérieurs aux seuils d'audibilité ($\mathrm{N}_i > \mathrm{T}_i$) ce qui fait qu'ils travaillent directement avec les niveaux du spectre équivalent de bruit i.e. $\mathrm{D}_i = \mathrm{N}_i$. La seule justification qui est fournie est le fait que SAUERT et al. ne l'ont pas fait, ce qui n'est pas le cas. Ainsi, vu que cela ne rajoute pas de complexité supplémentaire et que cela ne va pas à l'encontre de la philosophie de la méthode, nous nous permettrons donc d'utiliser les véritables niveaux de perturbations avec la prise en compte du masquage.

5.3.3 Approximation non-linéaire non-concave : STANTON et al.

Finalement, STANTON et al. proposent aussi une approximation non-linéaire des fonctions de coût f_i [183]. Ils remarquent qu'une difficulté majeure pour la résolution du problème de maximisation du SII est la non-dérivabilité des fonctions des coefficients à cause de l'utilisation de minima et maxima. Ainsi, ils proposent d'approximer ces fonctions en utilisant des extrema généralisé, et plus particulièrement ils proposent l'utilisation du maximum régularisé *LogSumExp* à savoir :

$$\max(a,b) \approx \ln(e^a + e^b)\,. \tag{5.45}$$

Le minimum généralisé correspondant s'obtient alors directement avec :

$$\min(a,b) \approx -\ln(e^{-a} + e^{-b})\,. \tag{5.46}$$

Ainsi les approximations des coefficients sont les suivantes :

$$\hat{\mathrm{L}}_i(\mathrm{E}_i) = -\ln(e^{-1} + e^{-(1-(\mathrm{E}_i-(\mathrm{U}_i+10))/160)})\,, \tag{5.47}$$

$$\hat{\mathrm{A}}_i(\mathrm{E}_i,\mathrm{D}_i) = -\ln(e^{-\alpha_\mathrm{A}} + e^{-\ln(1+e^{\alpha_\mathrm{A}(\mathrm{E}_i-\mathrm{D}_i^-)/30})})/\alpha_\mathrm{A}\,, \tag{5.48}$$

dont le produit donne l'approximation des fonctions de coût f_i à savoir :

$$\hat{f}_i(\mathrm{E}_i,\mathrm{D}_i) = \mathrm{I}_i \cdot \hat{\mathrm{A}}_i(\mathrm{E}_i,\mathrm{D}_i) \cdot \hat{\mathrm{L}}_i(\mathrm{E}_i)\,. \tag{5.49}$$

On remarque l'utilisation de la forme α-*quasimax*, pour les coefficients d'audibilité A_i, permettant d'améliorer la précision mais une valeur trop grande du paramètre α_A engendre des dérivées importantes aux niveaux des angles qui pourront porter préjudice à l'optimisation qui suit. Les auteurs préconisent l'utilisation de $\alpha_\mathrm{A} = 100$. En revanche, aucun traitement particulier n'est précisé pour les coefficients de distorsion. Et pourtant, si on calcule l'approximation des coefficients de distorsion en D_i^+ on obtient :

$$\hat{\mathrm{L}}_i(\mathrm{D}_i^+) = -\ln(e^{-1} + e^{-(1-(\mathrm{D}_i^+-(\mathrm{U}_i+10)))/160)})\,, \tag{5.50}$$

prenons alors les cas particuliers mais tout à fait probable où $\mathrm{D}_i^+ \approx (\mathrm{U}_i + 10)$, on obtient alors :

$$\hat{\mathrm{L}}_i(\epsilon) \approx -\ln(e^{-1} + e^{-1}) = 1 - \ln(2) \approx 0{,}3\,, \tag{5.51}$$

alors que :

$$\mathrm{L}_i(\mathrm{U}_i + 10) = 1\,. \tag{5.52}$$

L'approximation proposée n'est donc clairement pas adaptée et nous proposons alors d'utiliser aussi la forme α-*quasimax* pour les coefficients de distorsion, afin d'améliorer la précision, à savoir :

$$\hat{\mathrm{L}}_i(\mathrm{E}_i) = -\ln(e^{-\alpha_\mathrm{L}} + e^{-\alpha_\mathrm{L}(1-(\mathrm{E}_i-(\mathrm{U}_i+10))/160)})/\alpha_\mathrm{L}\,. \tag{5.53}$$

Avec $\alpha_\mathrm{L} = 100$, les approximations $\hat{f}_i$ sont visuellement indiscernables des fonctions de coût originales f_i, bien qu'elles soient mathématiquement différentes. C'est pourquoi nous n'avons pas tracé les f_i sur la figure 5.5 pour éviter la superposition des courbes et donc pour plus de clarté. Nous pouvons d'ailleurs observer sur cette figure que pour $\alpha_\mathrm{L} = 1$, il y avait effectivement un problème d'ajustement. Nous supposerons donc que c'était un oubli des auteurs et nous travaillerons alors avec ce nouveau paramétrage des $\hat{\mathrm{L}}_i$, nous vérifierons tout de même toujours que les résultats avec $\alpha_\mathrm{L} = 1$ sont bien moins performants et ce sera toujours le cas dans notre étude. Bien que les approximation des fonctions de coût $\hat{f}_i$ soient maintenant très précises, elles ne sont pas concaves. Les auteurs résolvent alors le problème de maximisation directement par l'utilisation d'un algorithme d'optimisation différentiable. Les dérivées se calculant de la façon suivante :

$$\frac{\partial \hat{f}_i}{\partial \mathrm{E}_i} = \mathrm{I}_i \cdot \left(\hat{\mathrm{A}}_i \cdot \frac{\partial \hat{\mathrm{L}}_i}{\partial \mathrm{E}_i} + \hat{\mathrm{L}}_i \cdot \frac{\partial \hat{\mathrm{A}}_i}{\partial \mathrm{E}_i}\right). \tag{5.54}$$

avec :

$$\frac{\partial \hat{\mathrm{L}}_i}{\partial \mathrm{E}_i} = -\frac{1}{160} \cdot \frac{e^{\alpha_\mathrm{L}(\mathrm{E}_i-(\mathrm{U}_i+10))/160}}{1+e^{\alpha_\mathrm{L}(\mathrm{E}_i-(\mathrm{U}_i+10))/160}} \tag{5.55}$$

$$\frac{\partial \hat{\mathrm{A}}_i}{\partial \mathrm{E}_i} = \frac{1}{30} \cdot \frac{e^{\alpha_\mathrm{A}(\mathrm{E}_i-\mathrm{D}_i^-)/30} \cdot (1+e^{\alpha_\mathrm{A}(\mathrm{E}_i-\mathrm{D}_i^-)/30})^{-2}}{e^{-\alpha_\mathrm{A}}+(1+e^{\alpha_\mathrm{A}(\mathrm{E}_i-\mathrm{D}_i^-)/30})^{-1}}. \tag{5.56}$$

Ayant conscience que le résultat puisse se trouver dans un maximum local, dû à la non-concavité des fonctions, ils proposent alors comme point de départ de l'optimisation le résultat obtenu avec la méthode de TAAL et al. introduit juste précédemment. Cela leur permet alors d'avoir une initialisation théoriquement proche du spectre optimal, et l'algorithme d'optimisation différentiable utilisé sur les approximations permet alors d'affiner la recherche.

5.4 Résultats objectifs

Pour un bruit et un auditeur donnés, il est possible de calculer les spectres équivalents optimaux pour plusieurs niveaux du signal de référence $\mathrm{S}_{ref}^{d\mathrm{BX}}$ sous une contrainte énergétique en dB(X). Afin d'obtenir une visualisation complète du processus d'optimisation, nous faisons varier $\mathrm{S}_{ref}^{d\mathrm{BX}}$ dans un intervalle allant du niveau minimal pour obtenir un SII non nul $\mathrm{S}_{min}^{d\mathrm{BX}}$, au niveau nécessaire pour obtenir un SII maximum $\mathrm{S}_{max}^{d\mathrm{BX}}$, par pas de 1 dB(X). Les formules de ces niveaux sont respectivement exprimées par les équations suivantes :

$$\mathrm{S}_{min}^{d\mathrm{BX}} = \min_i(\mathrm{D}_i - 15\,\mathrm{dB} + \mathrm{H}_i + 10\cdot\log(\mathrm{b}_i)), \tag{5.57}$$

$$\mathrm{S}_{max}^{d\mathrm{BX}} = 10\cdot\log(\sum_i \mathrm{b}_i \cdot 10^{(\mathrm{D}_i+15\,\mathrm{dB}+\mathrm{H}_i)/10}). \tag{5.58}$$

À partir des niveaux optimaux E_i^{opt} trouvés pour un $\mathrm{S}_{ref}^{d\mathrm{BX}}$ donné, on peut calculer le niveau en dB(Z) du spectre optimal directement par :

$$\mathrm{S}_{ref}^{d\mathrm{BZ}} = 10\cdot\log\sum_i \mathrm{b}_i \cdot 10^{\mathrm{E}_i^{opt}/10}, \tag{5.59}$$

ainsi $\mathrm{S}_{ref}^{d\mathrm{BX}} = \mathrm{S}_{ref}^{d\mathrm{BZ}}$ pour une contrainte en dB(Z) mais $\mathrm{S}_{ref}^{d\mathrm{BX}} \neq \mathrm{S}_{ref}^{d\mathrm{BZ}}$ pour une contrainte en dB(A) par exemple. Cette opération permet de ramener les résultats de plusieurs contraintes sur une même échelle afin de comparer les résultats. Le niveau $\mathrm{B}^{d\mathrm{BZ}}$ du bruit considéré se calcule aussi directement à partir des niveaux de son spectre équivalents à savoir :

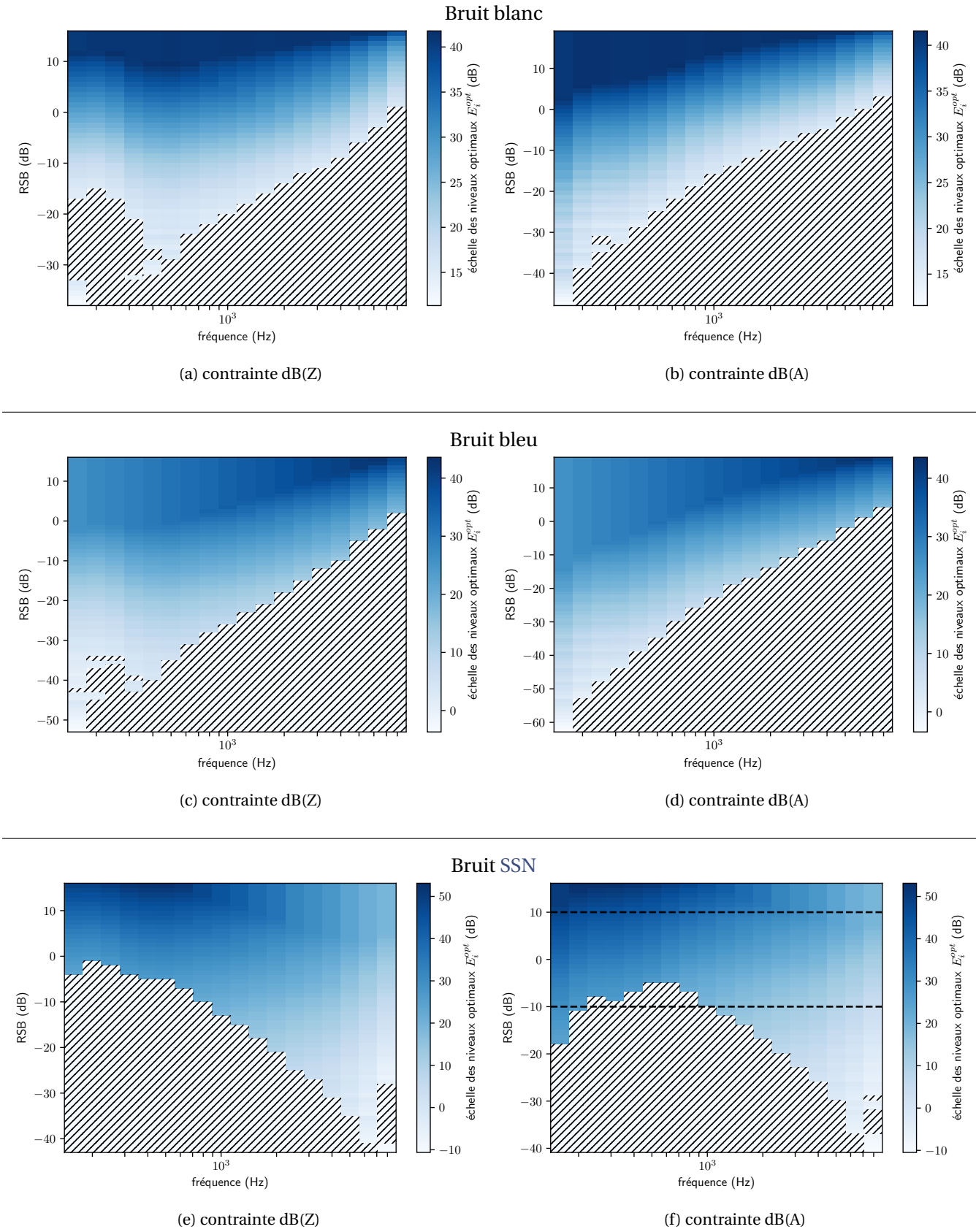
$$\mathrm{B}^{d\mathrm{BZ}} = 10\cdot\log\sum_i \mathrm{b}_i \cdot 10^{\mathrm{N}_i/10}, \tag{5.60}$$

ainsi le RSB se calcule de la façon suivante :

$$\mathrm{RSB} = \mathrm{S}_{ref}^{d\mathrm{BZ}} - \mathrm{B}^{d\mathrm{BZ}} \tag{5.61}$$

5.4.1 Analyse des spectres optimaux

Les niveaux optimaux du spectre équivalent de parole en fonction du RSB sont visibles figure 5.6 pour les trois bruits classiques (un bruit correspond à une ligne) et pour deux contraintes énergétiques différentes en dB(Z) (colonne de gauche) et en dB(A) (colonne de droite). Ces résultats ont été obtenus en supposant que les seuils d'audibilité de l'auditeur sont inférieurs au spectre de masquage i.e. $\mathrm{T}_i \leq \mathrm{Z}_i$ et donc $\mathrm{D}_i = \mathrm{Z}_i$ (voir équation 5.1). Deux exemples de niveaux E_i^{opt} sont consultables figure 5.7, ils ont été obtenus dans le bruit SSN, sous contrainte en dB(A), à des RSB de -10 dB et +10 dB : ces spectres équivalents optimaux correspondent aux deux coupes tracées sur la figure 5.6f.



(a) contrainte dB(Z)

(b) contrainte dB(A)

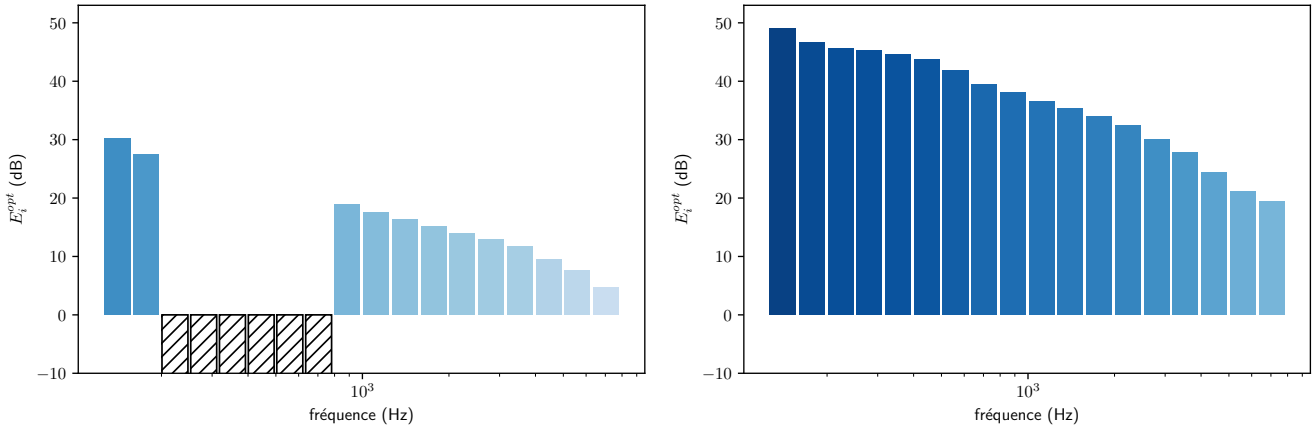
(c) contrainte dB(Z)

(d) contrainte dB(A)

(e) contrainte dB(Z)

(f) contrainte dB(A)

FIGURE 5.6 – Niveaux E_i^{opt} des spectres équivalents optimaux à différents RSB pour chaque bruit et pour chaque contrainte : dB(Z) à gauche et dB(A) à droite. Deux exemples de niveaux E_i^{opt}, obtenus dans le bruit SSN sous contrainte en dB(A), au niveau des deux coupes tracées sur la sous-figure (f), sont consultables figure 5.7



(a) RSB = -10 dB

(b) RSB = 10 dB

FIGURE 5.7 – Exemples des niveaux E_i^{opt} des spectres équivalents optimaux dans le bruit SSN, sous contrainte en dB(A), pour deux RSB différents. Visuellement, ils correspondent aux deux coupes tracées figure 5.6f.

Observations globales

Pour bien comprendre le comportement du processus d'optimisation pour les différents RSB il est intéressant d'observer les niveaux optimaux des spectres équivalents de parole des faibles vers les forts RSB (de bas en haut sur chaque sous-figure de la figure 5.6). On remarque globalement que pour de faibles RSB, de nombreuses bandes sont désactivées (fond rayé sur la figure) et toute l'énergie est allouée dans certaines bandes qui sont favorisées par l'optimisation. Pour des RSB plus élevés, et à mesure que la réserve d'énergie disponible augmente, de nouvelles bandes sont progressivement activées. La favorisation de certaines bandes plutôt que d'autres est due majoritairement au niveau de masquage présent dans ces bandes : moins il y a de masquage dans une bande, moins il sera coûteux d'y investir de l'énergie afin d'y faire émerger la parole au dessus du masquage. De plus, les différents facteurs de pondération introduits par la FIB, les H_i et les b_i vont aussi conditionner l'allocation énergétique en donnant plus ou moins d'importance aux bandes :

- la FIB donne plus d'importance aux bandes sensibles de l'oreille, autour de 2 kHz, avec une contribution décroissante lorsqu'on s'en éloigne de part et d'autre de ce pic,
- les largeurs de bande b_i qui croissent avec la fréquence centrale de chaque bande, font que le coût en énergie est de plus en plus important pour les bandes à la fréquence centrale de plus en plus élevée, elles favorisent donc les bandes étroites i.e. les bandes basse fréquence,
- les coefficients de pondération H_i, pour une pondération (Z) ils n'ont aucune influence puisqu'ils sont nuls, pour une pondération (A) il vont nuancer l'effet de la FIB en augmentant le coût en énergie dans les bandes où l'oreille est sensible.

Pour étudier l'influence des différents facteurs de pondération, le bruit blanc est certainement le plus adapté car ses niveaux équivalents de bruit (resp. de masquage) sont constants (resp. quasi-constants) dans chaque bande, comme nous pouvons le voir figure 5.1a. Ainsi, pour ce bruit, sa répartition spectrale n'a que peu d'influence sur l'optimisation.

Bruit blanc, contrainte dB(A) : influence des largeurs de bande

Avec la contrainte en dB(A), l'influence de la FIB est minimisée et il reste alors seulement le facteur de largeur de bande b_i qui influence majoritairement la maximisation du SII. Les résultats du bruit blanc avec contrainte en dB(A) visible figure 5.6b sont alors inversement corrélés avec la largeur de bande. En effet, on observe bien que ce sont d'abord les bandes étroites qui sont remplies puis, lorsque que l'énergie disponible augmente, ce sont les bandes plus larges qui sont investies.

Bruit blanc, contrainte dB(Z) : influence de la FIB

Avec la contrainte en dB(Z), plus rien ne compense la FIB qui va alors favoriser les bandes sensibles de l'oreille. Le facteur de largeur de bande b_i toujours présent favorise quant à lui les bandes étroites basse fréquence. Pour les résultats du bruit blanc avec contrainte en dB(Z) visible figure 5.6a, on observe alors un compromis qui s'installe avec une favorisation des bandes autour de 400 Hz où la FIB commence à croître et où la largeur des bandes est encore assez étroite.

Bruit bleu : influence d'un bruit haute-fréquence

Pour les bruits différents du bruit blanc, dont la densité spectrale n'est pas constante dans les bandes, le spectre va grandement influencer l'optimisation. Pour le bruit bleu par exemple, sa densité spectrale étant croissante avec la fréquence centrale des bandes, le coût en énergie nécessaire pour que les bandes haute fréquence contribuent à l'intelligibilité est renforcé. Cela pousse alors encore plus à la favorisation des bandes basse fréquence, déjà introduite par le facteur de largeur de bande b_i, que ce soit pour la contrainte en dB(A), figure 5.6d, ou pour la contrainte en dB(Z), figure 5.6c, pour laquelle le compromis commence maintenant plutôt vers les bandes autour de 200 Hz.

Bruit SSN : influence d'un bruit basse-fréquence

Enfin, le bruit SSN, possédant un spectre très basses fréquences accompagné d'une pente spectrale importante de −9 dB/octave, donne des résultats très différents des deux bruits précédents. En effet, cette répartition spectrale très localisée fait qu'il est bien plus intéressant d'investir l'énergie disponible dans les bandes où il y a très peu de bruit. Cela permet de faire émerger le signal de parole même si les facteurs de pondérations dans ces bandes ne sont pas très élevés. La différence entre les résultats pour la contrainte en dB(Z), figure 5.6e, et ceux pour la contrainte en dB(A), figure 5.6f, sont plus subtiles à analyser. On remarque bien une distribution différente, notamment vers les RSB autour de -10 dB pour lesquels les bandes basses fréquences sont favorisées plus tôt pour la contrainte en dB(A). Pour cette contrainte, on observe aussi un investissement plus tardif dans les médiums où la sensibilité auditive est importante : ces bandes sont activées assez tôt mais il faut attendre un RSB autour de 10 dB pour qu'il y ai beaucoup d'énergie (couleur foncée sur l'échelle des niveaux) contrairement à la contrainte en dB(Z) où cet investissement se fait plutôt autour de 0 dB.

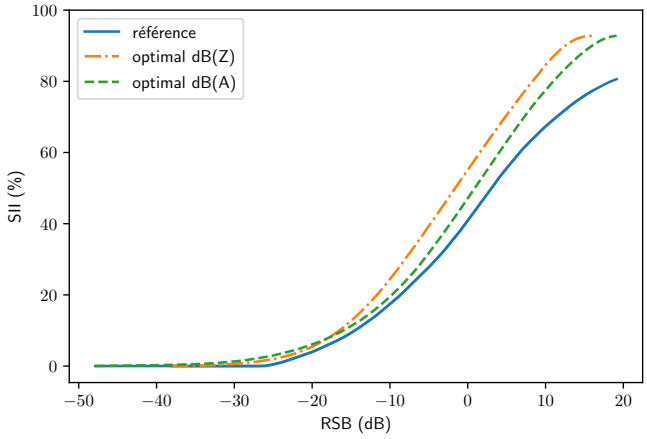
5.4.2 Améliorations optimales du SII

Afin d'observer l'amélioration du SII obtenu avec les spectres optimaux par rapport à un spectre de parole classique, nous générons un spectre équivalent de parole de référence normalisé au niveau S_{ref}^{dBZ}. Les niveaux U'_i de ce spectre se calculent de la façon suivante :

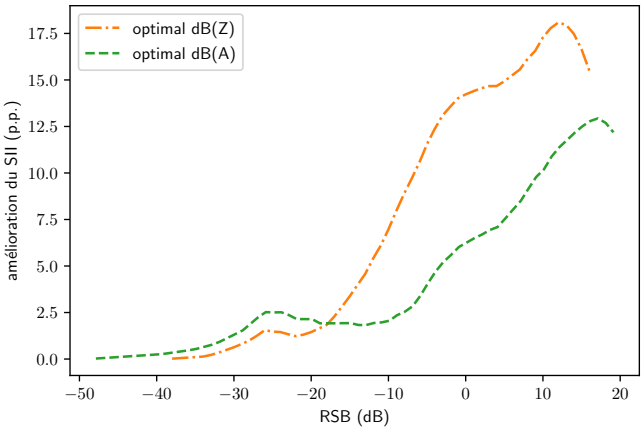
$$U'_i = U_i - 10 \cdot \log(\sum_i b_i \cdot 10^{U_i/10}) + S_{ref}^{dBZ}, \tag{5.62}$$

avec U_i les niveaux du spectre équivalent de parole de référence fourni dans la norme ANSI/ASA S3.5 [7] et visible sur la figure 5.3a. Les SII des spectres optimaux sous contrainte dB(Z) et dB(A), ainsi que ceux du spectre de parole de référence normalisé, sont visibles sur les sous-figures de gauche de la figure 5.8 pour chacun des trois bruits et pour l'ensemble des RSB. Les améliorations des SII en p.p., correspondant aux différences entre les SII des spectres optimaux et ceux de la parole de référence, sont tracées à côté de chaque sous-figure.

Bruit blanc

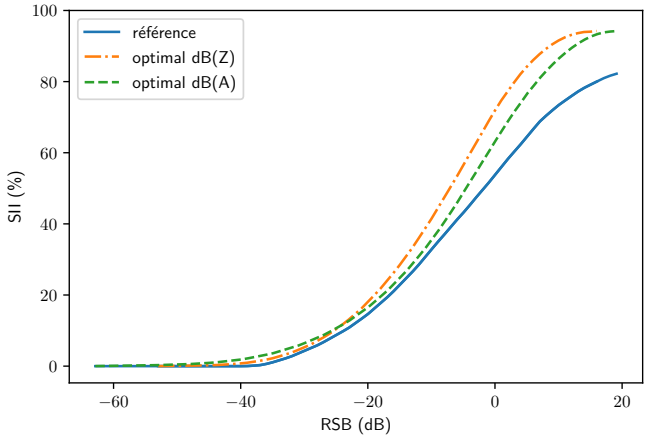


(a) comparaison des SII

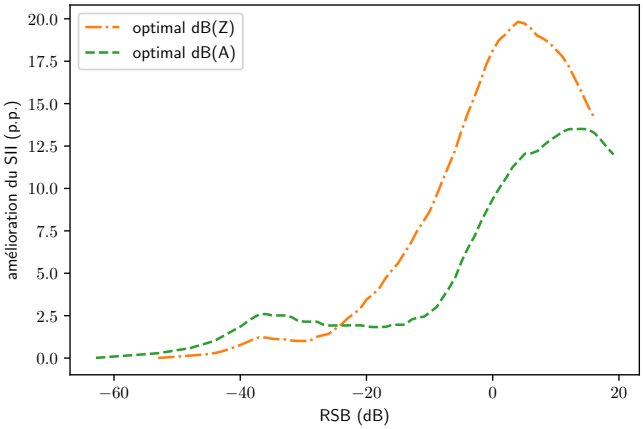


(b) amélioration du SII optimal/référence

Bruit bleu

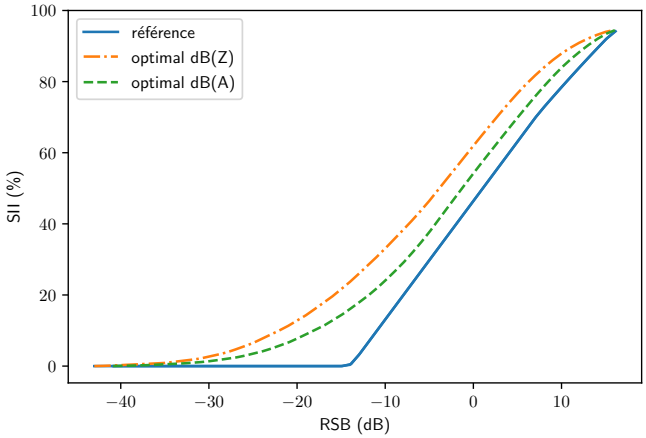


(c) comparaison des SII

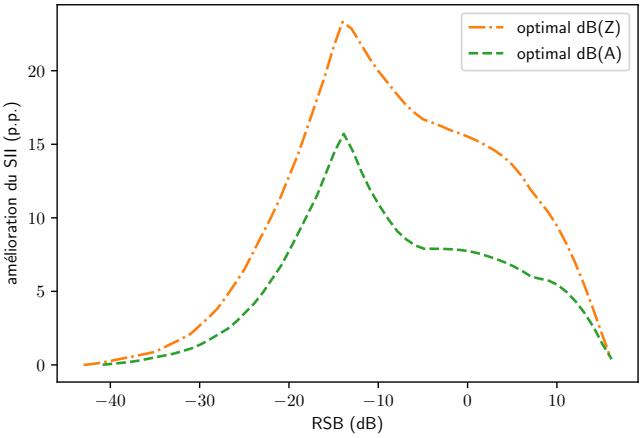


(d) amélioration du SII optimal/référence

Bruit SSN



(e) comparaison des SII



(f) amélioration du SII optimal/référence

FIGURE 5.8 – Comparaison, et amélioration, du SII calculé à partir d'un spectre de parole de référence normalisé, d'un spectre optimal de parole de même niveau en dB(Z) et d'un spectre optimal de parole de même niveau en dB(A), pour différents RSB et pour chaque bruit classique.

Pour la contrainte en dB(Z), on observe une nette amélioration du SII dans des zones de RSB qui diffèrent en fonction des bruits. En effet, on remarque que pour le bruit blanc (resp. bleu) l'amélioration du SII n'est pas très marquée pour les faibles RSB, cependant à partir de -20 dB (resp. -25 dB) l'amélioration croît rapidement et atteint un pic vers les RSB maximaux à 10 dB (resp. 5 dB). Cela s'explique par la densité spectrale de bruit très présente dans les bandes favorisées par la FIB, il faut donc attendre une réserve d'énergie suffisamment importante pour pouvoir investir dans ces bandes qui contribuent grandement à l'intelligibilité. Au contraire, pour le bruit SSN, le pic d'amélioration se situe à des niveaux bien plus faible autour de -15 dB. La densité spectrale de bruit étant bien moins présente dans les bandes d'importance, les niveaux où l'énergie disponible est suffisante pour investir ces bandes sont alors bien plus faibles. De plus, les spectres optimaux commencent à améliorer nettement le SII vers -30 dB alors que le spectre de parole de référence, étant relativement basse fréquence, ne présente un SII non-nul qu'à partir de -14 dB, cela justifie alors la localisation du pic d'amélioration.

L'amélioration du SII sous contrainte en dB(A) en fonction du RSB se comporte de façon très similaire, cependant elle est globalement bien moins importante qu'à dB(Z) constant. On se rend alors bien compte de l'effet néfaste que la contrainte perceptive proposée introduit concernant la maximisation du SII. Cependant, l'amélioration est toujours notable, la présence de pic d'amélioration est toujours présente, on peut donc toujours s'attendre à de bonnes performances subjectives de la méthode même sous la nouvelle contrainte perceptive.

5.4.3 Améliorations du SII par approximation et proposition d'extension

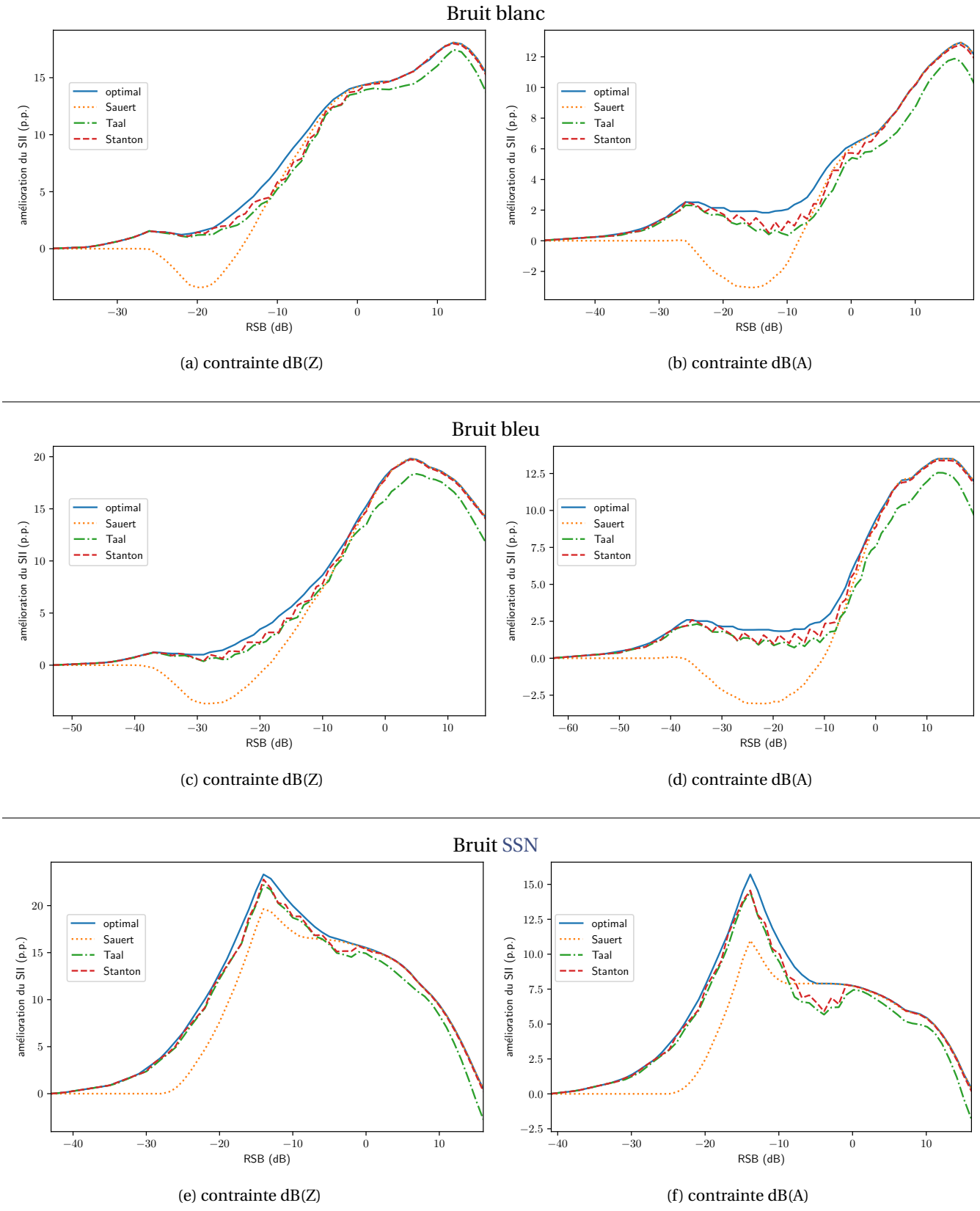
En ce qui concerne les spectres obtenus par les différentes méthodes d'approximation des fonctions de coût introduites section 5.3, les améliorations du SII en p.p. obtenus avec ces spectres par rapport au SII obtenu avec le spectre de référence normalisé sont visibles figure 5.9, pour chacun des trois bruits classiques et pour chaque contrainte, en dB(Z) et en dB(A).

Observations globales

Toutes les approches par approximation améliorent le SII quasi-systématiquement mais on remarque des performances très diversifiées sur toute la gamme de RSB. Nous rappelons que ces approches se basent sur des fonctions de coût différentes de celles utilisées par le calcul du SII. Ainsi, un spectre qui est optimal pour une approximation ne l'est pas forcément du point de vue du SII réel. On remarque que la contrainte énergétique choisie ne semble pas spécialement influencer les résultats obtenus, si ce n'est les remarques déjà faites pour les résultats optimaux précédemment.

Approximation linéaire : SAUERT et al.

L'approche de SAUERT et al. est quasi-optimale pour des RSB suffisamment élevés, cependant elle ne l'est plus du tout pour des RSB plus faibles, on observe même un SII moins important que le spectre de référence normalisé pour le bruit blanc vers -20 dB RSB et pour le bruit bleu vers -30 dB RSB. Cela s'explique par les approximations des fonctions de coût $\hat{f}_i$ proposées qui continuent d'être linéaires même en dessous de D_i^-. Là où la fonction de coût réelle f_i est nulle, et où les autres approximations tendent de façon asymptotique vers zéro, les $\hat{f}_i$ de SAUERT et VARY tendent linéairement vers $-\infty$, cela force alors l'optimisation à investir de l'énergie inutile dans ces bandes sans quoi le coût total serait extrêmement pénalisé. Au contraire, lorsque la réserve en énergie devient suffisante pour qu'il puisse y avoir naturellement suffisamment d'énergie dans chaque bande, l'approximation linéaire semble très performante pour générer des spectres quasi-optimaux dans les trois bruits classiques étudiés.



(a) contrainte dB(Z)

(b) contrainte dB(A)

(c) contrainte dB(Z)

(d) contrainte dB(A)

(e) contrainte dB(Z)

(f) contrainte dB(A)

FIGURE 5.9 – Amélioration du SII, entre un spectre de parole de référence normalisé et le spectre de parole issu de chaque approche, en fonction du RSB et pour chaque bruit classique.

Approximation non-linéaire concave : TAAL et al.

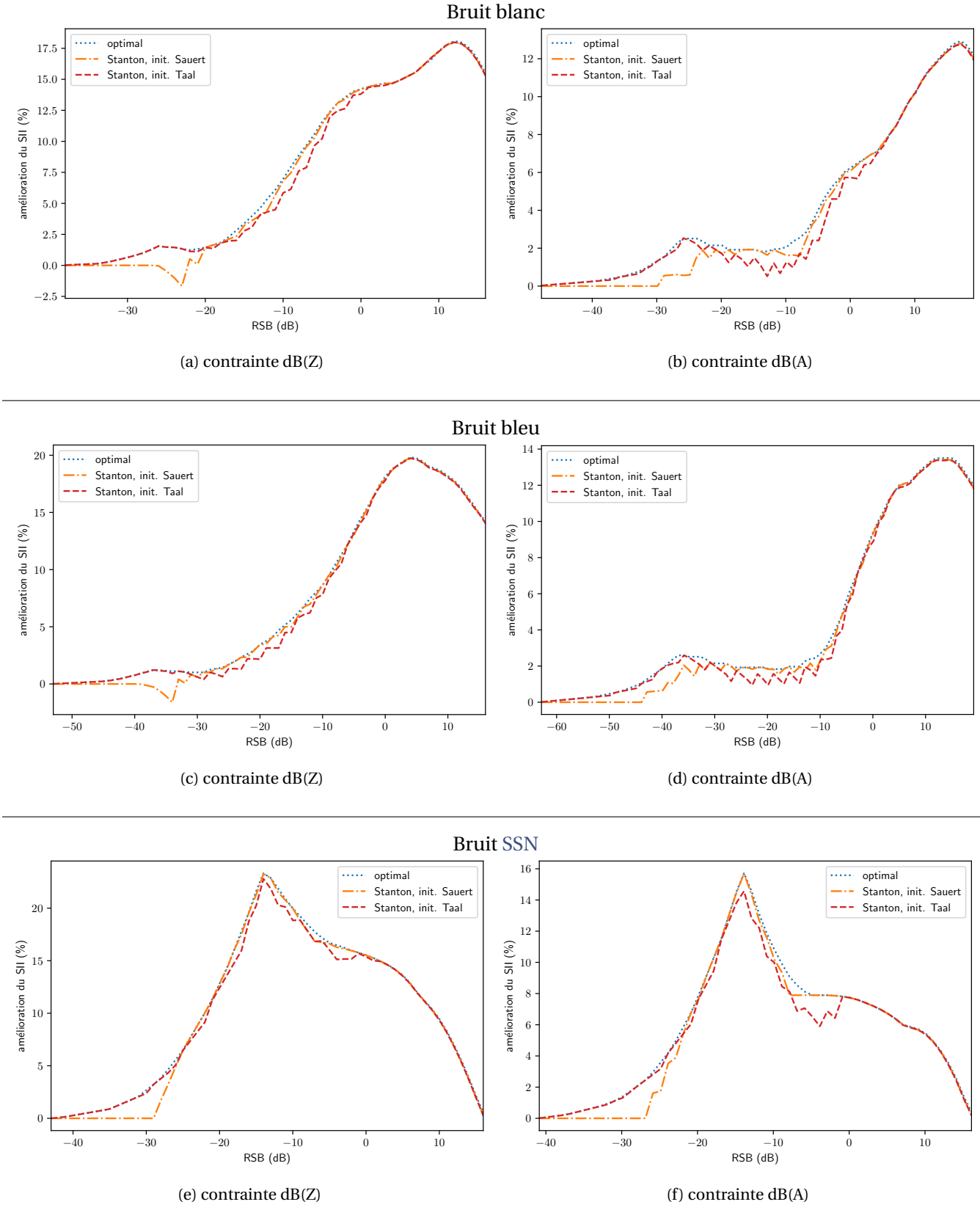
L'approche de TAAL et al. présente un comportement inverse. En effet, les SII obtenus pour des RSB très faibles sont quasi-optimaux alors que l'approche s'éloigne des résultats optimaux pour des RSB plus importants. Cela s'explique par la structure des approximations des fonctions de coût $\hat{f}_i$ proposées qui, cette fois, tendent de façon asymptotique vers zéro, autorisant alors l'optimisation à désactiver certaines bandes pour un moindre coût. Cela permet de ré-allouer l'énergie libérée dans des bandes qui apportent une bonne contribution à l'intelligibilité ce qui est primordial à faible RSB tant la réserve en énergie disponible est basse. Cependant, les approximations des fonctions de coût $\hat{f}_i$ proposées par TAAL et al., dont les allures sont visibles figure 5.5, ont un comportement qui diffère grandement des fonctions de coût d'origine f_i. L'amélioration du SII obtenue en fonction du RSB n'est donc pas nécessairement maximale, ni lisse, puisque c'est une approximation assez différente du critère qui est maximisée, cela explique donc les résultats en escalier obtenus pour les RSB intermédiaires. Enfin, à fort RSB, un autre phénomène se manifeste. L'absence de borne supérieure dans les approximations des fonctions de coût $\hat{f}_i$ qui tendent de façon asymptotique vers I_i suppose qu'il est toujours légèrement rentable d'investir de l'énergie dans les bandes au delà de D_i^+, ce qui est totalement faux du point de vue du SII. Ainsi, l'approche peut trouver qu'il est plus rentable d'investir de l'énergie au delà de D_i^+ dans certaines bandes plutôt que d'investir ce surplus d'énergie dans d'autres bandes qui contribueraient alors à l'amélioration du SII.

Approximation non-linéaire non-concave : STANTON et al.

Finalement, l'approche de STANTON et al. aura des améliorations du SII toujours plus grande que celle de TAAL et al. puisqu'ils prennent leurs spectres comme point initialisation du problème d'optimisation. Ainsi, l'approche de STANTON et al. dépend grandement des résultats de celle de TAAL et al. et hérite de ses dysfonctionnements. En effet, on remarque que, pour des RSB intermédiaires, l'affinement des spectres n'est pas très performant. On note parfois une légère amélioration, dûe aux approximations des fonctions de coût beaucoup plus proches des fonctions originales, mais il reste une marge importante avec l'amélioration optimale du SII. Comme nous l'avons vu juste précédemment, les bandes activées par l'approche TAAL et al. à un RSB donné ne sont pas nécessairement celles permettant de maximiser le SII à ce niveau, ainsi le problème d'optimisation va souvent être initialisé avec les mauvaises bandes désactivées. La dérivée des approximations $\hat{f}_i$ de STANTON et al. étant quasiment nulle dans ces bandes, l'algorithme d'optimisation différentielle ne pousse jamais à l'activation de celles-ci et c'est donc des maxima locaux, non-optimaux, qui sont souvent atteints. Par contre, à fort RSB, toutes les bandes sont activées à l'initialisation, ainsi l'algorithme d'optimisation est beaucoup plus propice à converger vers le maximum global. C'est bien ce que nous observons avec des améliorations du SII quasi-optimales à partir d'un RSB d'environ 0 dB pour tous les bruits.

Extension proposée

Le problème majeur de la méthode de STANTON et al. étant lié à une initialisation parfois inadaptée, nous proposons d'étudier les résultats lorsque nous utilisons plutôt les spectres optimaux obtenus avec la méthode de SAUERT et al. comme point de départ. En effet, nous avons vu que l'absence d'une borne inférieure dans cette dernière force à investir de l'énergie dans toutes les bandes, il y aura donc aucune bande complètement vide lors de l'initialisation du problème. Cela pourra alors aider l'algorithme d'optimisation différentielle à exploiter des bandes que la méthode de TAAL et al. aurait condamnées et à proposer une distribution peut-être plus adaptée. Les améliorations du SII obtenues par l'approche de STANTON et al. pour les deux initialisations sont visibles figure 5.10. Nous remarquons que nos suppositions sont vérifiées, malgré des performances initiales de l'approche de SAUERT et al. bien moins importantes que celle de TAAL et al. pour des RSB intermédiaires, le fait qu'aucune bande ne soit vide, lors de l'initialisation de l'algorithme d'optimisation de STANTON et al., permet d'approcher encore plus les résultats optimaux. Pour un RSB donné, ces deux initialisations complémentaires pourraient alors être utilisées séparément puis prendre le meilleur des deux spectres optimaux obtenus.



(a) contrainte dB(Z)

(b) contrainte dB(A)

(c) contrainte dB(Z)

(d) contrainte dB(A)

(e) contrainte dB(Z)

(f) contrainte dB(A)

FIGURE 5.10 – Amélioration du SII, entre un spectre de parole de référence normalisé et les spectres de parole issus de l'approche de STANTON et al. avec deux initialisations différentes, en fonction du RSB pour chaque bruit classique et pour chaque contrainte : dB(Z) à gauche et dB(A) à droite. Une initialisation est effectuée avec les spectres de SAUERT et al. (init. Sauert) et l'autre avec les spectres de TAAL et al. (init. Taal).

5.4.4 Exploitation des résultats pour le traitement des signaux de parole

Pour un bruit, un auditeur et un niveau de parole donnés, nous sommes maintenant capable de calculer le spectre équivalent optimal qui maximise le SII. Il reste alors à détailler comment exploiter ce spectre pour traiter efficacement un signal de parole diffusé dans ces conditions afin d'améliorer son intelligibilité.

L'objectif du traitement est de faire en sorte que le signal traité possède un spectre équivalent long terme identique au spectre optimal calculé. Cette égalisation fréquentielle peut s'effectuer en utilisant un banc de filtres qui décompose le signal $x(n)$ en composantes $x_i(n)$ sur les 18 bandes de tiers-d'octave exploitées par le SII. Le signal traité $y(n)$ est alors synthétisé en sommant les composantes préalablement multipliées par un gain g_i à savoir :

$$y(n) = \sum_i g_i \cdot x_i(n)\,. \tag{5.63}$$

Pour calculer les gains à appliquer aux différentes composantes, il suffit de calculer les niveaux du spectre équivalent E_i du signal de parole :

$$\mathrm{E}_i = 10 \cdot \log\left(\frac{1}{b_i} \cdot \frac{1}{\mathrm{N}} \cdot \sum_{n=1}^{\mathrm{N}} x_i^2(n)\right), \tag{5.64}$$

puis la contrainte visant à ce que les niveaux du spectre équivalent du signal traité soient égaux aux niveaux E_i^{opt} du spectre équivalent optimal s'exprime par :

$$\mathrm{E}_i^{opt} = 10 \cdot \log\left(\frac{1}{b_i} \cdot \frac{1}{\mathrm{N}} \cdot \sum_{n=1}^{\mathrm{N}} (g_i \cdot x_i(n))^2\right), \tag{5.65}$$

et on trouve alors :

$$g_i = 10^{(\mathrm{E}_i^{opt} - \mathrm{E}_i)/20}. \tag{5.66}$$

Pour généraliser, nous pouvons considérer que nous n'avons pas, ou difficilement, accès au spectre équivalent long terme de parole du locuteur, notamment pour une application en temps réel ou pour un signal avec plusieurs locuteurs par exemple. Les gains pourront alors être calculés à partir du spectre équivalent de parole de référence de niveaux U'_i normalisé au bon niveau.

Conclusion du chapitre 5

Dans ce chapitre, nous avons posé formellement le problème de maximisation du *Speech Intelligibility Index* (SII) sous contrainte énergétique et proposé une résolution exacte basée sur une division en sous-problèmes convexes. En réponse à la problématique présentée chapitre 4 sur l'intérêt d'une contrainte perceptive plutôt qu'une contrainte énergétique simple, nous avons aussi introduit dans la procédure d'optimisation la possibilité d'utiliser une contrainte pondérée.

Après avoir détaillé les différentes approches de maximisation du SII déjà existantes, basées sur des approximations du critère, et les avoir étendues à l'utilisation de contraintes pondérées, nous avons comparé et analysé en détails les spectres obtenus sur trois bruits classiques (blanc, bleu et SSN), sous deux contraintes énergétiques (simple et pondération A). Nous avons alors mis en évidence et expliqué les raisons pour lesquelles les améliorations du SII des approches par approximations ont des comportements et des performances très diversifiées en fonction du RSB sur les bruits classiques étudiés. En proposant une légère extension de la méthode de Stanton et al., exploitant les résultats de celle de Sauert et al. et de Taal et al., nous obtenons des améliorations quasi-optimales sur toute la plage de RSB.

Cette extension permet de résoudre le problème de maximisation du SII avec un nombre de calculs significativement plus faible qu'avec la procédure de résolution exacte proposée, sans perte significative de performances dans les bruits considérés. Cela peut être utile dans des applications où l'on souhaiterait faire ce calcul en temps réel sur du matériel équipé d'une faible puissance de calcul, dans un véhicule automobile par exemple.

Chapitre 6

Maximisation exacte d'un critère d'intelligibilité sous contrainte énergétique classique : *Hurricane Challenge 2*

Sommaire

Retour à la table des matières

Introduction du chapitre 6

Afin de tester les performances de la méthode de maximisation exacte du SII, nous avons participé à la deuxième édition du *Hurricane Challenge* [162] qui vise à évaluer et comparer plusieurs algorithmes de renforcement de la parole dans des conditions d'écoute dégradées. Le challenge consistait à modifier des signaux de parole enregistrés dans plusieurs langues, afin d'améliorer leur intelligibilité dans un bruit de cafétéria. Une nouveauté importante de cette édition est l'ajout de différentes conditions de réverbération afin d'évaluer la robustesse des traitements face à ce genre de dégradation. Des tests d'intelligibilité à grande échelle ont alors été menés afin de mesurer le gain d'intelligibilité des méthodes participantes.

La contrainte énergétique imposée repose sur l'énergie globale du signal et ne prend pas en compte d'échelle perceptive, il n'a donc pas été possible de tester les performances de la méthode avec notre nouvelle contrainte. Il était tout de même intéressant d'y participer afin d'étudier les performances de l'optimisation exacte, ainsi que l'influence de facteurs rarement étudiés que sont la langue et la réverbération.

La présentation générale du challenge sera détaillée dans la section 6.1, avec le protocole mis en place et les différents algorithmes qui ont participé. Les résultats de la maximisation du SII obtenu dans les conditions acoustiques du challenge seront ensuite présentés dans la section 6.2. Enfin, une présentation et une interprétation des résultats du challenge seront proposés dans la section 6.3.

6.1 Présentation du challenge

6.1.1 Mise en place et instructions

Des listes d'environ 100 phrases construites à partir d'une matrice de (5 x 10) mots, permettant d'obtenir des phrases grammaticalement correctes mais sémantiquement imprévisibles, ont été enregistrées par trois locuteurs de genre masculin chacun avec une langue différente, à savoir : allemand, anglais et espagnol. L'objectif du challenge était d'améliorer l'intelligibilité de ces signaux en prévision de leur dégradation par un bruit de conversation (avec plusieurs voix concurrentes) et par l'introduction de trois conditions de réverbération. Des tests perceptifs ont alors été mis en place afin de mesurer le score d'intelligibilité à trois niveaux de présentation fixés, pour chaque condition de réverbération, sur trois sites différents i.e. un pour chaque langue.

Pour assurer la répétabilité de l'expérience sur les trois sites, les conditions acoustiques ont été créées à partir d'enregistrement du bruit masquant et de réponses impulsionnelles, dans une salle au temps de réverbération d'environ 0,8s. Les différentes conditions de réverbération ont alors été obtenues en prenant les mesures sur une tête acoustique à une distance plus ou moins éloignée de la source : proche (1m), intermédiaire (2,5m) et lointaine (4m). Pour chaque condition de réverbération, les trois niveaux de présentation choisis ont été obtenus par des tests préliminaires et correspondent approximativement à des scores d'intelligibilité de 25% (RSB faible), 50% (RSB intermédiaire) et 75% (RSB élevé).

Les instructions du challenge nous invitaient à traiter l'ensemble des signaux de parole, indépendemment pour chaque condition (Langue x Réverbération x RSB), afin d'améliorer leur intelligibilité. Des exemples de bruits et de réponses impulsionnelles de la salle relativement proches, mais différents, de ceux utilisés lors des tests étaient mis à notre disposition. De plus, les valeurs exactes des RSB qui était fixées pour chaque condition (Langue x Réverbération) nous avaient aussi été communiquées.

6.1.2 Participations au challenge

Il y a eu 9 participations au challenge, leur dépendance au bruit et à la réverbération est indiquée dans le tableau 6.2, en voici les descriptions synthétiques :

- ACO [17] : Cet algorithme est une combinaison séquentielle de l'algorithme AdaptDRC [174], puis de l'algorithme OE (pour *Onset-Enhancement*) [72]. AdaptDRC se base sur une ré-allocation spectro-temporelle des bandes d'octave du signal en deux étapes : d'abord une amplification visant à maximiser le SII en se

basant sur une version modifiée des travaux de SAUERT et al., puis une étape de compression dynamique adaptative au bruit. OE cherche à réduire le masquage par recouvrement de la parole, ainsi qu'à renforcer ses attaques, afin d'améliorer l'intelligibilité dans les milieux réverbérants.

- ASE [39] : Cet algorithme ne prend en considération que le signal de parole lui même. Il se base aussi sur une ré-allocation spectro-temporelle des bandes d'octave du signal en deux étapes : d'abord une étape de compression dynamique, puis une amplification basée sur des connaissances liées à la perception auditive. Une compression dynamique large bande est finalement appliquée sur le signal résultant. Dans la version préliminaire qui a été proposée pour le challenge, les paramètres de compression et d'amplification ont été basés sur une expertise des stimuli et ont été fixés pour tous les signaux traités.
- exactMaxSII [67] : Notre approche basée sur une égalisation des bandes de tiers-d'octave, fixe pour chaque condition (Langue x Réverbération x RSB), visant à maximiser le SII de façon exacte.
- DeepSSC-Lomb [65] : Notre approche de conversion du style de la parole visant à imiter la parole Lombard qui sera détaillée chapitre 10. L'exploitation de la décomposition en ondelettes pour décrire certaines caractéristiques, couplée à l'utilisation de modèles récurrents adapté à l'analyse de séquences temporelles, permet d'améliorer objectivement l'apprentissage des transformations à appliquer. La version préliminaire qui a participé au challenge présentait de nombreux artefacts audibles.
- DSSC-L/eMSII : Combinaison séquentielle de nos deux autres entrées, à savoir DeepSSC-Lomb puis exactMaxSII.
- iMetricGAN [109] : Cet algorithme est composé d'un générateur (G) et un discriminateur (D). D essaye de prédire des scores d'intelligibilité (SIIB [30] et ESTOI [31]) des signaux de parole, et oriente alors G afin de traiter les signaux de manière à maximiser les scores d'intelligibilités prédits. G reçoit les signaux non-traités et génère des facteurs qui modifie le spectrogramme, obtenu par TFCT, point par point. Le signal traité est alors re-synthétisé par TFCT inverse.
- MS500 : L'algorithme cherche à estimer la FTM de l'environnement d'écoute afin de compenser l'étalement provoqué par la réverbération sur le spectre de modulation. L'inverse de la FTM étant difficilement obtenable, MS500 modifie le spectre de modulation de la parole naturelle sur certaines fréquences déterminantes à partir de relations entre le spectre de modulation original, la FTM estimée et le spectre de modulation dégradé.
- IISPA [173] : L'algorithme IISPA, pour *Intelligibility-Improving Signal Processing Approach*, consiste à optimiser des paramètres de traitement avec un modèle de reconnaissance automatique de la parole. Les paramètres d'optimisation sont les fréquences limites d'un filtre passe-bande, la pente et la courbe spectrales, et les paramètres de compression ou expansion du spectre de modulation.
- SSDRC [221] : Cet algorithme de référence en renforcement direct de la parole sans prise en compte du bruit a été détaillé chapitre 4. Il consiste à façonner le spectre en combinant du filtrage fixe (pré-accentuation) et adaptatif au voisement (pré-accentuation et affinement des formants), puis à appliquer une compression dynamique. Au regard de ses excellentes performances lors de la première édition du challenge, il a été introduit dans cette deuxième édition comme base de référence.

6.1.3 Évaluations subjectives

Les évaluations subjectives se sont déroulées sur trois sites : Oldenbourg en Allemagne, Édimbourg en Écosse et Vitoria-Gasteiz en Espagne. Sur chaque site, des sujets normo-entendants natifs de la langue du pays ont participé aux tests avec 62 sujets pour l'allemand, 62 pour l'anglais et 63 pour l'espagnol. Chaque sujet a entendu 2 phrases pour chaque condition (Réverbération x RSB x Traitement), le traitement comprend toutes les participations au challenge ainsi que la parole naturelle, pour un total de 180 phrases par sujet. Les présentations ont suivi un plan de mesures répétées et les sujets devaient indiquer les mots reconnus en les sélectionnant sur un écran présentant la matrice (5x10) de mots.

Un détail très important à noter est que les RSB communiqués initialement aux participants, ont été ajustés durant le test préliminaire final afin d'avoir un meilleur échantillonnage des fonctions psychométriques de la parole naturelle pour les différentes conditions. Les RSB utilisés, ainsi que ceux initialement communiqués, pour chaque condition sont consultables dans le tableau 6.1. Pour les conditions de réverbération proche, il n'y a pas eu de changement pour l'allemand et des légers changements de -1 dB pour l'anglais et +1 dB pour l'espagnol. Pour les conditions de réverbération intermédiaire, les changements sont très notables avec -4 dB pour l'allemand, -7 dB pour l'anglais et -5 dB pour l'espagnol. De même pour les conditions de réverbération lointaine avec -4 dB pour l'allemand, -8 dB pour l'anglais et -7 dB pour l'espagnol. On remarque donc une diminution globale et diversifiée des niveaux de présentation prévus pour les différentes conditions de réverbération, excepté pour la condition proche qui reste relativement proche des RSB communiqués initialement. Le changement de niveau de présentation est très important pour notre méthode qui est extrêmement sensible à ce facteur. Les traitements proposés ayant été paramétrés pour les RSB initiaux, il faudra donc prendre en compte cette information lors de l'analyse des résultats.

Langue	Réverbération	RSB(dB) utilisé			RSB (dB) communiqué			Δ RSB
		Faible	Inter.	Élevé	Faible	Inter.	Élevé	
Allemand	Proche	-15,0	-12,5	-10,0	-15,0	-12,5	-10,0	+0,0 dB
	Intermédiaire	-13,0	-10,0	-7,0	-9,0	-6,0	-3,0	-4,0 dB
	Lointaine	-13,0	-9,0	-5,0	-9,0	-5,0	-1,0	-4,0 dB
Anglais	Proche	-13,0	-8,5	-4,0	-12,0	-7,5	-3,0	-1,0 dB
	Intermédiaire	-11,0	-5,0	1,0	-4,0	2,0	8,0	-7,0 dB
	Lointaine	-10,0	-4,0	2,0	-2,0	4,0	10,0	-8,0 dB
Espagnol	Proche	-17,5	-14,5	-11,5	-18,5	-15,5	-12,5	+1,0 dB
	Intermédiaire	-17,0	-14,0	-11,0	-12,0	-9,0	-6,0	-5,0 dB
	Lointaine	-18,0	-14,0	-10,0	-12,0	-8,0	-4,0	-7,0 dB

TABLEAU 6.1 – RSB utilisés, ainsi que ceux initialement communiqués aux participants, pour chaque condition d'écoute.

6.2 Maximisation du SII

Les niveaux du spectre équivalent long-terme N_i du bruit masquant dépendent des trois conditions de réverbération et de l'oreille d'écoute (gauche ou droite). Lors des enregistrements, les bruits ayant été diffusés face aux quatre coins de la salle, la position de l'enregistrement ne devrait pas influencer grandement le spectre long-terme du bruit. Nous avons effectivement remarqué que les niveaux des six spectres équivalents calculés étaient très proches, nous avons donc décidé de travailler avec un spectre équivalent de bruit moyen visible figure 6.1 et son spectre équivalent de masquage correspondant. Le niveau du bruit est de 54 dB SPL / 54 dB(Z) / 53 dB(A).

6.2.1 Spectres optimaux

Tout d'abord, nous vérifions que les niveaux du spectre équivalent de masquage ne dépassent jamais les niveaux du spectre équivalent de bruit de plus de 9 dB. Ainsi l'hypothèse 2 est respectée et ce bruit est bien éligible à la procédure d'optimisation exacte. Ensuite, nous remarquons directement que ce bruit présente un creux de densité spectrale autour de 1 kHz où la FIB est maximale. Nous pouvons alors déjà prédire que la maximisation du SII va chercher à amplifier ces bandes où l'oreille est sensible afin de faire émerger de l'information au dessus du masquage. Les spectres équivalents optimaux en fonction du RSB sont visibles 6.2 et c'est effectivement ce qui se passe. Pour de très faibles RSB, les bandes hautes fréquences où le bruit est quasiment absent sont forcément favorisées, mais assez rapidement ce sont bien les bandes autour de 1 kHz qui le deviennent, suivies des bandes autour de 300 Hz où on remarque aussi un creux d'énergie dans le spectre équivalent de masquage.

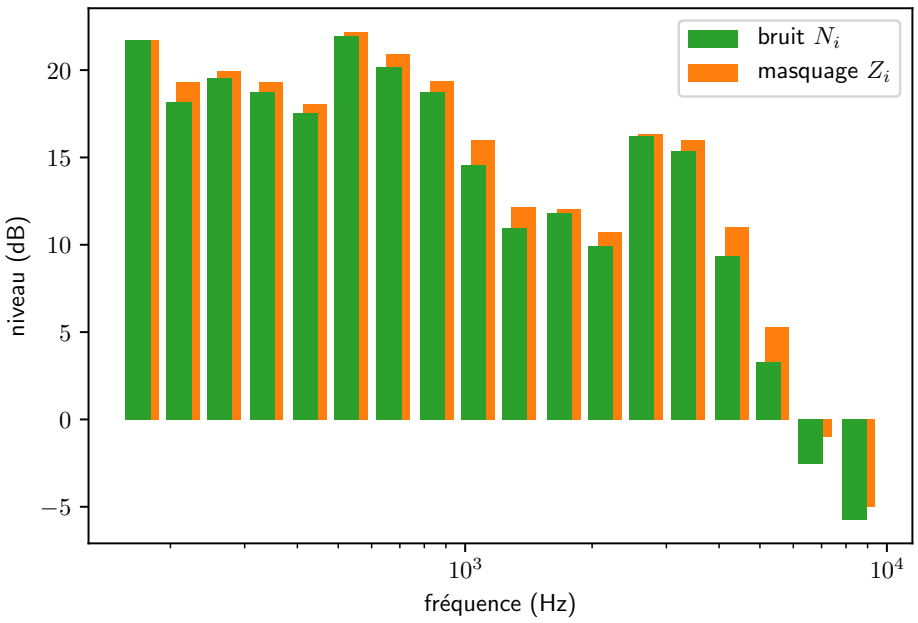


FIGURE 6.1 – Spectre équivalent du bruit de conversation utilisé pour le challenge et son spectre équivalent de masquage correspondant.

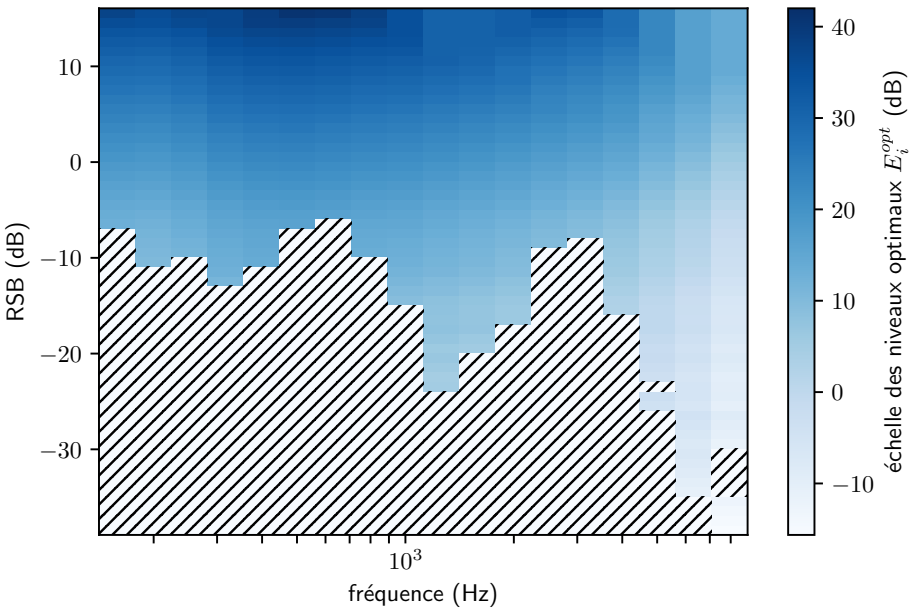


FIGURE 6.2 – Niveaux E_i^{opt} des spectres équivalents optimaux en fonction du RSB dans le bruit de conversation utilisé pour le challenge.

6.2.2 Amélioration du SII

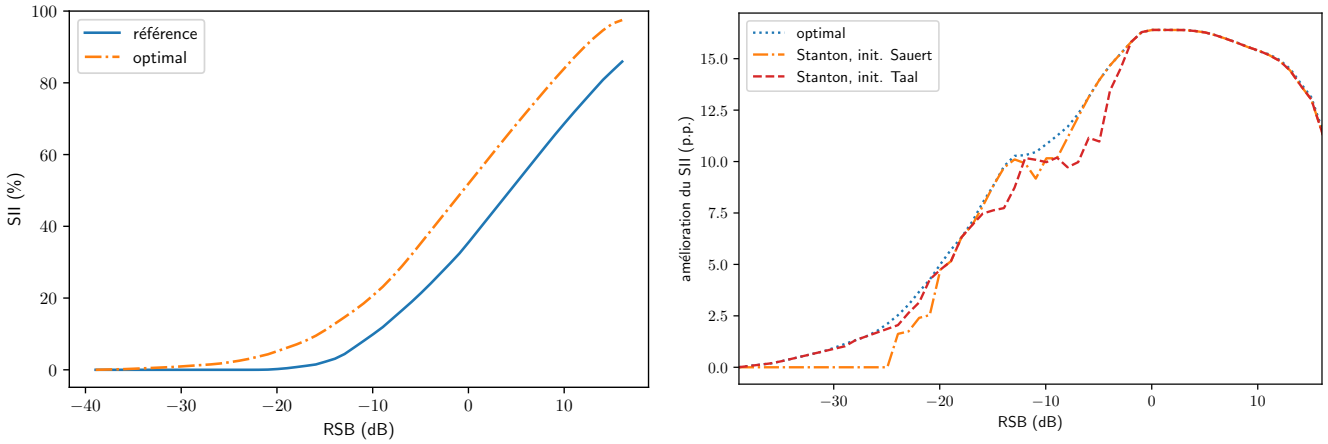
Nous pouvons aussi observer les SII optimaux par rapport au SII obtenu avec le spectre de référence normalisé sur un ensemble de RSB, ainsi que les améliorations correspondantes, figure 6.3. On note une amélioration croissante du SII jusqu'à 15 p.p. atteint vers un RSB de 0 dB. Les tests se déroulant entre -20 dB et 0 dB, l'intelligibilité devrait donc théoriquement être bien rehaussée.

Les améliorations obtenues avec la méthode de STANTON et al., pour les deux initialisations décrites section 5.4.3, sont aussi visibles sur la figure 6.3. On remarque encore une fois la complémentarité des deux initialisations pour ce nouveau bruit, en prenant le meilleur résultat des deux initialisations, on approche de près les résultats optimaux sur l'ensemble des RSB.

6.2.3 Calcul des gains

À partir de l'ensemble des signaux, nous calculons les niveaux E_i du spectre équivalent de chaque langue. Pour chacune des 27 conditions d'écoute, nous récupérons les niveaux E_i^{opt} des spectres optimaux au RSB correspondant, ainsi que les gains à appliquer grâce à l'équation 5.66.

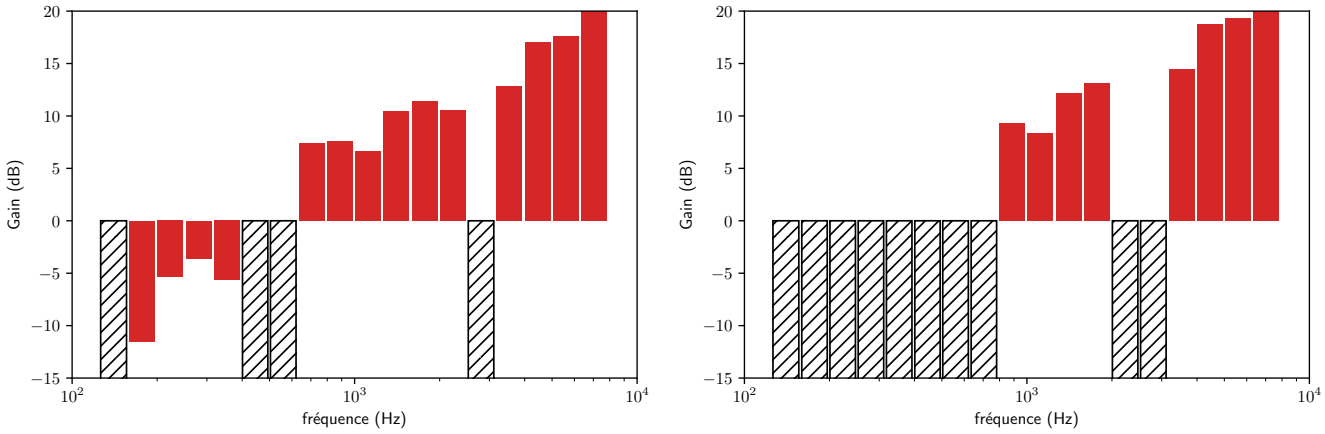
Des exemples de gains obtenus pour la condition (Espagnol x Réverbération intermédiaire x RSB Intermédiaire) pour le RSB initialement communiqué (-9 dB), ainsi que pour le RSB utilisé lors du test d'intelligibilité (-14 dB), est visible figure 6.4. En comparant les gains utilisés à ceux qui auraient dû l'être, l'influence du changement de niveau de présentation sur la méthode devient flagrant. Dans cette condition, avec une valeur de RSB bien plus faible que celle annoncée initialement, du point de vue de la maximisation du SII il est contre productif de conserver de l'énergie dans toutes les bandes qui seront, pour la plupart, totalement masquées par le bruit. Certaines composantes du signal auraient donc dû être annulées de façon à libérer de l'énergie à investir dans les bandes prioritaires afin de les faire émerger du bruit.



(a) Comparaison du SII en fonction du RSB.

(b) Amélioration du SII en p.p. en fonction du RSB.

FIGURE 6.3 – Comparaison du SII obtenu avec un spectre de parole de référence normalisé et celui obtenu avec le spectre optimal de parole dans le bruit de conversation utilisé pour le challenge (à gauche). Amélioration du SII vis-à-vis du spectre de parole de référence normalisé, pour le spectre optimal ainsi que les spectres de parole issus de l'approche de STANTON et al. avec les deux initialisations (à droite).



(a) RSB = -9 dB

(b) RSB = -14 dB

FIGURE 6.4 – Gains obtenus pour la condition (Espagnol x Réverbération intermédiaire x RSB faible), pour le RSB initialement communiqué (-9 dB), ainsi que pour le RSB utilisé lors du test d'intelligibilité (-14 dB).

6.3 Résultats des évaluations subjectives

Les différences entre les scores d'intelligibilité (pourcentage moyen de réponses correctes) des différentes participations et ceux de la parole naturelle pour toutes les conditions sont consultables dans le tableau 6.2 en p.p., les scores de référence de la parole naturelle sont donnés en italique. Les différences statistiquement non-significatives ont un fond gris clair, les diminutions statistiquement significatives ont un fond gris foncé et les augmentations statistiquement significatives ont un fond blanc. Enfin, les valeurs en gras correspondent à la meilleure augmentation pour chaque condition. Ces différences sont aussi visibles graphiquement sur la figure 6.5, pour une meilleure visibilité nous avons retiré nos deux autres participations DeepSSC-Lomb et DSSC-l/eMSII pour une raison que nous justifions ci-après.

6.3.1 Analyse préliminaire de nos participations

La première chose que nous remarquons sont les résultats catastrophiques de notre participation DeepSSC-Lomb basée sur de la conversion de voix naturelle vers voix Lombard. L'algorithme de conversion était encore dans une version préliminaire et la présence de nombreux artefacts, clairement audibles, est très probablement responsable de cette baisse importante d'intelligibilité. Au contraire, notre méthode exactMaxSII basée sur la maximisation exacte du SII améliore quasi-systématiquement l'intelligibilité de façon significative. Enfin, les résultats du couplage de nos deux méthodes dans la participation DSSC-L/eMSII montre que la maximisation du SII permet de compenser les dégradations introduites par la conversion de voix pour atteindre des scores d'intelligibilité proche de la parole naturelle.

Une amélioration de l'algorithme de conversion de voix utilisé est donc primordial pour pouvoir, à l'avenir, étudier l'intérêt d'une telle approche. Ainsi, nous ne traiterons pas les résultats des approches DeepSSC-Lomb et DSSC-l/eMSII par la suite.

		Bruit ?	Réverb. ?	Réverb. proche			Réverb. intermédiaire			Réverb. lointaine		
				RSB			RSB			RSB		
				Faible	Inter.	Élevé	Faible	Inter.	Élevé	Faible	Inter.	Élevé
Allemand	*Parole naturelle (score)*			*11,1*	*40,8*	*64,9*	*15,4*	*44,6*	*70,3*	*12,3*	*41,0*	*76,7*
	ACO	✓	✓	2,5	-0,8	0,0	9,3	10,7	6,4	9,0	22,6	8,2
	ASE	×	×	**50,0**	**45,9**	**29,7**	**43,8**	**44,3**	**26,7**	**31,8**	**44,4**	**20,7**
	exactMaxSII	✓	×	41,5	30,8	14,4	31,5	13,6	2,6	16,6	14,9	11,5
	DeepSSC-Lomb	×	×	-5,7	-31,5	-37,2	-10,2	-31,5	-36,6	-8,9	-23,4	-24,8
	DSSC-L/eMSII	✓	×	25,6	7,4	-10,5	-2,0	-22,6	-27,2	-5,1	-18,5	-33,6
	iMetricGAN	✓	×	47,0	33,6	21,8	25,7	29,3	19,2	13,6	23,0	8,7
	MS500	✓	✓	13,1	-2,8	-8,9	2,3	-7,5	-3,4	-4,1	-5,7	-4,8
	IISPA	✓	✓	43,6	31,3	20,3	27,5	21,8	6,6	17,5	12,0	-1,1
	SSDRC	×	×	47,0	42,1	29,3	36,4	39,3	21,5	20,8	33,9	16,7
Anglais	*Parole naturelle (score)*			*7,3*	*18,5*	*50,5*	*13,8*	*43,8*	*73,5*	*18,0*	*42,7*	*75,8*
	ACO	✓	✓	-0,5	8,3	17,8	12,8	26,2	14,0	17,5	27,5	15,8
	ASE	×	×	6,8	**42,8**	**40,5**	**27,0**	**42,7**	**23,2**	**22,8**	**42,0**	**18,8**
	exactMaxSII	✓	×	**10,0**	22,8	18,3	10,3	21,8	12,0	4,0	19,0	9,0
	DeepSSC-Lomb	×	×	-4,3	-10	-14,2	-4,2	-12,7	-8,7	-7,0	-6,0	-12,3
	DSSC-L/eMSII	✓	×	2,0	7,8	-1,3	-2,0	-4,3	1,8	-6,3	1,3	-2,7
	iMetricGAN	✓	×	6,7	27,8	27,5	18,5	26,8	14,8	13,5	34,0	13,5
	MS500	✓	✓	2,3	12,0	11,0	9,3	15,3	8,7	7,5	16,2	6,5
	IISPA	✓	✓	3,7	13,7	9,3	1,0	1,8	-2,5	-2,8	1,5	-9,2
	SSDRC	×	×	9,7	31,3	36,7	21,3	31,2	19,5	18,2	34,3	17,7
Espagnol	*Parole naturelle (score)*			*14,8*	*43,2*	*66,3*	*12,7*	*25,5*	*52,5*	*6,8*	*7,5*	*55,0*
	ACO	✓	✓	4,3	-2,7	6,3	1,2	13,0	12,8	0,8	11,0	19,8
	ASE	×	×	**58,8**	43,7	**29,0**	**46,7**	**56,5**	**41,2**	**30,2**	**45,7**	**38,2**
	exactMaxSII	✓	×	23,7	17,2	19,2	32,7	34,2	27,5	22,2	17,7	25,2
	DeepSSC-Lomb	×	×	-11,2	-36,8	-46	-10,8	-16,3	-34,7	-5,7	-21,7	-32,5
	DSSC-L/eMSII	✓	×	2,8	-0,2	-6,3	11,2	10,7	-2,2	7,5	-4,5	-6,8
	iMetricGAN	✓	×	42,2	34,0	23,5	28,0	37,0	23,2	15,2	23,0	15,8
	MS500	✓	✓	17,7	4,5	12,3	7,5	17,8	12,3	4,2	9,0	11,8
	IISPA	✓	✓	41,7	35,5	20,0	30,2	37,5	25,0	17,7	18,5	14,5
	SSDRC	×	×	49,0	**44,3**	27,8	39,0	49,2	39,8	15,0	32,0	28,0

TABLEAU 6.2 – Différences entre les scores des traitements et ceux de la parole naturelle (donnés en italique) en p.p. pour toutes les conditions.

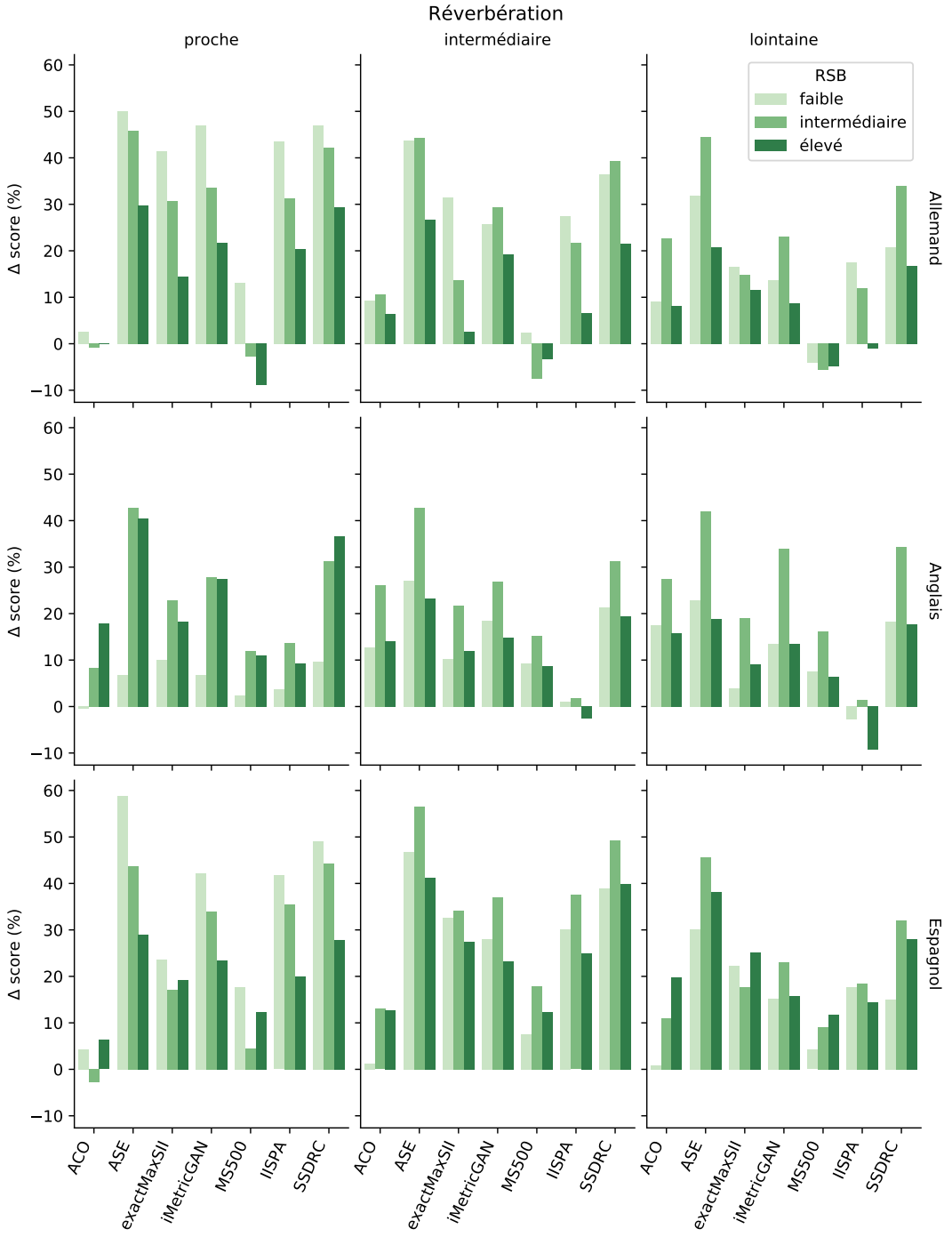


FIGURE 6.5 – Différences entre les scores des traitements et ceux de la parole naturelle en p.p. pour toutes les conditions. Les différences statistiquement non-significatives ont un fond gris clair, les diminutions statistiquement significative ont un fond gris foncé et les augmentation statistiquement significative ont un fond blanc. Les valeurs en gras correspondent à la meilleure augmentation pour chaque condition.

6.3.2 Présentation des résultats

Scores bruts

Nous remarquons que notre méthode engendre une amélioration significative de l'intelligibilité pour l'ensemble des conditions d'écoute, à part pour la condition (Anglais x Réverbération lointaine x RSB Faible) où le gain de 4 p.p. n'est pas significatif. De manière globale, les gains engendrés par notre méthode se placent dans la moyenne des autres participations.

On remarque aussi que dans les conditions de réverbération proche, à de faibles RSB, notre méthode obtient le meilleur score pour la langue anglaise avec un gain de 10 p.p. équivalent à l'algorithme de référence SSDRC qui obtient un gain de 9,7 p.p.. Pour les autres langues, dans la condition (Réverbération proche x RSB faible), les performances de notre méthode sont moindres avec un gain relativement moyen de 41,5 p.p. pour l'allemand et un gain relativement faible de 23,7 p.p. pour l'espagnol. Toujours en réverbération proche, mais pour des RSB plus importants, notre méthode obtient des performances plus mitigées.

Estimation du SRP

À partir des score moyens pour les trois RSB de chaque condition (Langue x Réverbération x Traitement), il est possible d'estimer des courbes psychométriques permettant d'observer l'évolution théorique des scores en fonction du RSB. Des exemples de courbes psychométriques obtenues pour la parole naturelle allemande, et pour chaque condition de réverbération, sont visibles figure 6.6. On observe un aplatissement logique des courbes psychométriques en fonction de l'intensité de la réverbération.

Avec ces courbes il est possible d'estimer le SRP qui correspondant à un score d'intelligibilité de 50%. La différence entre le SRP de la condition (Langue x Réverbération x Parole naturelle) et SRP de chaque condition (Langue x Réverbération x Traitement) permet d'obtenir une mesure de l'amélioration de l'intelligibilité plus facilement interprétable que les scores moyens qui dépendent grandement de la pente de la courbe psychométrique. Une différence positive pour un traitement correspond à une diminution du SRP et donc à une amélioration de l'intelligibilité. Les différences entre les SRP des traitements par rapport à ceux de la parole naturelle pour les différentes conditions sont visibles figure 6.7 pour toutes les conditions (Langue x Réverbération x Traitement). On observe toujours que notre méthode présente globalement des améliorations comparables aux autres participations.

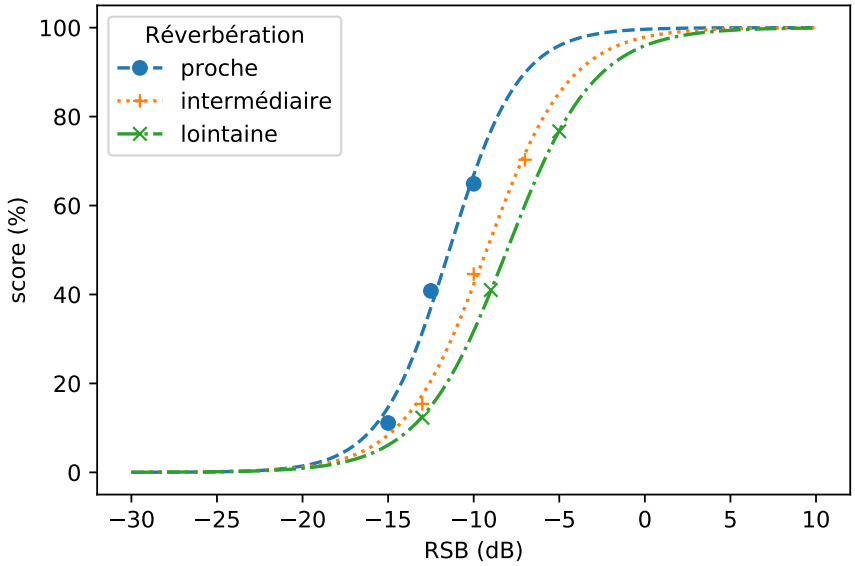


FIGURE 6.6 – Courbes psychométriques estimées à partir des scores moyens de la langue allemande sur les trois RSB des différentes conditions de réverbération.

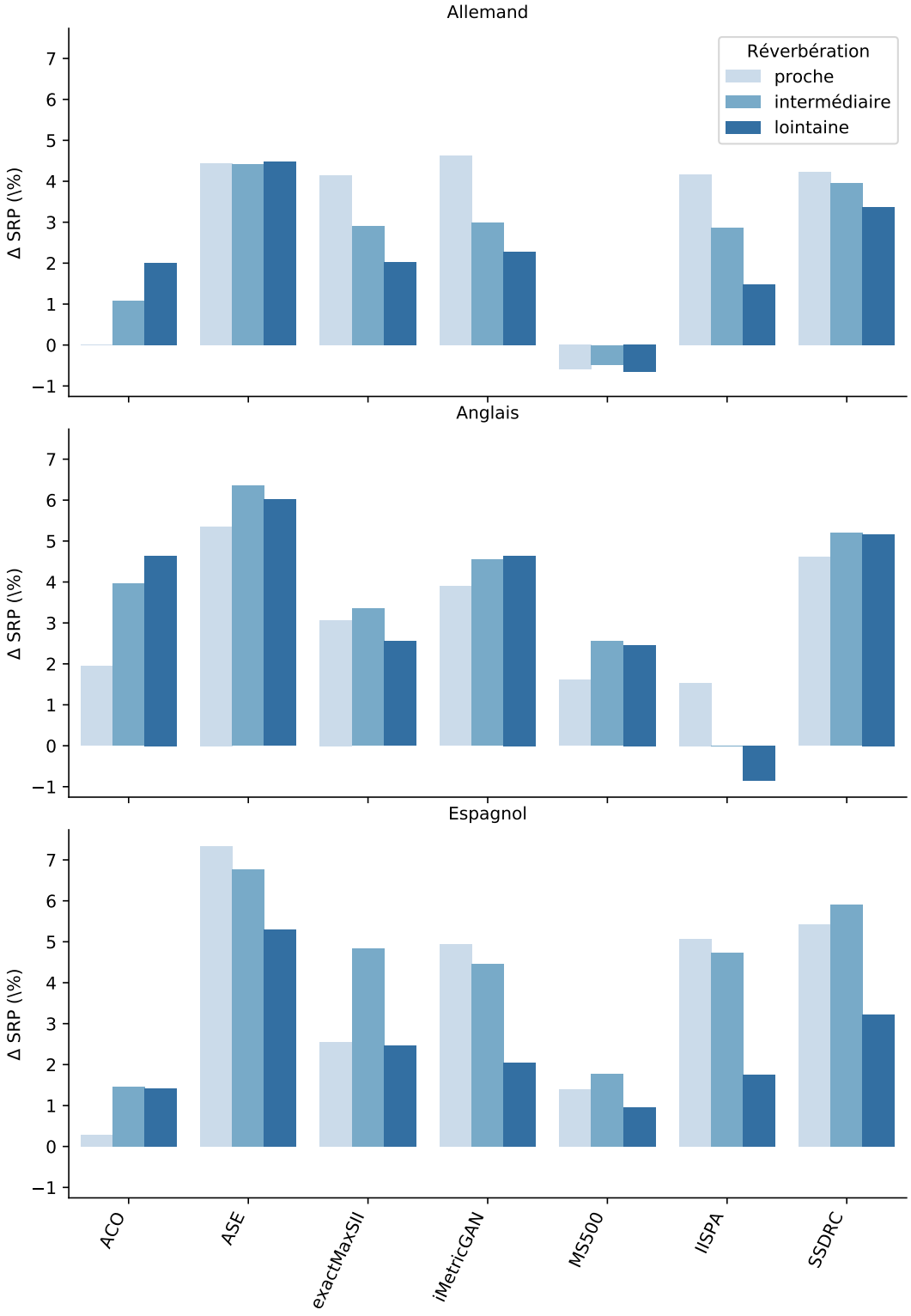


FIGURE 6.7 – Différences entre les SRP des traitements par rapport à ceux de la parole naturelle pour les différentes conditions.

Influence de la langue

Pour une réverbération proche, on remarque que notre méthode est plus performante pour l'allemand avec une amélioration du SRP de +4,1 dB, elle l'est un peu moins pour l'anglais avec une amélioration de +3,1 dB et elle l'est encore moins pour l'espagnol avec une amélioration de seulement +2,6 dB. On trouve d'autres tendances parmi les autres participations, par exemple, les résultats de la participation ASE présentent un comportement inverse avec des performances plus importantes pour l'espagnol (+7,3 dB), puis l'anglais (5,4 dB) et enfin l'allemand (+4,4 dB). Ainsi, le facteur de la langue semble interagir grandement avec les traitements et de façons très différentes.

Influence de la réverbération

Pour les conditions de réverbération intermédiaire et lointaine, les changements importants de niveaux de présentation prévus complexifient grandement l'analyse de nos résultats. En effet, les traitements appliqués ne correspondent pas exactement à ceux qu'aurait fournit notre approche avec les bons niveaux. Ces traitements conservent tout de même la tendance principale de l'approche qui est de renforcer l'énergie spectrale dans les bandes qui présentent une contribution importante à l'intelligibilité. Et malgré un paramétrage inadapté, les résultats entraînent toujours un gain d'intelligibilité notable. Pour l'allemand, qui est la langue dont les niveaux de présentation prévus ont le moins changé, une réverbération plus intense entraîne une diminution de l'amélioration du SRP ce qui semble cohérent vis-à-vis d'une méthode qui ne prend pas en compte la réverbération. Pour l'espagnol la réverbération intermédiaire se démarque avec une amélioration importante du SRP de +4,8 dB contre +2,5 dB pour les deux autres conditions qui l'entourent. Si on regarde les niveaux de présentation pour l'espagnol dans le tableau 6.1, le niveau prévu pour la réverbération intermédiaire était 5 dB supérieur à celui prévu pour la réverbération proche, le traitement était donc plus modéré avec moins de bandes désactivées, comme nous pouvons le voir figure 6.4. Cependant, les niveaux effectivement utilisés pour les deux conditions de réverbération sont très proches à 0,5 dB d'écart. Ainsi, les meilleures performances obtenues dans la condition réverbération intermédiaire seraient en grande partie liées au traitement plus modéré. Pour l'anglais, dont les niveaux de présentation prévus ont le plus changé, la réverbération ne semble pas avoir d'effet particulier, ce qui semble alors confirmer que des traitements plus doux entraînent de meilleures performances.

Concernant les autres participations, parmi celles qui prennent en compte la réverbération, ACO et MS500 présentent des résultats intéressants seulement pour l'anglais où l'amélioration du SRP est d'autant plus importante que la réverbération est intense, et IISPA est autant influencé négativement par la réverbération que les autres participations. Ce sont au contraire celles qui ne prennent pas en compte la réverbération, et qui sont déjà très efficaces en parole directe, qui semblent le plus robustes à ces dégradations, à savoir ASE et SSDRC.

6.3.3 Interprétation des résultats

Pour conclure, dans des conditions de réverbération proche, et pour toute les langues, notre approche produit des gains d'intelligibilité comparables aux autres algorithmes actuels de renforcement de la parole. Nous avons aussi remarqué que les algorithmes avec les meilleurs performances utilisent tous un module supplémentaire de compression dynamique et, comme nous l'avons vu chapitre 4, il est admis que cela procure systématiquement un gain d'intelligibilité notable. Il est donc fortement probable que l'utilisation d'un tel module couplé avec notre approche donnerait des gains d'intelligibilité encore plus intéressants. Il faudra tout de même faire attention à bien appliquer la compression avant le calcul des gains car celle-ci va venir modifier le spectre long-terme de la parole.

Enfin, le changement des niveaux de présentation annoncés nous a empêché d'analyser précisément l'influence de la réverbération sur la méthode d'optimisation exacte du SII. Cependant, cela nous a permis de mettre en évidence qu'un traitement trop brusque des signaux, bien qu'améliorant nettement l'intelligibilité, pourrait ne pas être le plus adapté. En effet, le SII mesure l'audibilité d'un signal de parole naturel et notre approche fait l'hypothèse majeure que la mesure est toujours pertinente après avoir transformé le signal, avec notamment

l'annulation de nombreuses bandes fréquentielles, qui n'a alors plus du tout les caractéristiques d'un signal naturel. Comme nous avons pu l'observer succinctement avec les résultats du challenge, notamment avec les conditions (Espagnol x Réverbération proche) et (Espagnol x Réverbération intermédiaire), cette hypothèse semble erronée. En effet, des traitements qui conservent de l'énergie dans des bandes supposées inutiles du point de vue du SII provoquent des gains d'intelligibilité supérieurs à ceux des modifications optimales qui annulent ces bandes pour renforcer celles plus importantes, dans des conditions d'écoute relativement proches. Il est difficile d'affirmer ces observations à cause de la grande variabilité des paramètres d'études (langue et réverbération), et, au contraire, de la faible variabilité de la source avec un locuteur unique pour chaque langue. Il faudrait donc mettre en place un protocole dédié pour confirmer ces suppositions qu'un traitement plus modéré pourrait alors effectivement procurer de meilleures performances.

Dans ce cas plusieurs possibilités d'extension de la méthode peuvent être imaginées. Comme il l'a été fait indirectement durant le challenge, paramétrer le traitement avec un niveau de référence supérieur à celui de l'écoute réelle, puis normaliser au bon niveau, est une possibilité. Du point de vue de l'optimisation, cela reviendrait finalement à translater toutes les fonctions de coût d'une constante vers des niveaux équivalent de parole plus faibles. Dans ce cas, les coefficients de distorsions L_i et les coefficients d'audibilité A_i sont tous deux décalés or la distorsion n'étant pas dépendante du bruit, une autre possibilité serait d'agir seulement sur les coefficients d'audibilité. Par exemple en posant :

$$\tilde{\mathrm{A}}_i(\mathrm{E}_i, \mathrm{D}_i) = \min(\max(\frac{\mathrm{E}_i - (\mathrm{D}_i + p_1)}{p_2}, 0), 1)\,, \tag{6.1}$$

avec p_1 le paramètre de décalage qui correspond au RSP à partir duquel la bande commence à contribuer à l'audibilité et p_2 le paramètre de largeur qui fixe sur quel intervalle du RSP la bande contribue linéairement à l'audibilité. Dans la définition du SII, on a $p_1 = -15\,\mathrm{dB}$, p_1 pourrait alors être diminué pour favoriser la distribution d'énergie dans plus de bandes. On a aussi $p_2 = 30\,\mathrm{dB}$, p_2 pourrait alors être diminué pour éviter un remplissage trop important de certaines bandes, ou il pourrait aussi être augmenté si on souhaite tout de même avoir quelques bandes pré-dominantes. Dans tous les cas, le choix et paramétrage de ces extensions demandera un protocole adéquat qui fera l'objet d'une étude complète.

Conclusion du chapitre 6

Dans ce chapitre, nous avons présenté les résultats de la deuxième édition du *Hurricane Challenge* auquel nous avons participé avec notre méthode de maximisation exacte du SII. Une nouveauté de cette deuxième édition était d'introduire différentes langues et conditions de réverbération pour étudier le comportement des algorithmes de renforcement de la parole face à ces facteurs. Malgré l'absence de compression dynamique, notre approche induit des gains d'intelligibilité comparables aux autres algorithmes de renforcement de la parole, pour une réverbération faible, et obtient même le meilleur gain d'intelligibilité pour la langue anglaise pour un faible RSB. C'est aussi le cas pour des conditions de réverbération plus importantes, mais un changement de dernière minute des niveaux de présentation annoncés par les organisateurs a complexifié l'analyse des résultats de notre méthode qui est très sensible au niveau de présentation. En revanche, les résultats obtenus avec des traitements paramétrés pour des conditions d'écoutes imprévues nous a permis d'aborder une perspective nouvelle pour notre approche.

En effet, l'utilisation du SII comme critère d'optimisation se base sur une hypothèse forte considérant que la mesure est toujours cohérente pour des signaux spectralement dénaturés. Certains résultats semblent montrer qu'il serait possible d'obtenir des gains d'intelligibilité plus importants en relaxant certains paramètres du SII de manière à éviter la désactivation précoces de certaines bandes qui participent encore à l'intelligibilité, même à très faible niveau. Des propositions d'extensions ont été faites et des tests plus orientés seront alors nécessaires pour vérifier ces suppositions.

Enfin, la contrainte énergétique du challenge portait sur l'énergie globale des signaux, nous n'avons donc pas pu tester la robustesse de notre approche face à une contrainte perceptive, cet aspect sera abordé dans le chapitre suivant avec des tests subjectifs que nous avons menés dans un contexte bruité automobile.

Chapitre 7

Maximisation exacte d'un critère d'intelligibilité sous contrainte énergétique pondérée en contexte bruité automobile

Sommaire

Retour à la table des matières

Introduction du chapitre 7

Comme nous l'avons introduit chapitre 4, les habitacles automobiles font partie des cas pratiques de renforcement de la parole où la contrainte énergétique classique consistant à maintenir les signaux au même niveau moyen peut être inadaptée. En effet, si le traitement augmente le niveau perçu des signaux, l'auditeur risque de ramener le volume sonore à un niveau plus confortable, pouvant alors engendrer une boucle rétroactive problématique.

Dans ce chapitre nous étudions donc l'intérêt d'un traitement visant à maximiser de manière exacte un critère objectif d'intelligibilité, le SII, sous contrainte perceptive afin d'améliorer l'intelligibilité de la parole dans un contexte bruité automobile. La section 7.1 présente le contexte d'étude et les résultats objectifs de la maximisation du SII pour trois types de bruit différents. La section 7.2 présente le protocole exigeant du test subjectif mis en place pour mesurer les performances de la méthode dans ces trois bruits habitacles automobiles. Enfin, la section 7.3 propose une interprétation des résultats, ainsi qu'une analyse des performances et des limites de la maximisation du SII sous contrainte perceptive.

7.1 Maximisation du SII dans un habitacle automobile

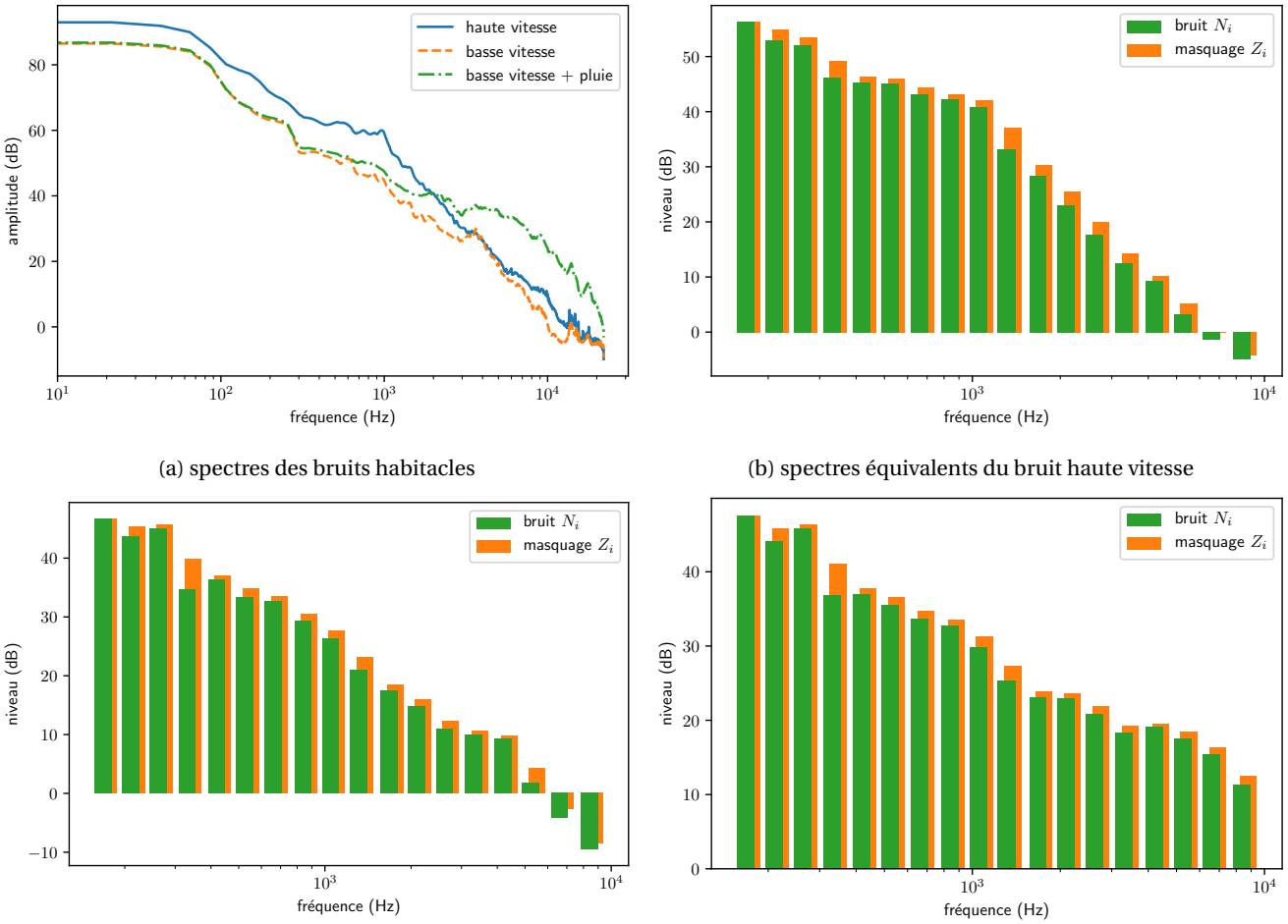
7.1.1 Bruits automobiles

Le bruit présent dans un habitacle automobile vient de trois sources physiques principales : le moteur, le roulement et les frottements aérodynamiques. Le bruit moteur est harmonique et sa fréquence fondamentale est fixée par le régime moteur, c'est la source principale de bruit à basse vitesse. Le bruit de roulement et le bruit aérodynamique sont des bruits large bande dont les niveaux augmentent différemment avec la vitesse. À partir d'environ 50 km/h, le bruit de roulement prend alors une part importante et c'est à partir d'environ 100 km/h que le bruit aérodynamique devient la source majeure de bruit dans l'habitacle. Ces bruits venant du véhicule ont un contenu spectral qui varie peu et qui se déplace doucement en fonction de la vitesse principalement. À ces bruits peuvent aussi s'ajouter des bruits extérieurs comme un bruit ambiant urbain, un phénomène météorologique ou encore des évènements sonores ponctuels à courte durée comme un klaxon ou une sirène passante. Cette courte synthèse des différentes sources de bruit dans un habitacle automobile montre que le bruit présent dans l'habitacle peut être très variable. Cependant, il est aisé d'avoir une bonne estimation des principaux bruits de l'habitacle en temps réel que ce soit par la mesure, via des microphones embarqués, ou une modélisation via les paramètres dynamiques du véhicule.

Nous pouvons observer figure 7.1a des exemples de spectre long-terme de trois bruits habitacles automobiles obtenus en moyennant les deux voies enregistrés en conditions réelles avec une tête acoustique (HMS IV, HEAD acoustics *GmbH*) située à l'emplacement du passager. Afin de balayer des cas d'utilisation assez larges nous avons sélectionné :

- un bruit Grande Vitesse (GV), enregistré à 130km/h, de niveau 92 dB(SPL) /90 dB(Z) /71 dB(A),
- un bruit Basse Vitesse (BV), enregistré à 50km/h, de niveau 85 dB(SPL) / 83 dB(Z) / 60 dB(A),
- un bruit Basse Vitesse avec Pluie (BV+P), enregistré à 50km/h par temps de pluie, de niveau 86 dB(SPL) / 84 dB(Z) / 63 dB(A).

Tout d'abord, nous vérifions que les niveaux du spectre équivalent de masquage ne dépassent jamais les niveaux du spectre équivalent de bruit de plus de 9 dB. Ainsi l'hypothèse 2 est respectée et ce bruit est bien éligible à la procédure d'optimisation exacte du SII. Ensuite, nous observons que les bruits concentrent beaucoup d'énergie dans les basses fréquences, cela explique l'importante différence entre les niveaux en dB(Z) et en dB(A) et cela expliquera aussi que les RSB avec lesquels nous travaillerons dans les sections suivantes sont très faibles. En comparant les spectres BV et GV, on remarque que l'augmentation du bruit aérodynamique, lorsque la vitesse augmente, rehausse le spectre et l'aplatit en atténuant le contraste harmonique alors moins visible à haute qu'à



(a) spectres des bruits habitacles

(b) spectres équivalents du bruit haute vitesse

(c) spectres équivalents du bruit basse vitesse

(d) spectres équivalents du bruit basse vitesse + pluie

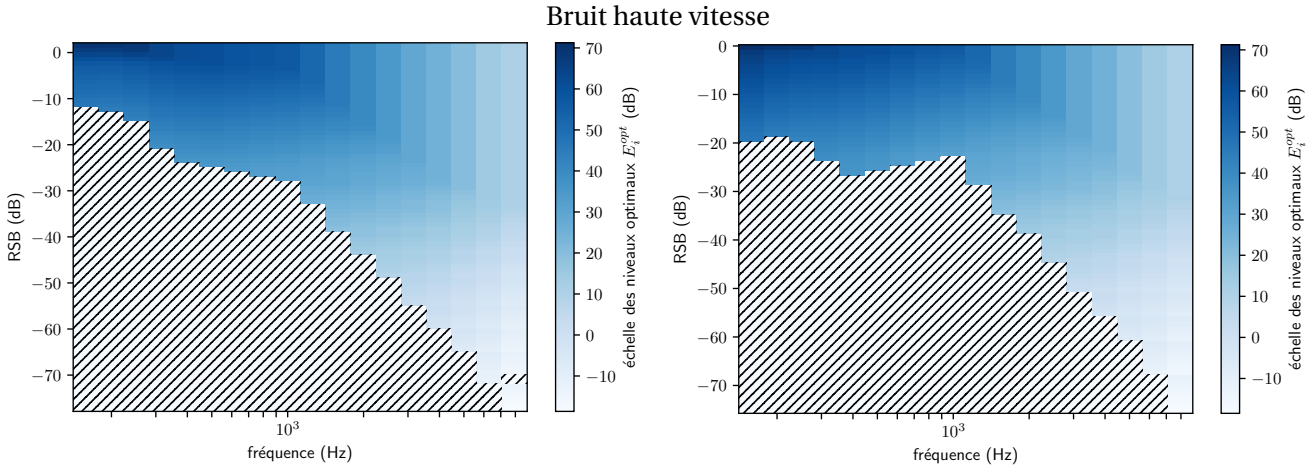
FIGURE 7.1 – Spectres de trois bruits habitacles automobiles et leurs spectres équivalents (bruit et masquage) correspondants.

basse vitesse. Enfin, le spectre BV+P est identique au spectre BV dans les basses fréquences, cependant la présence de pluie ajoute graduellement des composantes hautes fréquences, à partir de 300 Hz avec un maximum autour de 1kHz, ce qui a pour effet d'augmenter la pente spectrale du bruit.

7.1.2 Analyse des résultats objectifs de l'optimisation

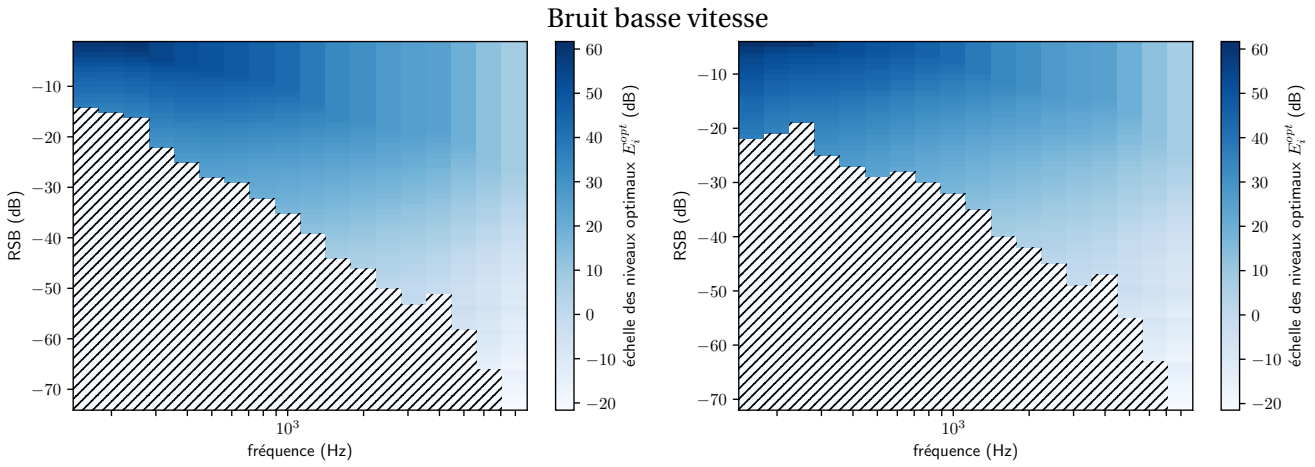
Les niveaux des spectres équivalents (bruit et masquage) utilisés pour le calcul du SII et correspondant au trois bruits introduits précédemment sont visibles figure 7.1. Les spectres équivalents étant calculés sur les bandes de tiers d'octave entre 140 Hz et 8900 Hz, les composantes de fréquences en dehors de cet intervalle ne participent donc pas au calcul, ni à l'optimisation du SII.

En suivant la méthode d'optimisation du SII introduite section 5.2, il est possible de calculer les spectres équivalents de parole optimaux qui maximisent le critère d'intelligibilité dans les trois bruits habitacles donnés. Les spectres optimaux de chaque bruit sont visibles figure 7.2 avec la contrainte énergétique classique en dB(Z) mais aussi la nouvelle contrainte perceptive en dB(A). Nous rappelons que la plage de RSB choisie se situe entre le RSB minimal permettant d'avoir un SII non-nul et le RSB permettant d'avoir un SII maximal. Le comportement de l'optimisation pour les différents bruits suit une logique qui rappellera grandement celle observée pour les bruits classiques section 5.4.1 avec tout de même quelques spécificités.



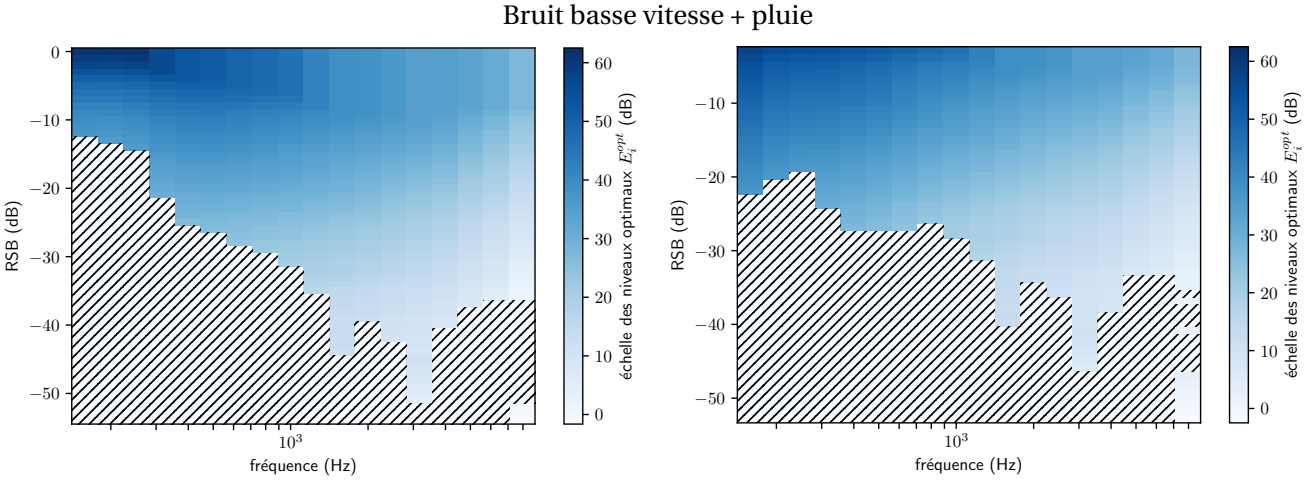
(a) contrainte dB(Z)

(b) contrainte dB(A)



(c) contrainte dB(Z)

(d) contrainte dB(A)



(e) contrainte dB(Z)

(f) contrainte dB(A)

FIGURE 7.2 – Niveaux E_i^{opt} des spectres équivalents optimaux à différents RSB pour chaque bruit automobile et pour chaque contrainte : dB(Z) à gauche et dB(A) à droite.

Spectres optimaux sous contrainte en dB(Z)

Pour l'optimisation avec une contrainte en dB(Z), comme pour le bruit SSN les bruits GV et BV sont spectralement très localisés, ce sont d'abord les bandes où le masquage est peu présent qui sont alors favorisées à faible RSB, puis plus ce dernier augmente, plus la réserve d'énergie disponible augmente et ainsi il devient intéressant d'investir dans des nouvelles bandes où le masquage est graduellement plus présent. Les résultats de l'optimisation dB(Z) pour le bruit BV+P sont plus surprenants, malgré un spectre de masquage avec une pente spectrale aussi décroissante, le remplissage des bandes lorsque le RSB augmente semble plus chaotique. Cela s'explique par son spectre de masquage plus aplati qui minimise l'intérêt d'investir aveuglement toute l'énergie dans les bandes où le masquage est cette fois légèrement moins présent. À cela s'ajoute la pondération plus importante des bandes autour de 2 kHz par la FIB dont l'influence, qui avait déjà été observée pour le bruit rose, est bien plus observable pour ce bruit BV+P que pour les bruits GV et BV.

Spectres optimaux sous contrainte en dB(A)

Pour l'optimisation avec une contrainte en dB(A), les différences observées sont globalement les mêmes que pour le bruit SSN. En effet, la pénalisation des bandes autour de 1 kHz introduite par la contrainte perceptive va influencer la distribution énergétique des spectres optimaux en favorisant les bandes éloignées. Ce phénomène est particulièrement observable vers les RSB autour de -15 dB(RSB) où les bandes basses fréquences sont favorisées par rapport aux résultats des optimisations dB(Z).

Améliorations théoriques du SII

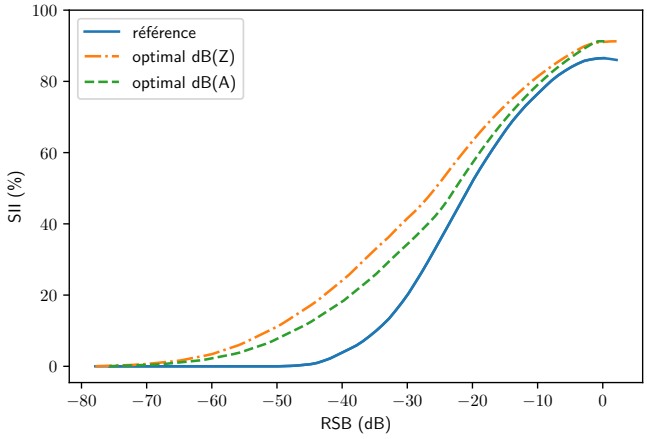
Les SII des spectres optimaux sous contrainte dB(Z) et dB(A), ainsi que ceux du spectre de parole de référence normalisé, sont visibles figure 7.3 pour chaque bruit et pour l'ensemble des RSB. Les améliorations théoriques qui sont alors la différences entre les SII des spectres optimaux et ceux de référence sont visibles figure 7.4. Encore une fois, on observe des comportements très similaires aux résultats obtenus pour le bruit SSN c'est à dire une nette amélioration du SII dans tous les cas avec un pic d'amélioration très important autour du RSB où le SII de référence commence à être non-nul.

On notera toujours une amélioration du SII globalement plus importante pour les optimisations dB(Z) car les bandes à forte pondération ne sont pas pénalisées par la contrainte en dB(A), cependant on remarque une très légère inversion à haut RSB où le SII maximal est toujours atteint d'abord par l'optimisation dB(A). Ce dernier résultat qui peut paraître surprenant s'explique en fait aisément en remarquant que le spectre de masquage présente beaucoup d'énergie dans les trois premières bandes où la FIB est minimale. L'énergie disponible lors de l'optimisation sera donc toujours investie en dernier dans ces bandes très coûteuses afin d'atteindre leur contribution maximale pour l'intelligibilité (zones foncées en haut à gauche des sous-figures 7.2). La contrainte perceptive en dB(A) minimise justement le coût énergétique dans ces bandes où l'oreille est très peu sensible ce qui permet d'atteindre alors légèrement plus rapidement la contribution maximale de ces bandes malgré le retard accumulé vis-à-vis des autres bandes pour les RSB plus faibles.

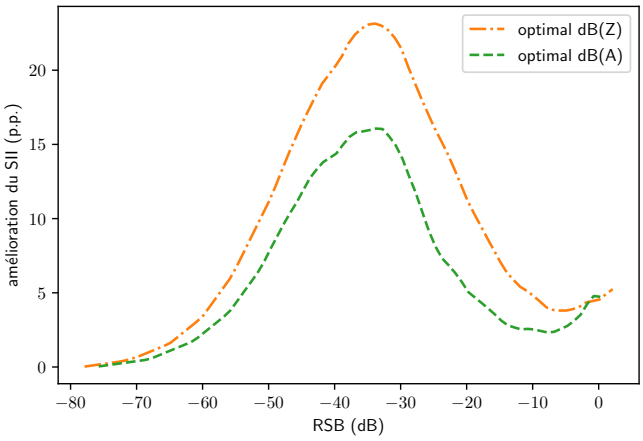
L'amélioration obtenue pour le bruit BV+P est plus constante que pour les autres bruits, elle présente un pic moins important et plus étalé. Pour ce bruit BV+P, l'optimisation en dB(Z) présente toujours une amélioration théorique importante du SII sur toute la plage de RSB mais bien moindre pour l'optimisation dB(A). Comme vu précédemment, le bruit BV+P ayant un spectre moins localisé que les bruits BV et GV, l'optimisation cherche à investir l'énergie disponible dans les bandes où la FIB est importante, l'optimisation dB(Z) y parvient mais l'optimisation dB(A) pénalise cette démarche via la contrainte perceptive minimisant alors l'amélioration possible du SII.

En ce qui concerne les résultats obtenus par approximation du SII, les améliorations théoriques obtenues avec la méthode de STANTON et al., pour les deux initialisations décrites section 5.4.3, sont aussi visibles figure 7.4. On remarque encore une fois la complémentarité des deux initialisations pour ces nouveaux bruits, en prenant le meilleur des deux on approche de près les résultats optimaux sur l'ensemble des RSB.

Bruit haute vitesse

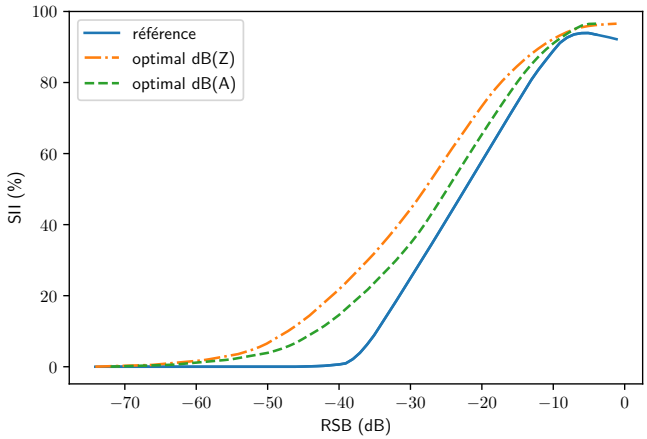


(a) comparaison des SII

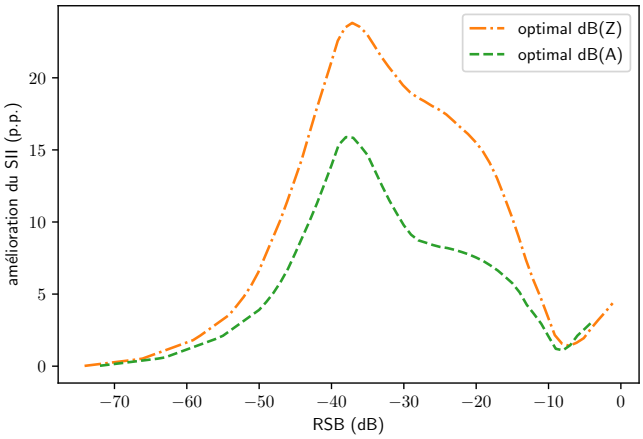


(b) amélioration du SII optimal/référence

Bruit basse vitesse

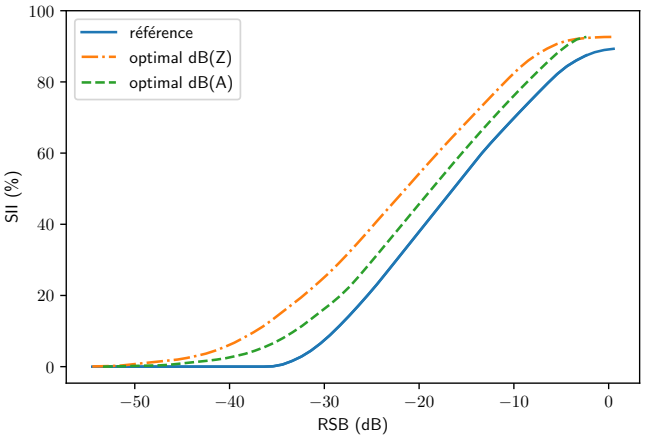


(c) comparaison des SII

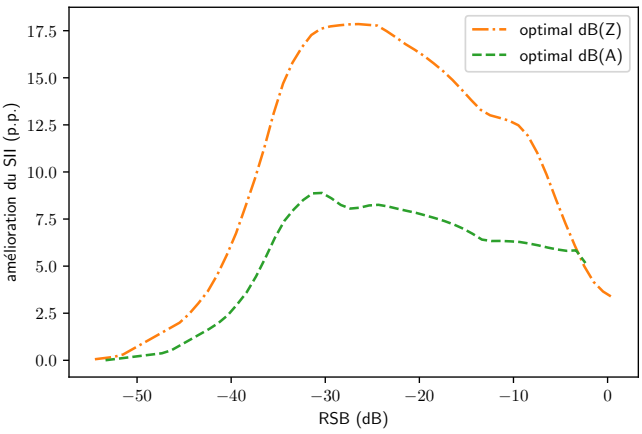


(d) amélioration du SII optimal/référence

Bruit basse vitesse + pluie

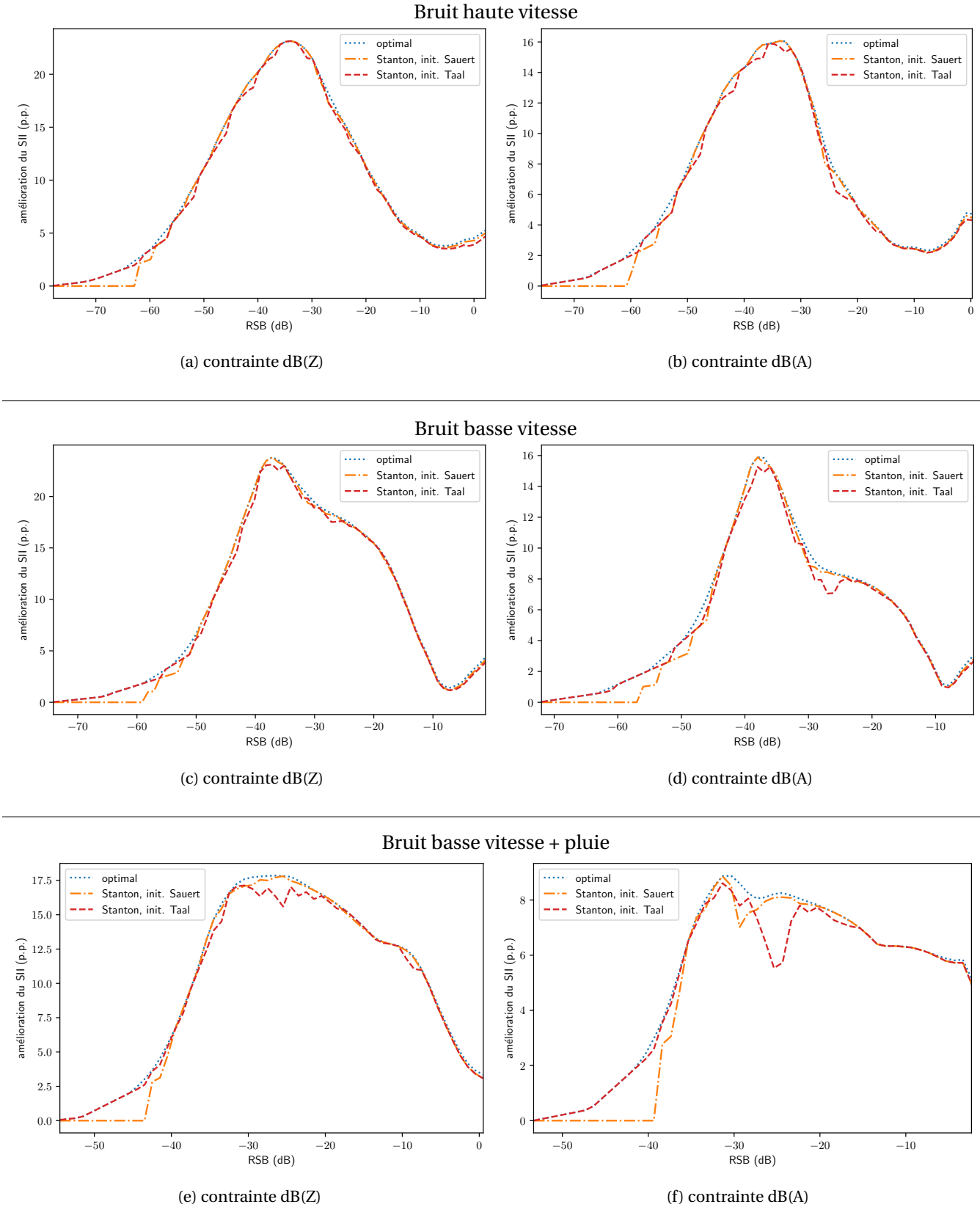


(e) comparaison des SII



(f) amélioration du SII optimal/référence

FIGURE 7.3 – Comparaison, et amélioration, du SII calculé à partir d'un spectre de parole de référence normalisé, d'un spectre optimal de parole de même niveau en dB(Z) et d'un spectre optimal de parole de même niveau en dB(A), pour différents RSB et pour chaque bruit automobile.



(a) contrainte dB(Z)

(b) contrainte dB(A)

(c) contrainte dB(Z)

(d) contrainte dB(A)

(e) contrainte dB(Z)

(f) contrainte dB(A)

FIGURE 7.4 – Amélioration du SII, entre un spectre de parole de référence normalisé et les spectres de parole issus de l'approche de STANTON et al. avec deux initialisations différentes, en fonction du RSB pour chaque bruit automobile et pour chaque contrainte : dB(Z) à gauche et dB(A) à droite. Une initialisation est effectuée avec les spectres de SAUERT et al. (init. Sauert) et l'autre avec les spectres de TAAL et al. (init. Taal).

7.2 Protocole du test subjectif

Dans cette section, nous détaillons les tests subjectifs mis en place pour valider notre approche et nous discutons des résultats obtenus. Nous prenons soins de détailler le choix de chaque variable introduites section 1.3. Les tests sont inspirés, d'une part, du *Hearing In the Noise Test* (HINT) [150] pour ses choix de paramètres et son déroulement, et d'autre part, des travaux de BRAND et al. pour la méthode adaptative d'estimation du SRP.

7.2.1 Stimuli et présentation

Le corpus de phrases choisi suit les recommandations du HINT [150] et est construit avec des phrases au vocabulaire standard représentatives d'un discours classique. Ce sont des phrases tirées du HINT franco-canadien [209], et des phrases de Fournier [58], enregistrées par le Collège National d'Audioprothèse [48]. Le corpus est prononcé par un unique locuteur (homme) de sorte à pouvoir comparer les résultats et chaque mot est considéré comme un mot-clé, même les mots-outils. Deux types de stimuli sont alors utilisés : avec et sans traitement par la méthode de maximisation exacte du SII sous contrainte perceptive.

Les bruits automobiles utilisés sont ceux introduits précédemment : GV, BV et BV+P. Les bruits GV et BV sont tous les deux quasi-stationnaires, cependant la présence de la pluie dans le bruit BV+P ajoute de la non-stationnarité au bruit qui pourrait introduire un biais non contrôlé dans nos tests perceptifs. Afin de prendre en compte l'éventuel effet de la non-stationnarité, nous synthétisons un quatrième bruit stationnaire Basse Vitesse avec Pluie Lissée (BV+PL) de même spectre long terme que le bruit BV+P.

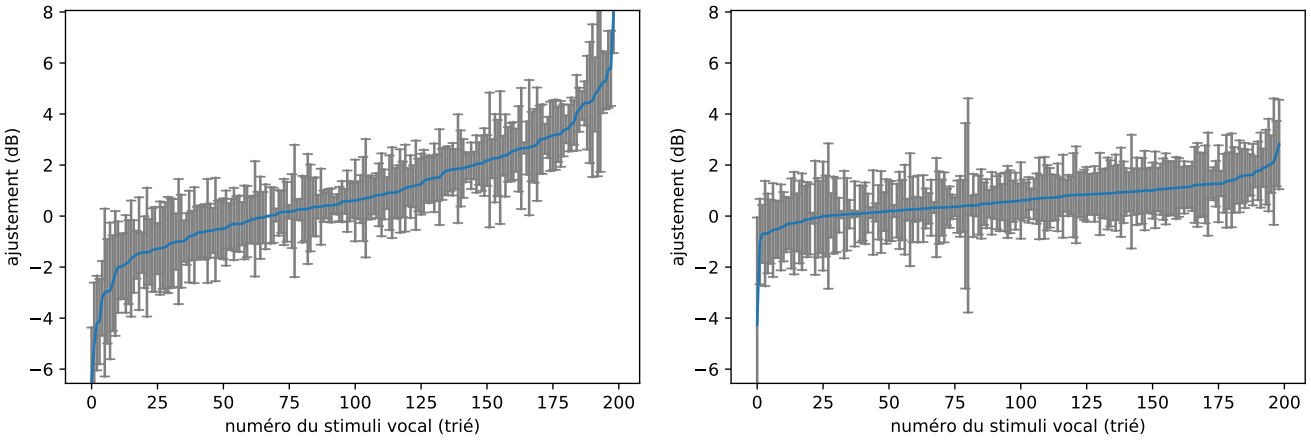
Durant les tests, tous les stimuli sont diffusés au moyen d'un égaliseur programmable (HEAD acoustics *GmbH* PEQ V) et d'un casque audio calibré (Sennheiser^{TM} HD 650).

7.2.2 Équilibrage des phrases

Comme nous l'avons vu section 1.3, une étape cruciale dans la mise en place d'un test d'écoute est l'équilibrage des stimuli vocaux. Effectuer cette étape sur chaque bruit habitacle étudié n'est pas concevable, c'est pourquoi nous avons décidé d'effectuer un équilibrage moyen sur un bruit synthétique dont le spectre correspond à la moyenne, en décibels, des trois bruits. L'équilibrage a été réalisé avec une population de 12 normo-entendants (3 femmes et 9 hommes) âgés entre 22 et 27 ans (pour une moyenne de 25,5 ans) et dont l'acuité auditive a été vérifiée par un examen d'audiométrie tonale [12].

La méthode employée est très proche de celle proposée par Nielsen et al. [148], pour la mise en place du *Hearing In the Noise Test* (HINT) danois [149], car elle permet d'obtenir de très bons résultats avec peu de sujets. La méthode se base sur un jugement subjectif des auditeurs qui doivent indiquer, à chaque présentation d'un stimulus vocal, si il était "facile" ou "difficile" de le comprendre ou bien si c'était "juste" compréhensible. Une session de test avec 12 phrases tirées d'un corpus arbitraire permettait aux auditeurs de s'entraîner à la tâche avant de commencer le véritable équilibrage. Au cours d'une première séquence d'écoute, tous les stimuli vocaux étaient présentés dans un ordre aléatoire au niveau initial de 60 dB(SPL) correspondant à un RSB de -28 dB. Si un stimulus était jugé "juste" il ne serait plus présenté au cours des séquences suivantes. En revanche, si un stimuli était jugé "difficile" (resp. "facile") il serait présenté plus fort (resp. moins fort) lors de la séquence suivante, en commençant par un pas de 2 dB diminuant de moitié à chaque inversion. Les séquences s'enchaînaient alors jusqu'à ce que tous les stimuli aient atteint un niveau de présentation "juste" et possèdent donc chacun un ajustement (en dB) qui lui est propre.

L'équilibrage a été réalisé en deux sessions successives avec deux groupes de six sujets chacun. La première session aura permis de déterminer des premiers ajustements, puis tous les stimuli vocaux présentés lors de la deuxième session commençaient directement avec ces ajustements. Les ajustements moyens du niveau de présentation de chaque stimulus vocal obtenus au cours des deux sessions sont visibles sur la figure 7.5. Les ajustements de la deuxième session sont, en moyenne, plus faibles que ceux de la première session indiquant alors que cette dernière a eu un effet positif sur l'équilibrage des stimuli vocaux. Finalement, les ajustements finaux sont calculés en prenant les moyennes sur l'ensemble des deux sessions.



(a) courbe du niveau de présentation

(b) score obtenu pour chaque phrase

FIGURE 7.5 – Ajustements moyens, et écarts-types, du niveau de présentation de chaque stimulus vocal obtenus au cours de la première session (gauche) et de la deuxième session (droite).

À l'issue de cette phase d'équilibrage nous avons créé dix listes de vingt phrases. Deux listes sont composées uniquement des phrases aux ajustements maximums (supérieurs à 2,5 dB), elles serviront de listes d'entraînement pour le test d'intelligibilité. Les huit autres listes sont composées de manière à conserver un maximum de phrases des listes originales ensemble car ces dernières sont équilibrées phonétiquement. Les listes originales qui ont cédé le plus de stimuli aux listes d'entraînement servent alors à remplir les autres listes et la répartition s'effectue de manière à minimiser la variance des ajustements entre les listes. Les nouvelles listes, ainsi que l'ajustement de chaque stimulus vocal, sont consultables dans l'annexe A.

7.2.3 Méthode d'estimation du seuil de réception de la parole

Les procédures adaptatives d'estimation du SRP ont été introduites 1.3.3 et celle qui a été choisie est celle détaillée dans les travaux de BRAND et al. [28] qui est une généralisation de la méthode adaptative de HAGERMAN et al. [73]. Elle consiste à présenter une première phrase à un niveau très bas relativement au SRP supposé et d'augmenter progressivement le niveau jusqu'à ce que le sujet soit capable de répéter au moins un mot. Les 19 phrases suivantes ne sont présentées qu'une seule fois et leur niveau dépend de la réponse donnée par le sujet pour la phrase précédente. Le pas en dB qui conditionne le changement de présentation d'une phrase sur l'autre est donné par l'équation suivante :

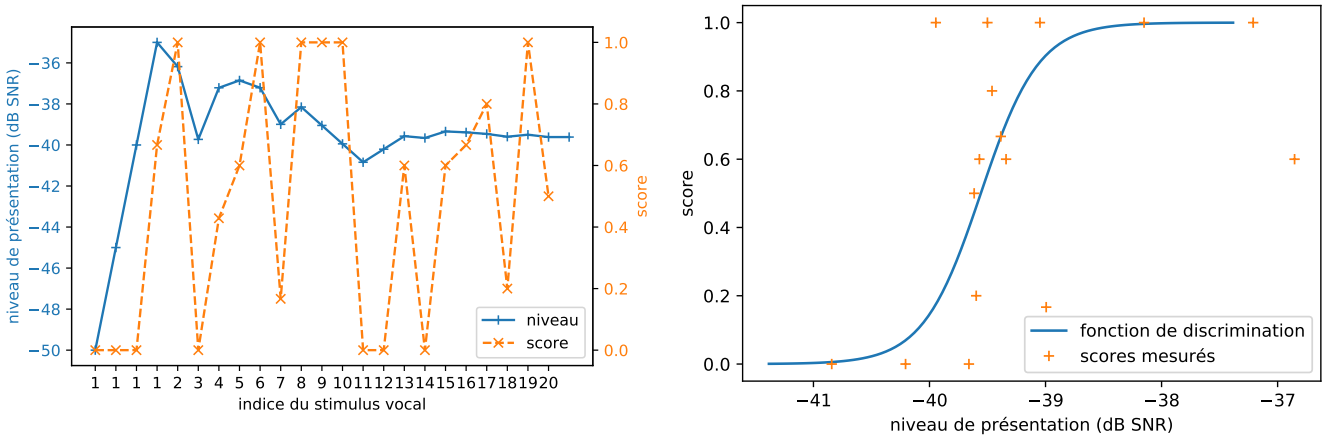
$$\Delta \mathrm{L} = \frac{10}{1,41^{r}} \cdot (0,5 - \mathit{prec}) \,. \tag{7.1}$$

L'indice r correspond au nombre d'inversions du niveau de présentation ayant eu lieu durant la procédure i.e. le nombre de changements de signe du ΔL. La variable *prec* correspond au score de la phrase précédente i.e. le taux de mots correctement répétés. Un exemple d'évolutions communes des scores et du niveau de présentation des stimuli vocaux pour une condition arbitraire du test est visible figure 7.6a.

La répétition d'un mot dont la phrase est présentée à un niveau L est considérée comme une épreuve de Bernoulli indépendante de probabilité p décrite par l'équation suivante :

$$p(\mathrm{L}, \mathrm{L}_{50}, s_{50}) = \frac{1}{1 + \exp(4 \cdot s_{50} \cdot (\mathrm{L}_{50} - \mathrm{L}))} \,. \tag{7.2}$$

Cette probabilité, aussi appelée fonction psychométrique, correspond à une fonction logistique centrée en L_{50} (qui correspond au SRP) et dont la dérivée au point d'inflexion est notée s_{50}. Le niveau de présentation ainsi que la réussite pour la répétition de chaque mot sont sauvegardés durant la présentation de la liste. À la fin de celle-ci, les paramètres L_{50} et s_{50} sont alors estimés en utilisant un estimateur du maximum de vraisemblance sur



(a) évolutions communes des scores et du niveau de présentation

(b) scores moyens et fonction psychométrique estimée

FIGURE 7.6 – Estimation du SRP pour une condition arbitraire du test avec (a) les évolutions communes des scores et du niveau de présentation des stimuli vocaux et (b) la fonction psychométrique estimée à partir des scores mesurés.

le processus de Bernoulli résultant. Pour une condition donnée, si m mots ont été présentés, la vraisemblance d'une fonction psychométrique donnée $p(\mathrm{L}, \mathrm{L}_{50}, s_{50})$ est alors :

$$l(p(\mathrm{L}, \mathrm{L}_{50}, s_{50})) = \prod_{k=1}^{m} p(\mathrm{L}_k, \mathrm{L}_{50}, s_{50})^{c(k)} \cdot (1 - p(\mathrm{L}_k, \mathrm{L}_{50}, s_{50}))^{c(k)-1}, \tag{7.3}$$

avec $c(k) = 1$ si le $k^{\text{ème}}$ mot présenté au niveau L_k a été correctement répété et sinon $c(k) = 0$. Un exemple de fonction psychométrique estimée est proposé figure 7.6b, on notera tout de même que ce sont bien les scores binaires aux mots qui ont été utilisés dans la maximisation de la vraisemblance, non pas les scores moyens de la phrases qui ont été ajoutés sur la figure à des fins d'interprétation des résultats. En effet, nous voyons bien sur cette figure que la courbe psychométrique modélise assez bien l'intelligibilité mesurée en minimisant l'influence des données aberrantes qui sont inéluctablement nombreuses lors de tests d'intelligibilité.

Une méthode plus classique consistant à moyenner les derniers niveaux de présentation ne prend pas directement en compte le score d'intelligibilité des derniers stimuli et se base seulement sur une hypothèse de convergence vers le SRP. L'équation 7.1 montre effectivement qu'il y aura convergence grâce à la diminution du pas à chaque changement de signe. Il arrive cependant qu'une diminution trop rapide du pas, alors que le niveau est encore éloigné du SRP réel, empêche de se rapprocher de ce dernier. Dans le cas de l'estimation du SRP en passant par la modélisation de l'intelligibilité par une fonction psychométrique, les scores de chaque mot sont pris en compte dans le calcul de la vraisemblance même si le niveau ne converge pas vers le SRP réel, la rendant alors bien plus robuste. On notera tout de même que cette approche profite aussi de la convergence vers le SRP réel car il est important d'avoir beaucoup de données au niveau du point d'inflexion de la courbe logistique pour obtenir une estimation plus précise des paramètres.

7.2.4 Configuration

Les tests ont été réalisés sur une population de 13 normo-entendants (2 femmes et 11 hommes) âgés entre 20 et 31 ans (pour une moyenne de 24,7 ans) et dont l'acuité auditive a été vérifiée par un examen d'audiométrie tonale [12]. Les tests se déroulent alors de la façon suivante. La procédure d'estimation du SRP est appliquée sur les deux listes d'entraînement afin d'habituer le sujet à la tâche. On applique ensuite la procédure aux huit autres listes, deux listes pour chaque bruit en conservant la voix originale pour l'une et en la traitant pour l'autre. Il est important que les listes, les phrases et le traitement soient présentés de façon pseudo-aléatoire en utilisant un carré latin équilibré afin d'éviter l'effet d'ordre.

7.3 Résultats du test subjectif et analyse de la méthode

7.3.1 Présentation des résultats

Les résultats des tests d'intelligibilité ont été synthétisés dans le tableau 7.1 et par des diagrammes en boîte de Tukey sur la figure 7.7. On remarque une nette diminution du SRP dans toutes les situations. Cependant, les performances varient entre les différents bruits : le SRP a une diminution moyenne de 6,9 dB pour le bruit GV et seulement 3,9 dB pour le bruit BV. Pour les bruits BV+P et BV+PL, la diminution est encore plus faible avec, respectivement, 1,7 dB et 1,1 dB.

Pour vérifier la significativité des résultats, une analyse de la variance (ANOVA), à deux facteurs sur mesures répétées a été effectuée. Les facteurs contrôlés sont le bruit utilisé (4 valeurs possibles) et la présence de traitement (2 valeurs possibles). Avant toute chose, nous fixons le niveau de significativité à 0,01 car nous souhaitons observer seulement les différences très significatives. Les résultats montre que le bruit [$F(3,36) = 742, p < 10^{-3}$], le traitement [$F(1,12) = 422, p < 10^{-3}$] et l'interaction entre ces deux facteurs [$F(3,36) = 42, p < 10^{-3}$] influencent significativement le SRP. Des comparaisons multiples par paire ont été effectuées sur les interactions, en utilisant des tests-t appariés avec la correction de Bonferroni, afin de détailler où les différences ont lieu et on retiendra trois observations intéressantes :

- l'amélioration du SRP grâce au traitement est statistiquement significative dans les bruits GV ($p < 10^{-4}$) et BV ($p < 10^{-4}$) mais pas dans les bruits BV+P ($p = 0,022$) et BV+PL ($p = 0,329$).
- le SRP n'est pas significativement différent entre les bruits GV et BV pour la parole non-traitée ($p = 0,46$) alors qu'il l'est pour la parole traitée ($10^{-4} < p < 10^{-3}$). Ainsi, les traitements sont significativement plus efficaces dans le bruit GV que dans le bruit BV.
- le SRP n'est pas significativement différent entre les bruits BV+P et BV+PL pour la parole non-traitée ($p = 1,0$) et il ne l'est pas non plus pour la parole traitée ($p = 0,54$). Ainsi, la non-stationnarité du bruit BV+P ne semble pas avoir d'influence, ni sur l'intelligibilité de la parole naturel, ni sur les performances des traitements. Les interprétations des résultats se feront donc uniquement sur le bruit BV+PL.

Bruit	Traitement	$\overline{\text{SRP}}$ (dB)	σ(SRP) (dB)
Grande Vitesse (GV)	non	-34,0	1,03
	oui	-40,9	1,8
Basse Vitesse (BV)	non	-33,6	0,72
	oui	-37,5	0,81
Basse Vitesse avec Pluie (BV+P)	non	-26,4	0,79
	oui	-28,1	0,74
Basse Vitesse avec Pluie Lissée (BV+PL)	non	-26,4	0,94
	oui	-27,5	1,34

TABLEAU 7.1 – Synthèse des résultats des tests d'intelligibilité avec les SRP moyens estimés et leurs écarts-types pour chaque condition (Bruit x Traitement).

7.3.2 Interprétation et analyse de la méthode

En premier lieu, nous pouvons noter que les SRP estimés pour toutes les conditions sans traitement ont un écart-type faible, inférieur à 1 dB, ce qui confirme la robustesse de la méthode adaptative d'estimation choisie. Au contraire, les écarts-type SRP estimés pour toutes les conditions avec traitement sont très variables montrant alors les traitements ont engendré des réactions assez diversifiées de la part des sujets.

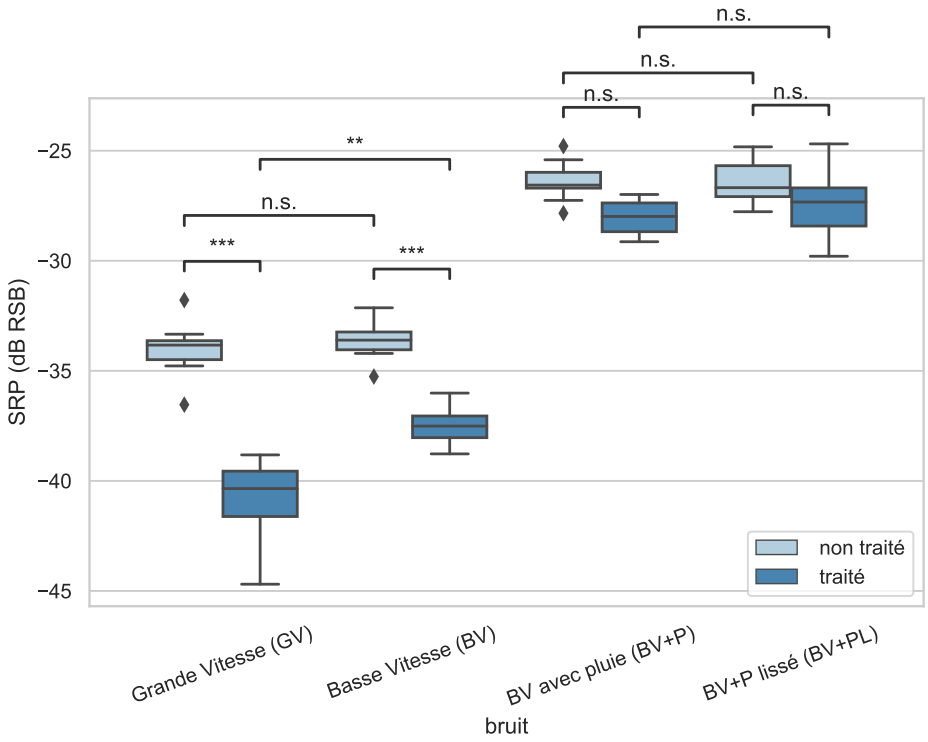


FIGURE 7.7 – Profil statistiques des SRP estimés pour les signaux non traités et traités dans les bruits automobiles testés. La significativité des différences entre certaines conditions sont indiqués par : (***) si $p < 10^{-4}$, (**) si $p < 10^{-3}$, (*) si $p < 10^{-2}$ et n.s. si $p \geq 10^{-2}$.

Concernant le bruit GV, l'énergie étant ré-allouée vers les hautes fréquences où le bruit est beaucoup moins présent, nous favorisons donc l'émergence du signal au détriment des composantes basses fréquences qui, de toute façon, seraient masquées par le bruit trop important dans cette zone. Cette procédure permet d'améliorer grandement l'intelligibilité tout en conservant un niveau perçu constant lorsque le bruit est localisé dans une partie du spectre, ici les basses fréquences. Mais si le spectre est plus aplati, l'amélioration est moins visible (bruit BV), voire n'est plus significative (bruit BV+PL).

Ces résultats confirment bien les améliorations théoriques de l'intelligibilité visibles sur la figure 7.4. En effet, pour les bruits GV et BV, l'amélioration théorique est maximale pour des RSB entre −30 dB et −40 dB, exactement dans la zone où le SRP a été estimé. Cependant, ces niveaux ne sont clairement pas représentatifs d'une écoute réelle car ils sont bien trop bas pour une écoute confortable et demandent beaucoup de concentration pour comprendre les phrases. Pour des RSB plus grands que −30 dB, la marge d'amélioration d'intelligibilité théorique diminue lorsque le RSB augmente. L'amélioration du SRP dans ces zones serait probablement bien plus faible, voire plus du tout significative. Bien que l'estimation du SRP soit un outil très utilisé dans les tests subjectifs, son amélioration à faible RSB ne prouve pas son amélioration à des niveaux raisonnables, et plus particulièrement pour des approches comme la maximisation d'un critère d'intelligibilité qui ont un traitement très sensible au RSB. Pour le bruit BV+PL, l'amélioration théorique était déjà beaucoup plus mitigée et l'amélioration non significative de l'intelligibilité confirme ces résultats. N'ayant pas de bandes où le bruit est peu présent dans lesquels ré-allouer l'énergie, un traitement visant à maximiser le SII sous contrainte en dB(Z) aurait tout de même concentré l'énergie du signal dans les bandes autour de 2 kHz où la FIB est importante et aurait très probablement obtenu une amélioration significative. Cependant la contrainte perceptive empêche cette astuce visant à exploiter la sensibilité auditive en augmentant le niveau perçu du signal et le traitement proposé n'est donc plus du tout efficace.

Conclusion du chapitre 7

Dans ce chapitre, nous avons testé une approche d'amélioration de l'intelligibilité de la parole dans le bruit basée sur la maximisation exacte du critère d'intelligibilité SII avec une contrainte énergétique prenant en compte la sensibilité auditive. Nous avons montré à travers des tests subjectifs rigoureux que cette approche améliore significativement l'intelligibilité vis-à-vis du SRP dans des bruits automobiles basse fréquence. Par contre, dans des bruits au spectre plus étalé, l'amélioration n'est plus significative. Cela met donc en évidence la limite de cette approche d'optimisation d'un critère d'intelligibilité lorsqu'elle est couplée avec la nouvelle contrainte perceptive.

Concernant la résolution du problème d'optimisation, le fait d'utiliser l'extension des procédures par approximation, proposée section 5.4.3, aurait permis d'obtenir des résultats équivalents. Bien que ce ne soit pas une preuve suffisante pour valider le comportement quasi-optimal de cette extension dans d'autres types de bruit, elle présente jusqu'ici des performances assez stables dans tout type de bruit avec trois bruits classiques (chapitre 5), un bruit de conversation (chapitre 6), ou encore trois bruits relativement différents d'habitacles automobiles dans ce chapitre.

Ce chapitre clôture donc nos travaux sur le renforcement direct de la parole basés sur la maximisation exacte du SII et son étude sous une contrainte perceptive nouvelle. La maximisation exacte du critère a obtenu des performances très intéressantes avec des tests d'intelligibilité classiques lors de notre participation au challenge et ces performances sont maintenues avec la contrainte perceptive dans des bruits habitacles automobiles au spectre localisé. Cependant, les interprétations des résultats et des limites de l'approche laissent place à de nombreuses perspectives d'approfondissement citées tout au long de cette partie comme une validation de la pertinence de la contrainte perceptive proposée ou une adaptation du critère proposant des traitements plus doux et potentiellement plus performants.

Troisième partie

Renforcement paramétrique par conversion du style de parole dans le bruit avec amélioration du traitement des aspects temporels

Chapitre 8

Approches actuelles de renforcement paramétrique de la parole dans le bruit et négligence des aspects temporels

Sommaire

Retour à la table des matières

Introduction du chapitre 8

Le renforcement paramétrique de la parole dans le bruit se base sur la manipulation de paramètres de la parole afin d'améliorer son intelligibilité dans des environnements bruyants. Les paramètres à modifier peuvent être issus d'outils d'analyse-modifications-synthèse de la parole ou directement accessibles dans le cas d'une synthèse vocale paramétrique. Une étude approfondie des différents modèles d'analyse-modifications-synthèse de la parole a été faite chapitre 2. Les détails de fonctionnement des synthèses vocales paramétriques sortent de notre cadre d'étude, en revanche la manipulation des paramètres de la parole s'effectue de façon similaire aux modèles d'analyse-modifications-synthèse.

STYLIANOU propose une appellation pour la manipulation paramétrique générale de la parole, il parle alors de "transformation de la parole" et deux grands domaines d'application en émergent [186]. Celui de la "conversion de parole" qui cherche à transformer les paramètres d'une parole source vers une parole cible pour, par exemple, de la conversion d'identité ou de la conversion d'émotion. Et celui de la "modification de parole" qui manipule les paramètres sans chercher à imiter de parole cible. Ces deux domaines sont utilisés en renforcement paramétrique de la parole. En renforcement par modification de la parole les transformations se basent majoritairement sur des observations et connaissances sur l'intelligibilité de la parole. En renforcement par conversion de parole on cherchera plutôt à imiter des styles de parole visant l'amélioration de l'intelligibilité comme la parole Lombard ou la parole claire, introduites chapitre 3, on parle alors de conversion du style de parole.

Dans tous les cas, les méthodes de renforcement paramétriques de la parole peinent à obtenir d'aussi bonnes performances que les méthodes directes. Les dégradations liées à la synthèse est la principale raison évoquée justifiant ce résultat. On notera d'ailleurs que nous étions les seuls à proposer une méthode paramétrique lors du deuxième challenge *Hurricane*, présenté chapitre 6, et que nos piètres performances sont certainement liées au dégradation très audibles de la synthèse. Mais nous pensons aussi que la négligence de nombreux aspects temporels, parfois évoqués dans les études actuelles, pourraient aussi contribuer à ce manque à gagner.

Dans ce chapitre nous présentons alors en détails les différentes approches de renforcement paramétrique de la parole dans le bruit avec un regard critique sur la négligence actuelle de certains aspects temporels. Nous étudierons d'abord, dans la section 8.1, les méthodes de modification de parole basées sur des règles fixées à l'avance issues de connaissances ou d'observations psychoacoustiques de l'intelligibilité de la parole. Les méthodes de renforcement par conversion du style de parole seront ensuite étudiées dans la section 8.2.

8.1 Renforcement par modification de la parole

Les traitements appliqués en renforcement par modification de la parole dans le bruit sont généralement largement inspirées de ceux observés en renforcement naturel de la parole comme la parole Lombard et la parole claire détaillées chapitre 3. La majorité des approches combinent plusieurs modifications afin de maximiser le gain d'intelligibilité rendant alors la contribution de chacune difficile à interpréter.

8.1.1 Manipulations du fondamental

Comme nous l'avons vu chapitre 3, une observation importante de l'effet Lombard est l'augmentation quasi-systématique de la fréquence moyenne du fondamental. Les raisons de cette augmentation peuvent être une conséquence physiologique de l'augmentation de l'intensité globale ou une volonté de concentrer le contenu spectral dans les médiums où l'oreille est la plus sensible. Une autre modification naturelle introduite par l'effet Lombard est la modulation du fondamental (vibrato) favorisant la séparation entre la parole et le bruit. De nombreux travaux se sont donc inspirés de ces observations en proposant des modifications du fondamental afin d'améliorer l'intelligibilité des signaux de parole dans le bruit. Dans cette section, nous détaillerons les principales approches de renforcement paramétrique de la parole dans le bruit basées sur la manipulation du fondamental et leurs résultats.

Augmentation de la valeur moyenne

L'approche la plus directe consiste à augmenter la valeur moyenne du fondamental comme observé dans la parole Lombard. Cependant, comme nous l'avons vu chapitre 3, l'augmentation moyenne du fondamental dans la parole Lombard naturelle ne semble pas engendrer de gain d'intelligibilité dans un bruit stationnaire large bande de type SSN. Cette constatation est appuyée par les résultats de VALENTINI-BOTINHAO et al. obtenus en augmentant et en diminuant la valeur moyenne du fondamental [211]. Aucune des deux approches n'améliore, ou ne dégrade, l'intelligibilité de manière significative dans de nombreuses situations différentes : bruit SSN, bruit de conversation (avec plusieurs voix concurrentes), bruit habitacle automobile et bruit haute fréquence. PATEL et al. trouvent tout de même une amélioration significative de l'intelligibilité en augmentant de 20 Hz la valeur moyenne du fondamental dans un bruit de conversation (avec plusieurs voix concurrentes) [151], mais la présence d'autres transformations spectro-temporelles ne permet pas de quantifier l'influence de l'augmentation du fondamental.

Manipulations de la trajectoire

Si des modifications moyennes de la fréquence fondamentale ne semble pas procurer de gain d'intelligibilité, MILLER et al. proposent une étude sur les effets de différentes manipulations de la trajectoire du fondamental sur l'intelligibilité de la parole [124]. Les manipulations proposées sont :

- un aplatissement du fondamental, en fixant une trajectoire constante à la valeur médiane du fondamental du signal,
- une exagération, en amplifiant les variations du fondamental d'un facteur 1,75 autour de la valeur médiane,
- une inversion, en prenant l'inverse de la trajectoire normalisé par le carré de la valeur médiane,
- deux modulations du fondamental, en ajoutant à la trajectoire une composante sinusoïdale à 2,5 Hz et 5 Hz d'amplitude égale à l'écart-type de la fréquence du fondamental du locuteur.

Les modifications sont effectuées par une approche type TD-PSOLA, via le logiciel *Praat* [185]. Les résultats des tests subjectifs sur des normo-entendants dans un bruit SSN ne montrent pas d'amélioration de l'intelligibilité qui, au contraire, s'en retrouve dégradée significativement : l'aplatissement ou l'exagération du fondamental provoquent une diminution de 13 p.p. de l'intelligibilité et l'inversion ou les modulations provoquent une diminution de 23 p.p.. Les conclusions de l'étude n'affirment pas que de telles manipulations de la trajectoire du fondamental ne puissent pas améliorer l'intelligibilité mais plutôt que :

Remarque 1 *de telles manipulations de la trajectoire du fondamental pourraient améliorer l'intelligibilité mais des manipulations synthétiques trop primaires ont tendance à introduire des incohérences intonatives qui seraient responsables de la baisse d'intelligibilité.*

Adaptation locale

Enfin, VILLEGAS et al. proposent un entre-deux en modifiant la valeur moyenne locale du fondamental visant à maximiser une mesure, notée GP, de la proportion spectro-temporelle du signal qui échappe au masquage énergétique. Cette approche, intitulée F_0*-shift*, a plutôt tendance à réduire la fréquence du fondamental car cela entraîne une concentration des harmoniques dans les zones spectro-temporelles d'intérêt. Des tests perceptifs ultérieurs montrent finalement que cette approche dégrade l'intelligibilité dans un bruit de conversation (avec une seule voix concurrente) et un bruit SSN [45].

8.1.2 Manipulation des formants

D'autres observations de l'effet Lombard et dont l'implémentation est très appréciée en renforcement de la parole sont deux manipulations spécifiques de l'enveloppe spectrale à savoir l'aplatissement de la pente spectrale et les manipulations formantiques. Ces types de transformation se font déjà indirectement en renforcement

direct de la parole par l'utilisation de filtres passe-haut de pré-accentuation introduits dans le chapitre 4. En effet, un filtrage passe-haut suivi d'une normalisation de l'énergie entraîne un aplatissement de la pente spectrale et une augmentation relative à partir des seconds formants connus pour leur importance vis-à-vis de l'intelligibilité de la parole. Les approches paramétriques n'apportant rien de plus pour l'aplatissement de la pente spectrale, ce sera surtout pour leur capacité à manipuler les formants qu'elle seront utilisées.

McLoughlin et al. proposaient déjà en 1997 d'utiliser le codage LPC, et plus particulièrement la manipulation des LSP, pour modifier les formants [123]. Les modifications proposées sont un décalage des formants vers les hautes fréquences par un facteur linéairement dégressif, passant de 50% à 0% sur la plage 0-4kHz, et une augmentation d'un facteur 2,4 de la largeur des formants. Pour le décalage, le schéma de modification proposé est visible figure 8.1. Des tests d'intelligibilité portant sur l'écoute de voyelles dans un bruit habitacle automobile attestent d'une augmentation de 15 p.p. pour le décalage et de 21 p.p. pour l'élargissement des formants.

Nathwani et al. proposent aussi d'utiliser le codage LPC, sans passer par les LSP cette fois, afin de déplacer les formants [144]. Plusieurs schémas de décalage sont étudiés à savoir une modification en rampe, deux modifications lisses très similaires et une modification lisse adaptative au bruit qui adapte son amplitude à l'énergie instantanée de celui-ci. Le schéma de la modification en rampe (MR) et celui d'une modification lisse non-adaptative (ML) sont visibles figure 8.1. Des tests subjectifs portant sur l'écoute de phrases dans un bruit habitacle automobile (quasi-stationnaire) attestent d'un gain d'intelligibilité significatif pour les modifications lisses, supérieur à 1 dB pour 48,3% des sujets, mais pas de gain significatif pour la modification en rampe. Les schémas non-adaptatifs au bruit semblent plus performants mais la présence d'artefacts notables impacte la qualité des signaux traités. Ce problème sera corrigé ultérieurement en ne décalant les formants que sur les segments voisés et en lissant les modifications [145], le gain d'intelligibilité résultant conservera le même ordre de grandeur.

Une des caractéristiques de la parole claire, introduite chapitre 3, est la sur-articulation. Alors que l'effet Lombard se caractérise par un décalage de l'espace vocalique, la sur-articulation s'interprète comme un élargissement de l'espace vocalique. Godoy et al. proposent alors [69] de modifier les formants afin d'introduire un élargissement semblable à celui observé dans la parole claire. Le schéma proposé a été élaboré à partir des modifications de formants observées dans des exemples de parole claire et est visible figure 8.1. Contrairement aux autres travaux, on voit bien que les formants sont maintenant déplacés de façon bilatérale et on peut se représenter les deux bosses dans la fonction $\Delta(f)$, au niveau des premiers et deuxièmes formants, responsables de l'élargissement de l'espace vocalique. Le déplacement des formants s'effectue par une déformation directe des fréquences de l'enveloppe spectrale estimée par *True Envelope* [166] sur des fenêtres de 30ms. Sur chaque fenêtre l'emplacement des formants est estimé en localisant les maxima de l'enveloppe spectrale, le déplacement associé à chaque formant est donné par $\Delta(f)$ puis ceux associés aux fréquences entre les formants sont obtenues par interpolation linéaire. En appliquant sur chaque fenêtre la déformation de l'échelle spectrale obtenue puis en re-synthétisant le signal par addition-recouvrement, des analyses montrent bien un élargissement de l'espace vocalique. Des tests subjectifs dans un bruit SSN présenté à 0 et -4 dB RSB ne relève pourtant pas de gain d'intelligibilité significatif suite aux modifications. Une explication proposée par les auteurs de l'étude est que :

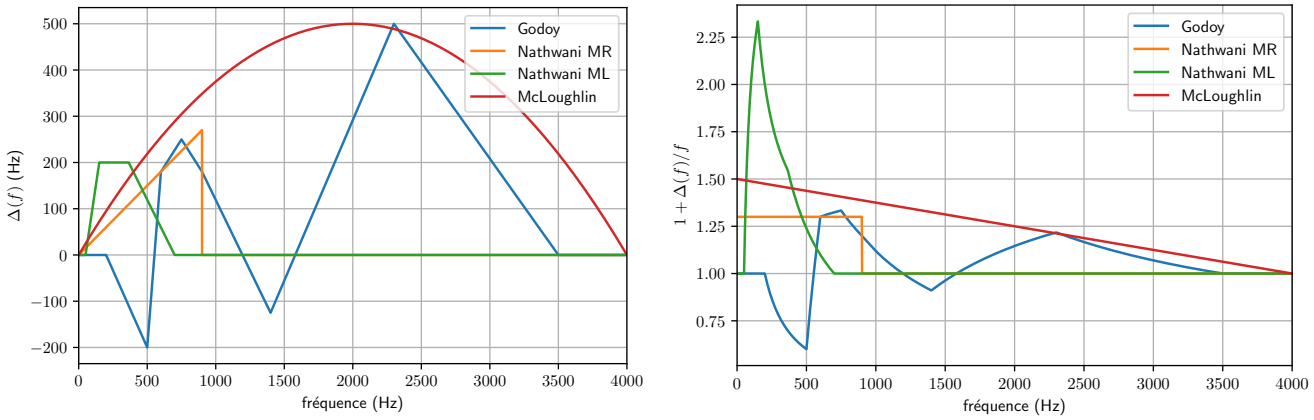
Remarque 2 *le traitement ne se concentre pas assez sur le contexte temporel et les spécificités des phonèmes, ainsi un traitement moyen ne permet pas de capturer les subtilités de l'hyper-articulation.*

8.1.3 Manipulation du débit

Dans le renforcement naturel de la parole, on observe aussi d'importantes modifications du débit, ainsi de nombreuses tentatives de manipuler synthétiquement le débit de parole afin d'améliorer l'intelligibilité des signaux dans le bruit ont été proposées.

Modifications uniformes

L'approche la plus simple à mettre en oeuvre est une modification uniforme du débit de parole. Adams et al. proposent d'étudier l'influence d'une accélération, ou un ralentissement, uniformes du débit de parole sur l'intelligibilité de la parole dans un bruit de conversation (avec plusieurs voix concurrentes) [4]. Les modifications



(a) fonction $\Delta(f)$ représentant le décalage des formants

(b) facteur d'échelle associé à $\Delta(f)$

FIGURE 8.1 – Schémas de modification des formants proposés par MCLOUGHLIN et al. [123], GODOY et al. [69] et NATHWANI et al. [144]

sont effectuées par une approche type TD-PSOLA, via le logiciel *Cool Edit Pro* et consistent à fixer les signaux à 130, 170 et 234 mots par minute (MPM). Des tests subjectifs dans du bruit de conversation (avec une seule voix concurrente) montrent une augmentation significative de l'intelligibilité liée à la diminution du débit. Cependant, n'ayant pas d'informations sur le débit original des signaux, il est difficile de savoir si c'est l'accélération des signaux qui cause une réduction de l'intelligibilité ou si c'est leur ralentissement qui cause une augmentation de celle-ci. En supposant que la condition 170 MPM corresponde au débit original, un ralentissement de 30% engendre une légère amélioration significative de l'intelligibilité d'environ 1 dB.

Modifications non-uniforme

Une autre approche consiste à appliquer une modification du débit non uniforme basée sur les caractéristiques des segments de parole. Les conditions et les valeurs sont choisies de façon empiriques en s'inspirant des résultats d'études de différents styles de parole. NATHWANI et al. proposent de rallonger les segments voisés de 40% et les segments non-voisés de 20% [145]. Les modifications sont effectuées par une approche type TD-PSOLA. Des tests d'intelligibilité portant sur l'écoute de phrases dans un bruit habitacle automobile montrent que les modifications ne procurent pas d'amélioration significative de l'intelligibilité.

Synthèse des modifications uniforme ou non dans des bruits stationnaires et fluctuants

COOKE et al. proposent une étude complète analysant les effets de certaines modifications du débit de parole sur l'intelligibilité dans des bruits de différentes natures [41]. Les modifications sont implémentées avec l'approche temporelle WSOLA. La première modification proposée est une élongation uniforme des signaux de 30%. La deuxième modification est non-uniforme et consiste à calculer une trajectoire de modification temporelle entre le signal de parole et celui du bruit favorisant l'audibilité des segments de parole contenant de l'information, cette méthode est surnommée *GCRetime* [10]. Trois bruits sont considérés : un bruit stationnaire SSN, et deux bruits fluctuants à savoir un autre bruit large bande (de type SMN) et un bruit de conversation (avec une seule voix concurrente).

Les tests d'intelligibilité montrent que la modification uniforme tend vers un léger gain d'intelligibilité de 3 p.p. non significatif dans le bruit SSN alors que dans les bruits fluctuants le gain est significatif et bien plus important : 8,3 p.p. dans le bruit de conversation et 9,0 p.p. dans le bruit SMN. Pour la méthode *GCRetime*, les différences sont encore plus marquées avec une baisse très significative de l'intelligibilité de -14,9 p.p. dans le bruit SSN, un gain significatif de 10,3 p.p. dans le bruit SMN et un gain significatif encore plus important de 16,3 p.p. dans le bruit de conversation. Ces résultats confirment les études précédentes en appuyant l'inefficacité

des modifications synthétiques du débit de parole actuelles dans des bruits stationnaires. Ils montrent aussi qu'au contraire, en exploitant les fluctuations des bruits non-stationnaires, des modifications du débit telles que proposées par *GCRetime* permettent une nette amélioration de l'intelligibilité de la parole dans ces conditions.

L'influence du bruit SMN et du bruit de conversation sur la méthode *GCRetime* a aussi été étudié en mesurant l'intelligibilité des signaux dans chaque bruit en leur appliquant le traitement destiné à l'autre bruit. Pas de changement significatif n'a été relevé dans le bruit SMN alors qu'une baisse d'intelligibilité significative de 5,6 p.p. a été relevée dans le bruit de conversation. Cette différence s'explique probablement par la capacité de *GCRetime* d'exploiter pleinement la structure spectro-temporelle parcimonieuse du bruit de conversation et perd de son efficacité face à la structure spectrale étalée du bruit SMN.

8.2 Renforcement par conversion du style de parole

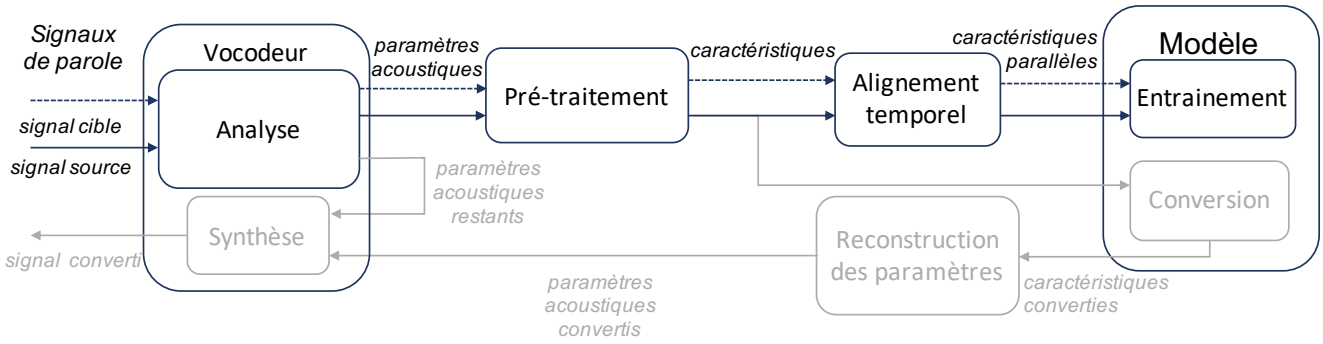
Les approches de renforcement par modification de la parole introduites précédemment se basent majoritairement sur des observations de signaux de paroles produites dans des environnements bruyants. Une autre façon de procéder consiste à exploiter directement des bases de données de ces signaux afin de chercher à convertir le style de parole traitée en celui de la parole visée e.g. convertir une parole neutre en parole Lombard ou en parole claire. Le principe de la conversion du style, qui est un sous-domaine de la conversion de parole, est détaillé dans cette section. Nous verrons alors quels sont les fonctions de conversion qui sont principalement utilisées et les caractéristiques acoustiques qui sont manipulées.

8.2.1 Principe de la conversion de parole

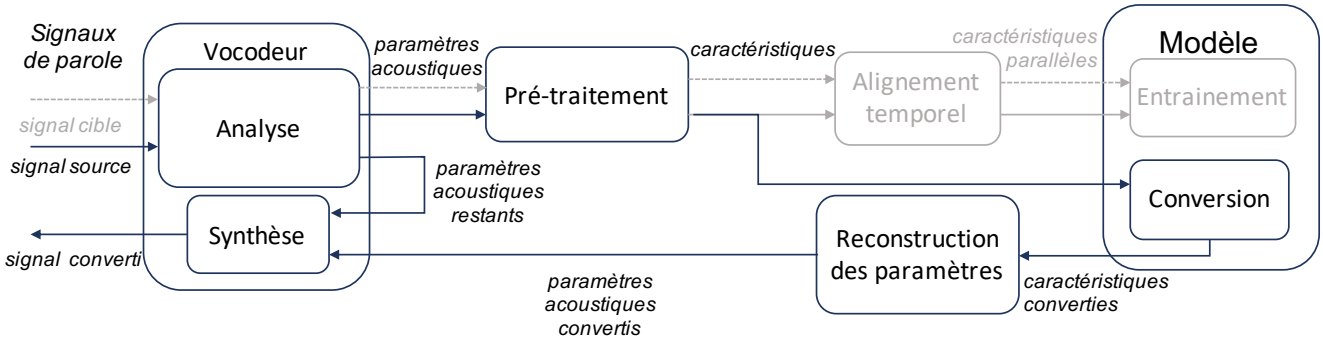
La conversion de parole consiste à construire une fonction de conversion transformant certaines caractéristiques acoustiques d'une parole source afin d'approcher celles d'une parole cible. Pour y parvenir, des bases de données comportant des signaux de la parole source et cible sont exploitées afin d'entraîner la fonction à convertir les caractéristiques. Généralement, les données des deux paroles ont le même contenu syntaxique ce qui permet d'aligner temporellement les signaux et d'appliquer ce qu'on appelle un apprentissage parallèle. De nombreuses approches parallèles ont été proposées en conversion de parole pour le type de fonction de conversion à utiliser que ce soit par quantification vectorielle [1], par régression linéaire multivariée [210], par modèle de mélange gaussien [187] ou par réseau de neurones artificiels [143, 189]. Ces approches portaient majoritairement sur de la conversion d'identité qui est l'application première du domaine, en conversion du style les travaux se sont concentrés sur les approches les plus performantes d'abord par modèle de mélange gaussien [95, 199, 5, 125, 114] puis par réseau de neurones à propagation avant [175, 126]. Ce sont donc ces fonctions de conversion que nous détaillerons dans cette section. Notons qu'il existe aussi des approches non-parallèles de conversion de parole [140, 176] qui ne seront pas abordées ici.

Un diagramme décrivant le flux d'information dans un système de conversion de parole classique est visible figure 8.2. Durant une phase d'entraînement, mise en évidence figure 8.2a, pour chaque signal de parole source et celui de parole cible correspondant, le vocodeur extrait des paramètres acoustiques. Les paramètres sont éventuellement pré-traités puis alignés temporellement donnant alors les caractéristiques acoustiques parallèles à associer sur lesquelles la fonction de conversion est entraînée. Une fois l'entraînement terminé, le système peut être utilisé pour convertir de nouveaux signaux de la parole source vers la parole cible. Au cours de cette phase de conversion, mise en évidence figure 8.2b, seules les caractéristiques acoustiques de la parole source sont transmises à la fonction de conversion qui estime des caractéristiques converties. Une étape de reconstruction des paramètres utilisés par le vocodeur peut être nécessaire en fonction du pré-traitement qui a été appliqué. Enfin, des modifications temporelles, non représentées sur le diagramme, peuvent être introduites durant la synthèse du signal converti.

Soit un ensemble de données parallèles composé de M paires de signaux source/cible. Après avoir extrait les caractéristiques acoustiques pour chaque paire de signaux alignés (x_m, y_m), on note $\mathbf{X}_m \in \mathbb{R}^{\mathrm{N}_m \times \mathrm{C}_x}$ le vecteur de C_x caractéristiques du signal source, et $\mathbf{Y}_m \in \mathbb{R}^{\mathrm{N}_m \times \mathrm{C}_y}$ le vecteur de C_y caractéristiques du signal cible associé. Si les



(a) phase d'entraînement



(b) phase de conversion

FIGURE 8.2 – Diagramme classique d'un système de conversion de parole durant les phases d'entraînement et de conversion.

caractéristiques sources et cibles sont identiques alors $C_x = C_y$, et c'est systématiquement le cas dans un système de conversion de parole classique. Cependant, ce n'est pas une nécessité comme nous le verrons, dans le chapitre suivant, avec notre proposition d'intégrer les modifications temporelles dans les caractéristiques apprises.

8.2.2 Conversion de parole par modèle de mélange gaussien

La procédure de conversion de parole par modèle de mélange gaussien, usuellement abrégé par l'acronyme anglais pour *Gaussian Mixture Model* (GMM), base son apprentissage sur une concaténation des vecteurs de caractéristiques des M signaux de la base de données d'entraînement à savoir :

$$\boldsymbol{\mathcal{X}} = [\mathbf{X}_1^\top, \mathbf{X}_2^\top, ..., \mathbf{X}_M^\top]^\top \in \mathbb{R}^{N \times C_x}, \tag{8.1}$$

avec $N = \sum_{m=1}^{M} N_m$, et :

$$\boldsymbol{\mathcal{Y}} = [\mathbf{Y}_1^\top, \mathbf{Y}_2^\top, ..., \mathbf{Y}_M^\top]^\top \in \mathbb{R}^{N \times C_y}. \tag{8.2}$$

Elle consiste alors à modéliser la densité de probabilité jointe de ces deux vecteurs par une somme pondérée de lois normales multivariées $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ [87] :

$$P(\boldsymbol{\mathcal{Z}} \mid \boldsymbol{\theta}) = \sum_{g=1}^{G} w_g \mathcal{N}(\boldsymbol{\mathcal{Z}}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \tag{8.3}$$

avec $\boldsymbol{\mathcal{Z}} = [\boldsymbol{\mathcal{X}}^\top, \boldsymbol{\mathcal{Y}}^\top]^\top$ le vecteur joint et $\boldsymbol{\theta} = \{w_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \mid g \in [\![1, G]\!]\}$ le jeu de paramètre du GMM.

Apprentissage et conversion

Les paramètres du GMM sont estimés par l'algorithme Espérance-Maximisation (EM) lors de la phase d'entraînement. En décomposant par blocs les paramètres estimés de la g$^{\text{ème}}$ composante du mélange, on note :

$$\boldsymbol{\mu}_g = \begin{bmatrix} \boldsymbol{\mu}_g^{\mathrm{X}} \\ \boldsymbol{\mu}_g^{\mathrm{Y}} \end{bmatrix} \quad \text{et} \quad \boldsymbol{\Sigma}_g = \begin{bmatrix} \boldsymbol{\Sigma}_g^{\mathrm{XX}} & \boldsymbol{\Sigma}_g^{\mathrm{XY}} \\ \boldsymbol{\Sigma}_g^{\mathrm{YX}} & \boldsymbol{\Sigma}_g^{\mathrm{YY}} \end{bmatrix}. \tag{8.4}$$

Une fois le modèle entraîné, il peut être utilisé pour convertir des échantillons de caractéristiques acoustiques $\hat{\mathbf{X}}(n) \in \mathbb{R}^{1\times \mathrm{C}_x}$ d'un nouveau signal source en caractéristiques $\hat{\mathbf{Y}}(n) \in \mathbb{R}^{1\times \mathrm{C}_y}$ qui devraient approcher celles de la parole cible. Pour cela, l'estimation ces caractéristiques s'effectue par la méthode des moindres carrés et on obtient :

$$\hat{\mathbf{Y}}(n) = \sum_{g=1}^{\mathrm{G}} \mathrm{P}(g \mid \mathbf{X}(n), \boldsymbol{\theta}) \left(\boldsymbol{\mu}_g^{\mathrm{Y}} + \boldsymbol{\Sigma}_g^{\mathrm{YX}} (\boldsymbol{\Sigma}_g^{\mathrm{XX}})^{-1} (\mathbf{X}(n) - \boldsymbol{\mu}_g^{\mathrm{X}}) \right) \tag{8.5}$$

avec :

$$\mathrm{P}(g \mid \mathbf{X}(n), \boldsymbol{\theta}) = \frac{w_g \mathcal{N}(\mathbf{X}(n); \boldsymbol{\mu}_g^{\mathrm{X}}, \boldsymbol{\Sigma}_g^{\mathrm{XX}})}{\sum_{h=1}^{\mathrm{G}} w_h \mathcal{N}(\mathbf{X}(n); \boldsymbol{\mu}_h^{\mathrm{X}}, \boldsymbol{\Sigma}_h^{\mathrm{XX}})} \tag{8.6}$$

Remarques complémentaires

Il existe deux problèmes majeurs avec cette méthode classique de conversion de parole par GMM :

1. bien que les trajectoires des caractéristiques acoustiques estimées soient globalement proche de celles visées, la conversion ne prend pas en compte la corrélation entre les fenêtres ce qui a tendance à introduire localement des motifs anormaux,
2. la conversion a tendance a placer les caractéristiques acoustiques sur des valeurs proches des moyennes des composantes du mélange ce qui a pour conséquence de réduire leur variances globales et ainsi d'aplatir les trajectoires converties.

Pour traiter ces deux problèmes, TODA et al. proposent (1) d'incorporer les caractéristiques dynamiques *delta* détaillées ci-après, dans le modèle accompagné d'un algorithme qui estime les paramètres par maximum de vraisemblance et (2) d'introduire les variances globales des caractéristiques directement dans la fonction de vraisemblance pour qu'elles soient prises en compte dans l'optimisation. Cet algorithme s'appelle *Maximum Likelihood Parameter Generation* (MLPG) [204] et permet d'améliorer grandement la qualité et les performances de la conversion de parole par GMM.

Pour toute caractéristique de dimension L $\mathbf{x} \in \mathbb{R}^{\mathrm{N}_m \times \mathrm{L}}$, ses caractéristiques *delta* notées $\mathbf{x}^{\delta} \in \mathbb{R}^{\mathrm{N}_m \times \mathrm{L}}$ se calculent de la façon suivante :

$$\forall (n,l) \in [[1,\mathrm{N}]] \times [[1,\mathrm{L}]], x^{\delta}(n,l) = \frac{\sum_{d=1}^{\mathrm{D}} (x(n+d,l) - x(n-d,l)) \cdot d}{2 \cdot \sum_{d=1}^{\mathrm{D}} d^2}, \tag{8.7}$$

où typiquement D = 2. Il est aussi possible de calculer ses caractéristiques *delta-delta* notées $\mathbf{x}^{\delta^2} \in \mathbb{R}^{\mathrm{N}_m \times \mathrm{L}}$ directement par $\mathbf{x}^{\delta^2} = (\mathbf{x}^{\delta})^{\delta}$. Un exemple de caractéristiques *delta* et *delta-delta* calculées à partir de l'enveloppe énergétique d'un signal de parole est présenté figure 8.3.

Notons que les performances de l'approche par GMM dépendent grandement du choix du nombre M de composantes dans le mélange. Le *Variational Bayesian GMM* (VBGMM) est une variante basée sur des méthodes variationnelles qui sont une extension de l'algorithme EM intégrant une régularisation des paramètres à partir de distributions a priori. Concernant les poids, les VBGMM ont donc une tendance naturelle à fixer le poids de certaines composantes à zéro permettant alors d'obtenir un nombre de composantes adapté.

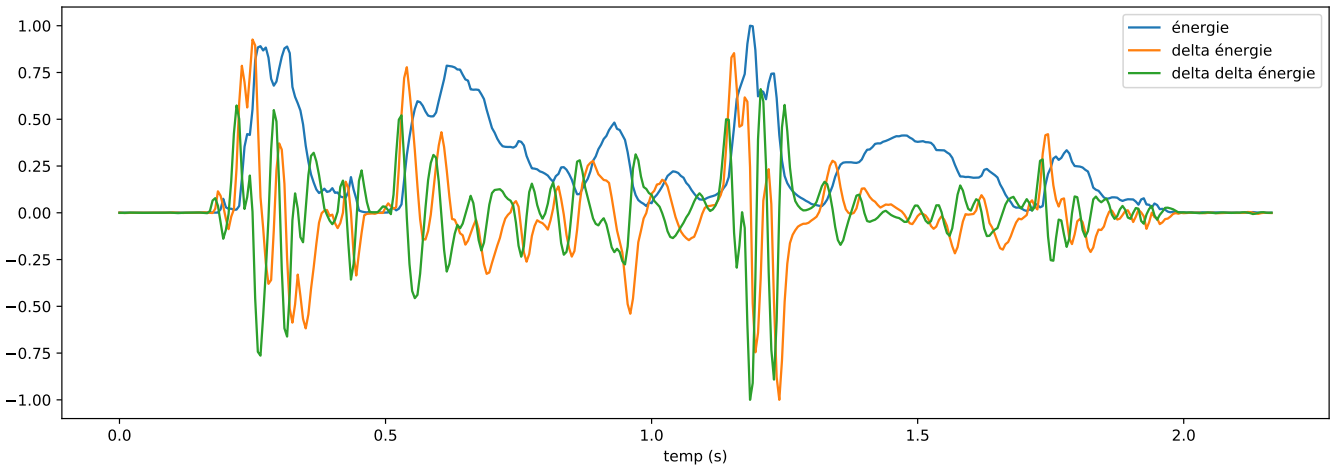


FIGURE 8.3 – Exemple de caractéristiques *delta* et *delta-delta* calculées à partir de l'enveloppe énergétique d'un signal de parole, avec D=2.

8.2.3 Conversion de parole par réseau de neurones à propagation avant

Un réseau de neurones artificiels, usuellement abrégé par l'acronyme anglais pour *Artificial Neural Network* (ANN), consiste en de multiples neurones inter-connectés par des liaisons auxquelles sont associés des poids. Ces poids sont modulés durant l'apprentissage par une minimisation de l'erreur quadratique moyenne sur les données d'entraînement. Le modèle de régression non-linéaire obtenu peut alors servir de fonction de conversion en conversion de parole. Le réseau de neurones à propagation avant, usuellement abrégé par l'acronyme anglais pour FeedForward Neural Network (FFNN), est la topologie d'ANN majoritairement utilisée pour sa simplicité d'utilisation.

Apprentissage et conversion

Pour un FFNN de L couches, la propagation des vecteurs d'activation dans le réseau se calcule par la formule itérative suivante :

$$\forall l \in [[1, \mathrm{L}-1]], \quad \mathbf{a}_l = \sigma(\mathbf{W}_l \mathbf{a}_{l-1}), \tag{8.8}$$

avec $\mathbf{a}_l$ les activations sur la $l^{\text{ème}}$ couche du réseau, $\mathbf{W}_l$ les poids associés initialisés aléatoirement et σ la fonction d'activation des neurones. On remarque que $\mathbf{a}_0$ correspond à l'entrée du réseau et ainsi :

$$\mathbf{a}_0 = [\mathbf{X}_1^\top, \mathbf{X}_2^\top, ..., \mathbf{X}_\mathrm{M}^\top]^\top \in \mathbb{R}^{\mathrm{N} \times \mathrm{C}_x}, \tag{8.9}$$

avec $\mathrm{N} = \sum_{m=1}^{\mathrm{M}} \mathrm{N}_m$. La taille de la première couche est donc nécessairement C_x. De plus, $\mathbf{a}_\mathrm{L}$ correspond à la sortie du réseau, on note alors $\hat{\boldsymbol{\mathcal{Y}}} = \mathbf{a}_\mathrm{L}$ et la fonction de coût à minimiser est l'erreur quadratique obtenue qui s'exprime par :

$$\mathrm{J} = \|\boldsymbol{\mathcal{Y}} - \hat{\boldsymbol{\mathcal{Y}}}\|^2 = \|\boldsymbol{\mathcal{Y}} - \mathbf{a}_\mathrm{L}\|^2, \tag{8.10}$$

avec :

$$\boldsymbol{\mathcal{Y}} = [\mathbf{Y}_1^\top, \mathbf{Y}_2^\top, ..., \mathbf{Y}_\mathrm{M}^\top]^\top \in \mathbb{R}^{\mathrm{N} \times \mathrm{C}_y}. \tag{8.11}$$

La taille de la dernière couche est donc nécessairement C_y. Une descente de gradient est alors appliquée sur les poids afin de minimiser la fonction de coût, pour cela on utilise une astuce mathématique de calcul efficace par rétropropagation du gradient. Les poids sont alors simplement mis à jour par la formule suivante :

$$\forall l \in [[1, \mathrm{L}]], \quad \mathbf{W}_l = \mathbf{W}_l - \alpha \frac{\partial \mathrm{J}}{\partial \mathbf{W}_l}, \tag{8.12}$$

avec α le taux d'apprentissage conditionnant la vitesse d'actualisation des poids.

Une fois le modèle entraîné, il peut être utilisé pour convertir des échantillons de caractéristiques acoustiques $\hat{\mathbf{X}}(n) \in \mathbb{R}^{1\times \mathrm{C}_x}$ d'un nouveau signal source en caractéristiques $\hat{\mathbf{Y}}(n) \in \mathbb{R}^{1\times \mathrm{C}_y}$ qui devraient approcher celles de la parole cible. Pour cela, l'estimation ces caractéristiques s'effectue directement en propageant l'échantillon dans le réseau c'est à dire en fixant :

$$\mathbf{a}_0 = \hat{\mathbf{X}}(n) \in \mathbb{R}^{1\times \mathrm{C}_x}, \tag{8.13}$$

puis en appliquant l'équation itérative 8.8, on obtient alors :

$$\hat{\mathbf{Y}}(n) = \mathbf{a}_\mathrm{L} \in \mathbb{R}^{1\times \mathrm{C}_y}. \tag{8.14}$$

Remarques complémentaires

En pratique, et contrairement au modèle GMM, toutes les données ne servent pas à l'actualisation des poids en même temps, elles sont divisées en paquets repartis aléatoirement appelés mini-lots. L'optimisation s'effectue donc avec une descente de gradient stochastique dont les deux principaux intérêts sont une réduction de la puissance de calcul instantanée nécessaire et une tendance à échapper aux minima locaux. Soit B le nombre de mini-lots utilisés, la nouvelle fonction de coût s'exprime alors :

$$\mathrm{J} = \sum_{b=1}^{\mathrm{B}} \|\boldsymbol{\mathcal{Y}}_b - \hat{\boldsymbol{\mathcal{Y}}}_b\|^2. \tag{8.15}$$

Il est aussi commun de désactiver aléatoirement certains neurones à chaque étape pour éviter une dépendance trop forte entre les neurones qui risquerait d'introduire du sur-apprentissage. À chaque passage d'un mini-lot, l'activation d'un pourcentage de neurones de chaque couche est annulée, ce pourcentage s'appelle le taux d'abandon. Il est plus commode de fixer le même taux pour toutes les couches mais il est tout a fait possible d'allouer des taux spécifiques à chaque couche. De plus, l'affaiblissement des pondérations est une technique de régularisation qui consiste à ajouter une pénalité à la fonction de coût qui dépend de l'amplitude des poids, celle-ci s'exprime alors :

$$\mathrm{J} = \sum_{b=1}^{\mathrm{B}} \|\boldsymbol{\mathcal{Y}}_b - \hat{\boldsymbol{\mathcal{Y}}}_b\|^2 + \frac{\lambda}{2}\sum_{l=1}^{\mathrm{L}} \|\mathbf{W}_l\|^2, \tag{8.16}$$

avec λ le taux d'affaiblissement des poids. Finalement, l'actualisation des poids s'effectue maintenant par la formule suivante :

$$\forall\, l \in [\![1,\mathrm{L}]\!], \quad \mathbf{W}_l = \mathbf{W}_l - \alpha\left(\lambda\mathbf{W}_l + \sum_{b=1}^{\mathrm{B}} \frac{\|\boldsymbol{\mathcal{Y}}_b - \hat{\boldsymbol{\mathcal{Y}}}_b\|^2}{\partial \mathbf{W}_l}\right). \tag{8.17}$$

Lors de l'utilisation d'un ANN, il est aussi courant d'appliquer une standardisation de chaque caractéristique (moyenne nulle et variance unitaire). En effet, l'optimisation par descente de gradient, de même que les fonctions d'activation, sont généralement sensibles à l'échelle des données. Une standardisation permet donc d'équilibrer l'importance de chaque caractéristique et favorise la convergence de l'apprentissage. Les GMM par leur nature même ne sont pas soumis à cette contrainte, les ensembles recherchés n'ont pas nécessairement la même variance, contrairement aux K-moyennes, et n'ont pas besoin d'être centrés.

Enfin, le problème (1) du GMM introduit dans le section précédente, portant sur la création de motifs anormaux dans les trajectoires des caractéristiques converties à cause de l'hypothèse d'indépendance entre les échantillons, est aussi présent dans le FFNN. On utilisera alors généralement les caractéristiques *delta* et *delta-delta* afin d'introduire un aspect dynamique dans l'apprentissage.

8.2.4 Caractéristiques acoustiques exploitées

Dans cette sous-section, nous détaillons les caractéristiques acoustiques utilisées en conversion de parole et plus particulièrement en conversion du style. Nous voyons aussi ce que les approches de renforcement par conversion du style de parole peuvent apporter de plus aux traitements des caractéristiques acoustiques que les approches par modification de parole.

Enveloppe spectrale

Les travaux de conversion de parole se concentrent majoritairement sur la conversion de l'enveloppe spectrale, les caractéristiques extraites à ce but peuvent être de plusieurs natures :

- Les LSP, utilisées comme caractéristiques à convertir elles permettent une représentation simplifiée de l'enveloppe spectrale. Elles sont appréciées en conversion d'identité pour leur grande intercorrélation [86, 106].
- Les coefficients cepstraux sur l'échelle de Mel, usuellement abrégé par l'acronyme anglais pour Mel Frequency Cepstral Coefficients (MFCC), permettent aussi une représentation simplifiée de l'enveloppe spectrale. Ils sont plutôt préférés en conversion du style pour leur capacité à capturer la forme globale de l'enveloppe spectrale qui influence fortement la perception d'un style [95, 49, 189, 114, 175, 176].
- L'enveloppe spectrale, elle est parfois utilisée directement comme caractéristique à convertir [210], cependant la dimension associée est souvent très grande et cela force à contraindre la fonction de transformation priorisant ainsi la qualité de re-synthèse. Des approches hybrides proposent une conversion directe de l'enveloppe spectrale combinée à une conversion d'une représentation simplifiée de celle-ci et permettent alors un compromis entre qualité et similarité [53]. Enfin, avec le développement des ANN et leur capacité à traiter des données avec beaucoup de dimensions, l'utilisation directe de l'enveloppe spectrale s'envisage de plus en plus [125].

Ainsi, en renforcement par conversion du style de parole, la fonction de conversion va apprendre à transformer l'enveloppe spectrale de manière beaucoup plus spécifique que les approches par modification de parole vues section 8.1.2. Cela pourrait répondre partiellement à la remarque 2 qui suggère que la dégradation de l'intelligibilité introduite par la manipulation des formants pourrait venir d'un traitement qui ne se concentre pas assez sur les spécificités des phonèmes. En revanche, concernant le contexte temporel des phonèmes, les modèles utilisés actuellement, à savoir le GMM et le FFNN, peuvent prendre en compte sommairement l'environnement local avec l'utilisation des caractéristiques delta, mais omettent le contexte à des échelles plus importantes comme la place du phonème dans le mot, ou même dans la phrase. C'est pourquoi dans le chapitre suivant nous proposerons l'utilisation d'une topologie différente d'ANN qu'est le réseau de neurones récurrents, usuellement abrégé par l'acronyme anglais pour *Recurrent Neural Network* (RNN), prenant en compte la dépendance entre les échantillons à l'échelle du signal entier.

Fondamental

En conversion d'identité, on se contente généralement d'une simple transformation linéaire du fondamental. Cela consiste à apposer la moyenne et variance du fondamental du locuteur cible sur le locuteur source, sur une échelle logarithmique [49, 189]. Cependant, la trajectoire du fondamental étant la caractéristique principale permettant de percevoir une émotion [142], le traitement de cette caractéristique est différent en conversion d'émotion et en conversion du style. Ainsi, les valeurs logarithmiques du fondamental sont généralement ajoutées dans les caractéristiques fournies à la fonction de conversion [106, 114, 175, 176].

Ainsi, en renforcement par conversion du style de parole, la fonction de conversion devrait apprendre à transformer la trajectoire du fondamental de manière beaucoup plus naturelle que les approches par modification de parole vues section 8.1.1. Cela pourrait répondre partiellement à la remarque 1 qui suggère que la dégradation de l'intelligibilité introduite par la manipulation de la trajectoire du fondamental pourrait venir des manipulations synthétiques trop primaires qui auraient tendance à introduire des incohérences intonatives. En revanche, encore une fois, les modèles utilisés actuellement, à savoir le GMM et le FFNN, peuvent prendre en compte sommairement l'environnement local avec l'utilisation des caractéristiques *delta*, mais omettent le contexte temporel à des échelles plus importantes comme l'évolution de la trajectoire du fondamental à l'échelle du mot, ou même de la phrase.

Par conséquent, l'utilisation d'un RNN qui sera proposée dans le chapitre suivant devrait améliorer cette prise en compte du contexte temporel. De plus, pour aider le modèle à capturer les tendances de la trajectoire du fondamental à différentes échelles, nous proposerons aussi une représentation différente du fondamental en lui appliquant une transformée en ondelettes, usuellement abrégé par l'acronyme anglais pour *Continuous Wavelet Transform* (CWT), et en exploitant les coefficients résultants.

Énergie instantanée

En conversion de parole, l'enveloppe énergétique est majoritairement conservée ou convertie directement avec la conversion de l'enveloppe spectrale comme une variable cachée. En effet, les approches qui, par exemple, manipule l'enveloppe spectrale en passant par les MFCC conservent généralement le premier coefficient [95, 49, 189, 125, 114, 175, 176] qui est une image de l'énergie instantanée.

En revanche, les critiques qui ont été faites sur le traitement du fondamental peuvent aussi s'appliquer ici. C'est pourquoi, dans le chapitre suivant, nous proposerons d'extraire cette caractéristique afin de lui appliquer aussi une transformée en ondelette afin d'aider le modèle à capturer les tendances de la trajectoire énergétique à différentes échelles.

Débit de parole

En conversion de parole, le débit de parole est soit conservé [49, 106, 125, 189, 126, 114], soit converti par une des approches de modification de parole introduites section 8.1.3 e.g. par une simple transformation linéaire de la durée en fonction du voisement [175, 176].

En effet, les modifications du débit de parole ne s'intègrent pas aussi naturellement dans les fonctions de conversion que les autres caractéristiques qui, elles, sont parallèles. Cependant, comme nous l'avons indiqué au début de cette section, rien ne nous empêche d'avoir plus de caractéristiques de sortie que de caractéristiques d'entrée. Dans le chapitre suivant, nous proposons alors l'idée nouvelle de créer une caractéristique de modification du débit directement intégrable en sortie de la fonction de conversion. L'idée de cette démarche est de soumettre les modifications du débit au même apprentissage contextuel que les autres caractéristiques acoustiques.

Algorithme d'analyse-modifications-synthèse

Finalement, il est important de noter que les caractéristiques acoustiques à convertir dérivent des paramètres acoustiques extraits par la méthode d'analyse-modifications-synthèse, le choix de celles-ci conditionnera donc grandement le choix du vocodeur. Par exemple, si on souhaite manipuler les formants via les LSP, on préférera le modèle LPC. Dans notre étude, vu que l'on souhaite manipuler l'enveloppe spectrale via les MFCC, on s'orientera vers le vocodeur STRAIGHT.

8.2.5 Conversion du style et intelligibilité

Nous pouvons remarquer que l'aspect intelligibilité a peu été abordé dans cette section. Le renforcement par conversion du style de parole étant un domaine encore assez jeune, le nombre de travaux est assez limité. Dans les études de conversion de parole existantes, aucun travaux n'existent encore sur la parole claire et ceux sur la parole Lombard se concentrent sur des tests subjectifs de similarité et de qualité [175, 176, 108]. Concernant l'intelligibilité, les mesures objectives utilisées, comme le SII [108], ou le SIIB [175, 176], attestent systématiquement d'un gain d'intelligibilité théorique significatif. En revanche, le seul test subjectif d'intelligibilité recensé [176] n'obtient pas de résultats significatifs. Les auteurs suggèrent que le faible nombre de sujets serait principalement responsable mais que des dégradations liées à la synthèse du vocodeur pourrait aussi influencer les résultats.

Conclusion du chapitre 8

Dans ce chapitre, nous avons présenté les approches actuelles de renforcement paramétrique de la parole dans le bruit basées sur des outils d'analyse-modification-synthèse de la voix. D'abord, par modification de parole consistant à manipuler les paramètres avec des règles prédéfinies, ce qui nous a permis de mettre en évidence un manque important de prise en compte des spécificités des phonèmes et de leur contexte temporel dans les traitements proposés. Ensuite, par conversion du style consistant à convertir les paramètres d'une parole neutre à une parole d'un style donné en se basant sur des modèles de conversion : le GMM et le FFNN. Basés sur un apprentissage statistique, ces modèles prennent bien plus en compte les spécificités des phonèmes. En revanche, en supposant les échantillons comme étant indépendants, le contexte temporel n'est toujours pas pris en compte. Un artifice possible est alors d'utiliser les caractéristiques *delta*, et *delta-delta*, introduisant des notions dynamiques dans l'apprentissage mais à une échelle très locale. Ainsi, dans le chapitre suivant, nous proposerons plutôt l'utilisation d'une topologie d'ANN prenant directement en compte l'aspect séquentiel des signaux permettant alors une prise en compte du contexte temporel bien plus étendue.

Aussi, nous avons vu que la trajectoire du fondamental est le facteur principal perceptif permettant de caractériser une émotion, et que des modifications trop synthétique de celle-ci pourrait être responsable d'une dégradation de l'intelligibilité. Une représentation sur différentes échelles temporelles par une transformée en ondelettes sera alors proposée afin d'améliorer l'apprentissage de cette caractéristique capitale. Cette représentation pourra aussi être testée sur l'enveloppe énergétique afin d'étudier son intérêt.

Finalement, si le renforcement par conversion de parole est prometteur, la manipulation complexe des paramètres qu'il propose, et les soucis de synthèse liés au vocodeur, sont autant de facteurs qui risquent de s'opposer à l'amélioration de l'intelligibilité qui est initialement visée. Il convient donc d'abord de proposer des structures de conversion de parole performantes et adaptées aux styles étudiés, pour espérer atteindre des gains d'intelligibilité satisfaisants.

Chapitre 9

Propositions d'améliorations du traitement des aspects temporels en renforcement par conversion de la parole dans le bruit

Sommaire

Retour à la table des matières

Introduction du chapitre 9

Dans le chapitre précédent, nous avons détaillé en quoi les méthodes de renforcement par conversion du style de parole, présentées section 8.2, avait le potentiel d'améliorer les approches de renforcement par modification de la parole, présentées section 8.1. En prenant en compte les spécificités des phonèmes et leur contexte temporel local, les approches par conversion de parole permettraient une meilleure maîtrise des modifications inspirées par les styles naturels comme la parole claire et la parole Lombard. En revanche,nous avons aussi vu que le traitement de multiples aspects temporels pouvaient être bien plus développé.

Premièrement, l'utilisation de modèles comme le GMM et le FFNN, supposant les échantillons traités comme indépendants, est problématique pour la prise en compte du contexte temporel des phonèmes. L'utilisation des caractéristiques *delta* et *delta-delta* ne résout que partiellement le problème en ajoutant des caractéristiques dynamiques locales. Ainsi, nous introduisons dans ce chapitre l'utilisation d'une topologie récurrente d'ANN prenant en compte les relations entre les échantillons en traitant les séquences dans leur entièreté : le RNN et ses extensions.

Deuxièmement, la trajectoire du fondamental étant ce qui, perceptivement, caractérise le plus les émotions, un traitement spécifique de cette caractéristique devrait être considérée. Ainsi, afin d'aider le modèle d'apprentissage à capturer le comportement de cette caractéristique à différentes échelles, nous proposons de lui appliquer une transformée en ondelette et d'utiliser les coefficients comme caractéristique à convertir. Les effets de ce pré-traitement sur l'enveloppe énergétique pourrait aussi s'envisager.

Troisièmement, le traitement des modifications temporelles n'a jamais été abordé autrement que d'utiliser des règles empiriques tirées des méthodes de modifications de la parole. Nous proposerons donc dans ce chapitre une nouvelle modélisation des modifications temporelles sous forme d'une caractéristique directement intégrable dans les fonctions de conversion et ayant fait l'objet d'un dépôt de brevet d'invention.

L'adaptation des fonctions de conversion et des caractéristiques acoustiques sera présentée dans la section 9.1. Puis, la nouvelle modélisation des modifications temporelles proposée sera détaillée dans la section 9.2.

9.1 Adaptation des fonctions de conversion et des caractéristiques acoustiques

9.1.1 Conversion par réseau de neurones artificiels récurrents

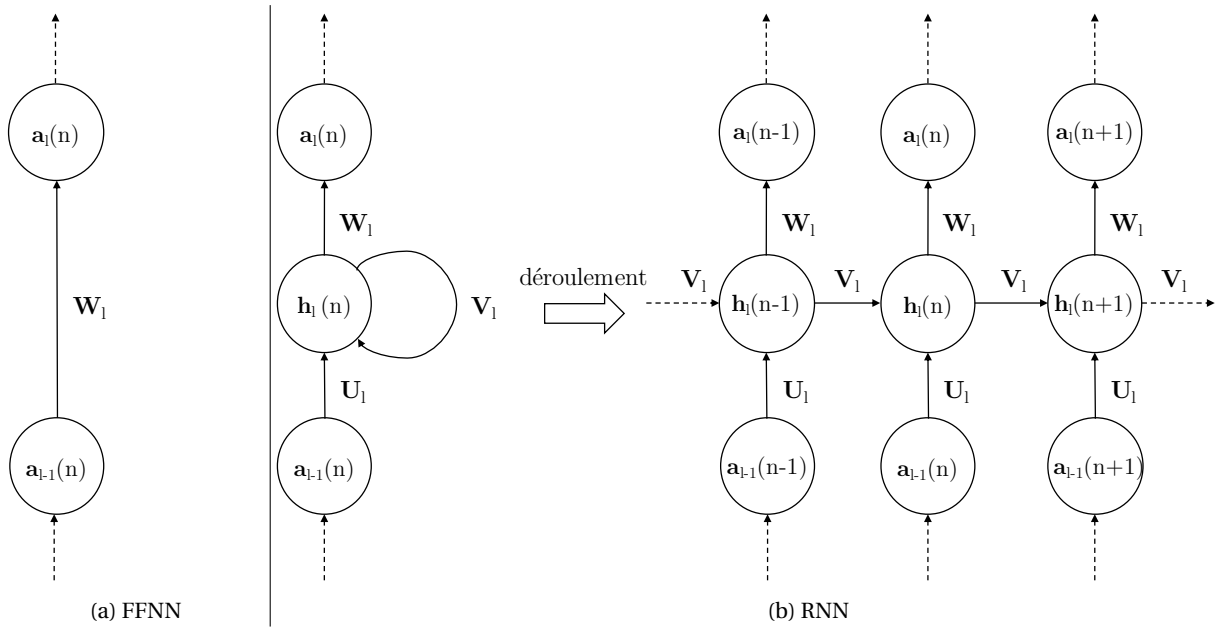
Le FFNN considère chaque échantillon de caractéristiques comme une observation indépendante, ainsi lorsque l'on traite des séquences temporelles comme en conversion de la parole, les informations sur la dynamique de la séquence sont perdues. Comme pour le GMM, il est possible d'introduire les caractéristiques dynamiques *delta* directement dans les données d'apprentissage. Comme nous l'avons vu équation 8.7, le calcul de ces caractéristiques s'effectue localement à partir de quelques échantillons dans le voisinage de celui qui est traité. Une autre possibilité est d'utiliser un RNN qui possède un fonctionnement similaire au FFNN mais prenant en compte les états de tous les échantillons précédents de la séquence dans son optimisation. Cela passe par la création d'états cachés $\mathbf{h}_l(n)$ pour chaque échantillon d'index n qui dépend des activations de ce même échantillon, mais aussi de l'état caché de l'échantillon précédent. En reprenant les notations de la section 8.2, on a :

$$\mathbf{h}_l(n) = \sigma^{(h)}\Big(\mathbf{U}_l\mathbf{a}_{l-1}(n) + \mathbf{V}_l\mathbf{h}_l(n-1)\Big), \tag{9.1}$$

$$\mathbf{a}_l(n) = \sigma^{(a)}\Big(\mathbf{W}_l\mathbf{h}_l(n)\Big). \tag{9.2}$$

avec $\mathbf{a}_l(n)$ les activations de l'échantillon d'index n sur la $l^{\text{ème}}$ couche du réseau, ($\mathbf{U}_l$, $\mathbf{V}_l$, $\mathbf{W}_l$) les poids associés initialisés aléatoirement et ($\sigma^{(a)}$, $\sigma^{(h)}$) les fonctions d'activation des neurones. On remarque que les poids sont partagés entre les échantillons, il ne dépendent pas de la position dans la séquence. Nous pouvons observer figure 9.1 une comparaison graphique entre les topologies des réseaux FFNN et RNN. Enfin, les $\mathbf{a}_0(n)$ correspondent aux échantillons des séquences de caractéristiques à l'entrée du réseau et s'expriment :

$$\forall n \in [\![1, \mathrm{N}_{min}]\!], \quad \mathbf{a}_0(n) = [\mathbf{X}_1(n)^\top, \mathbf{X}_2(n)^\top, ..., \mathbf{X}_\mathrm{M}(n)^\top]^\top \in \mathbb{R}^{\mathrm{M}\times \mathrm{C}_x}, \tag{9.3}$$



(a) FFNN (b) RNN

FIGURE 9.1 – Réseaux de neurones artificiels : propagation en avant et récurrent

avec $N_{min} = \min_m(N_m)$. La restriction sur le nombre d'échantillons vient du fait que les séquences qui vont servir à l'actualisation des poids au même passage doivent avoir la même taille. La solution proposée ici est de tronquer les séquences trop longues mais il est aussi possible de faire du bourrage de zéros pour allonger les séquences trop courtes, le raisonnement qui suit resterait inchangé. En pratique, l'utilisation de mini-lots permet de minimiser ce problème en regroupant des séquences de longueurs comparables pour chaque passage.

Vu que dorénavant une séquence entière correspond à une observation, la fonction de coût à minimiser est la somme des erreurs quadratiques sur chaque échantillon et s'exprime par :

$$J = \sum_{n=1}^{N_{min}} \|\boldsymbol{\mathcal{Y}}(n) - \mathbf{a}_L(n)\|^2, \tag{9.4}$$

avec :

$$\boldsymbol{\mathcal{Y}}(n) = [\mathbf{Y}_1(n)^\top, \mathbf{Y}_2(n)^\top, ..., \mathbf{Y}_M(n)^\top]^\top \in \mathbb{R}^{M \times C_y}. \tag{9.5}$$

Une descente de gradient est alors appliquée sur les poids afin de minimiser la fonction de coût. La dépendance temporelle va compliquer les calculs mais développant les dérivées des équations de récurrence, l'influence des états précédents peut être aisément prise en compte dans le calcul des gradients par rétropropagation et on parle alors de rétropropagation du gradient à travers le temps. En utilisant B mini-lots, un taux d'affaiblissement des poids λ, et un taux d'apprentissage α, l'actualisation des poids s'effectue alors par la formule suivante :

$$\forall\, l \in [[1, L]], \quad \mathbf{W}_l = \mathbf{W}_l - \alpha \left(\lambda \mathbf{W}_l + \sum_{b=1}^{B} \sum_{n=1}^{N_{min}} \frac{\|\boldsymbol{\mathcal{Y}}_b(n) - \hat{\boldsymbol{\mathcal{Y}}}_b(n)\|^2}{\partial \mathbf{W}_l} \right), \tag{9.6}$$

Le fonctionnement du RNN présenté ici est unidirectionnel (UD) car il prédit chaque échantillon de la séquence basé sur les échantillons passés. Cependant, il peut être intéressant de travailler avec un RNN bidirectionnel (BD) qui les prédirait sur les échantillons passés et futurs. La bidirectionnalité est simplement obtenue en combinant la sortie de deux RNN UD traitant la séquence dans les deux directions.

Un problème récurrent dans les ANN est qu'en multipliant les dérivées partielles à la chaîne, les gradients calculés peuvent diverger (*exploding gradient*) ou se dissiper (*vanishing gradient*). L'explosion des gradients peut être aisément repérée et maîtrisée par des méthodes de saturation. La dissipation des gradients quant à elle demande un intérêt plus particulier. Un facteur important de ce phénomène concerne les fonctions d'activations σ

utilisées, on évitera les fonctions dont la dérivée s'annule de façon symétrique comme les sigmoïdes : la fonction unité linéaire rectifiée, usuellement abrégé par l'acronyme anglais pour *Rectified Linear Unit* (ReLU), s'est imposée comme l'alternative la plus populaire par la stabilité qu'elle procure à l'optimisation. Bien que cette solution procure des résultats satisfaisant dans la majorité des applications du FFNN, le RNN reste toujours très impacté par ce phénomène qui l'empêche d'apprendre des dépendances long-terme en dissipant l'influence des états à des instants trop éloignés. De nouvelles cellules de calcul de l'état caché ont été proposées afin de corriger ce problème par un mécanisme de portes sélectionnant l'information à conserver ou non dans la mémoire du réseau permettant alors de véhiculer des informations sur de grandes distances.

RNN et cellule LSTM

La cellule de mémoire à long/court terme, usuellement abrégé par l'acronyme anglais pour *Long Short-Term Memory* (LSTM), permet cela en introduisant un état supplémentaire appelé état de la cellule $\mathbf{c}_l(n)$ et trois portes permettant de moduler les mouvements d'information qui y transite. La structure d'une cellule LSTM est visible figure 9.2. Les trois portes introduites sont les suivantes :

- une porte d'oubli, dont le vecteur $\mathbf{o}_l(n)$ permet de sélectionner l'information de l'état de la cellule précédente $\mathbf{c}_l(n-1)$ à conserver dans l'état de la cellule actuelle $\mathbf{c}_l(n)$,
- une porte d'entrée, dont le vecteur $\mathbf{e}_l(n)$ permet de sélectionner l'information d'un candidat $\mathbf{q}_l(n)$ à introduire dans l'état de la cellule actuelle $\mathbf{c}_l(n)$,
- une porte de sortie, dont le vecteur $\mathbf{s}_l(n)$ permet de sélectionner l'information dans l'état final de la cellule actuelle $\mathbf{c}_l(n)$ à introduire dans l'état caché $\mathbf{h}_l(n)$,

Les vecteurs de sélection des différentes portes se calculent classiquement à partir de l'activation actuelle et de l'état caché précédent par les formules suivantes :

$$\mathbf{o}_l(n) = \sigma^{(p)}\left(\mathbf{U}_l^{(o)}\mathbf{a}_{l-1}(n) + \mathbf{V}_l^{(o)}\mathbf{h}_l(n-1)\right), \tag{9.7}$$

$$\mathbf{e}_l(n) = \sigma^{(p)}\left(\mathbf{U}_l^{(e)}\mathbf{a}_{l-1}(n) + \mathbf{V}_l^{(e)}\mathbf{h}_l(n-1)\right), \tag{9.8}$$

$$\mathbf{s}_l(n) = \sigma^{(p)}\left(\mathbf{U}_l^{(s)}\mathbf{a}_{l-1}(n) + \mathbf{V}_l^{(s)}\mathbf{h}_l(n-1)\right), \tag{9.9}$$

avec $\sigma^{(p)}$ une sigmoïde car ses valeurs bornées entre 0 et 1 sont parfaitement adaptées pour permettre aux portes de remplir leur rôle de sélection d'information. Les calculs appliqués par une cellule LSTM sont les suivants :

$$\mathbf{q}_l(n) = \sigma^{(q)}\left(\mathbf{U}_l^{(q)}\mathbf{a}_{l-1}(n) + \mathbf{V}_l^{(q)}\mathbf{h}_l(n-1)\right) \tag{9.10}$$

$$\mathbf{c}_l(n) = \mathbf{c}_l(n-1) \circ \mathbf{o}_l(n) + \mathbf{q}_l(n) \circ \mathbf{e}_l(n), \tag{9.11}$$

$$\mathbf{h}_l(n) = \sigma^{(h)}\Big(\mathbf{c}_l(n) \circ \mathbf{s}_l(n)\Big) \tag{9.12}$$

L'activation est ensuite calculée classiquement par :

$$\mathbf{a}_l(n) = \sigma^{(a)}\Big(\mathbf{W}_l\mathbf{h}_l(n)\Big). \tag{9.13}$$

On remarque que si on force les poids d'oubli à 0 (on fait abstraction de l'état de la cellule précédente) ainsi que les poids d'entrée et de sortie à 1, on obtient une cellule RNN classique.

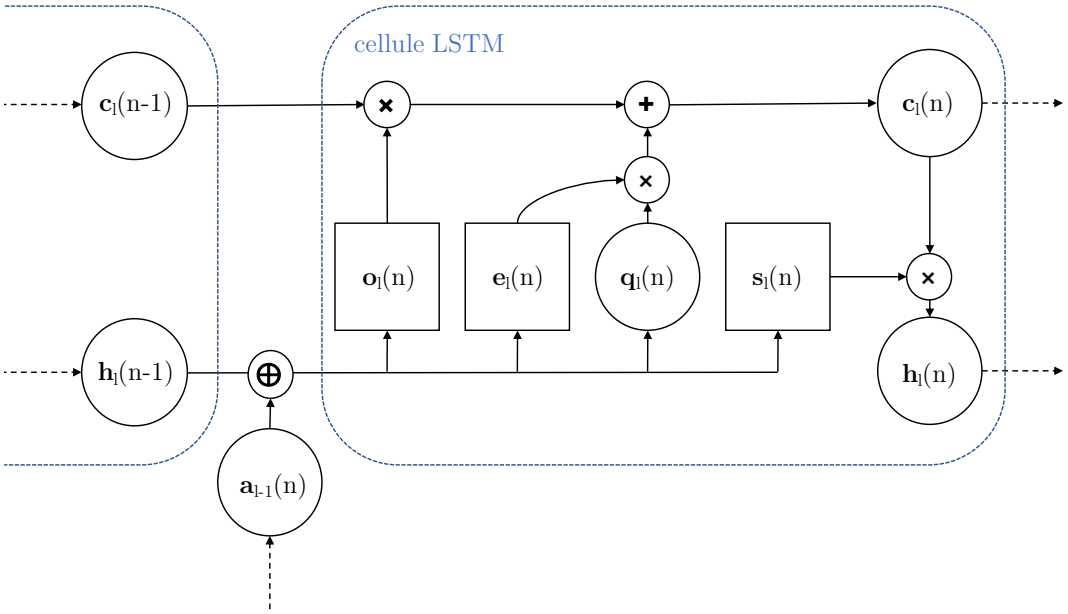


FIGURE 9.2 – Structure d'une cellule LSTM.

RNN et cellule GRU

La cellule de porte récurrente, usuellement abrégé par l'acronyme anglais pour *Gated Recurent Unit* (GRU), propose une version simplifiée de la cellule LSTM mais avec une approche très similaire. La structure d'une cellule GRU est visible figure 9.3. Il n'y a pas de création d'état de la cellule et seulement deux portes sont utilisées :

- une porte de ré-initialisation, dont le vecteur $\mathbf{r}_l(n)$ permet de sélectionner l'information de l'état caché précédent $\mathbf{h}_l(n-1)$ à combiner avec l'activation précédente $\mathbf{a}_{l-1}(n)$,
- une porte de mise à jour, dont le vecteur $\mathbf{m}_l(n)$ permet de sélectionner l'information de l'état caché précédent $\mathbf{h}_l(n-1)$ à conserver dans l'état caché actuel $\mathbf{h}_l(n)$.

Voici les équations qui en résultent :

$$\mathbf{r}_l(n) = \sigma^{(p)}\left(\mathbf{U}_l^{(r)}\mathbf{a}_{l-1}(n) + \mathbf{V}_l^{(r)}\mathbf{h}_l(n-1)\right), \tag{9.14}$$

$$\mathbf{m}_l(n) = \sigma^{(p)}\left(\mathbf{U}_l^{(u)}\mathbf{a}_{l-1}(n) + \mathbf{V}_l^{(u)}\mathbf{h}_l(n-1)\right), \tag{9.15}$$

$$\mathbf{q}_l(n) = \sigma^{(q)}\left(\mathbf{U}_l^{(q)}\mathbf{a}_{l-1}(n) + \mathbf{V}_l^{(q)}(\mathbf{h}_l(n-1) \circ \mathbf{r}_l(n))\right) \tag{9.16}$$

$$\mathbf{h}_l(n) = \mathbf{h}_l(n-1) \circ \mathbf{m}_l(n)) + \mathbf{q}_l(n) \circ (1 - \mathbf{m}_l(n)) \tag{9.17}$$

$$\mathbf{a}_l(n) = \sigma^{(a)}\left(\mathbf{W}_l\mathbf{h}_l(n)\right). \tag{9.18}$$

On remarque que si on force les poids de ré-initialisation à 1 (on conserve l'intégralité de l'état caché précédent) ainsi que les poids de mise à jour à 0, on obtient une cellule RNN classique.

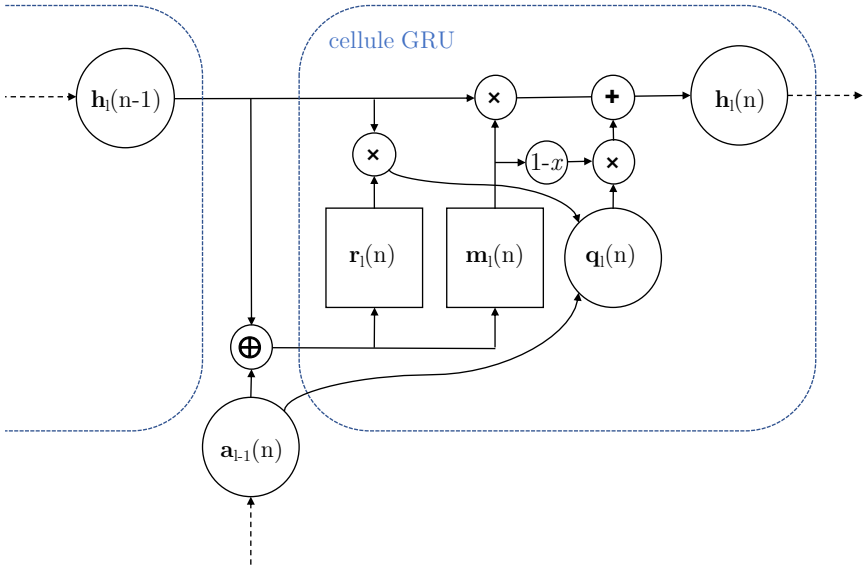


FIGURE 9.3 – Structure d'une cellule GRU.

9.1.2 Problèmes d'échelles temporelles et représentation de caractéristiques par transformée en ondelettes continue

L'utilisation des caractéristiques *delta*, ou d'un RNN, pour prendre en compte la dynamique des caractéristiques ont été introduites précédemment. En revanche, on sait que les modifications du fondamental et de l'intensité de la parole interviennent à de multiples échelles temporelles (phonème, syllabe, mot, phrase) et ces nombreuses dépendances temporelles ne sont pas toujours bien capturées par ces méthodes. Ainsi, il peut être intéressant de les assister par une représentation plus pertinente. Il a déjà été suggéré d'utiliser une transformée en ondelettes continue, usuellement abrégée par l'acronyme anglais CWT, pour représenter ces paramètres prosodiques sur différentes échelles temporelles [164] et cela a été utilisé avec succès en conversion du style en conversion d'émotion [125, 126] mais jamais en renforcement de la parole. Voici les détails de l'utilisation de la CWT proposée.

Pour une séquence de caractéristique unidimensionnelle $\mathbf{x} \in \mathbb{R}^{\mathrm{N}_m \times 1}$, sa CWT réelle à l'échelle $s \in \mathbb{R}_+^*$ se calcule de la façon suivante :

$$\forall n \in [\![1, \mathrm{N}_m]\!], w_s(n) = \sum_{n'=0}^{\mathrm{N}_m - 1} x(n') \psi\left(\frac{n' - n}{s}\right) \tag{9.19}$$

avec $\psi(t)$ l'ondelette mère. D'abord, si la caractéristique étudiée n'est pas définie sur toute la séquence, une interpolation linéaire est appliqué sur ces intervalles comme les segments non-voisés et les silences pour la trajectoire du fondamental. Ensuite, la caractéristique est standardisée (centrée, réduite) pour une meilleure interprétation de l'analyse en ondelette. Enfin, une CWT sur O octaves est appliquée sur la caractéristique avec une ondelette mère de taille s_0. Les nouvelles caractéristiques résultantes, notées $\mathbf{x}^{cwt} \in \mathbb{R}^{\mathrm{N}_m \times \mathrm{O}}$, sont représentées par les composantes :

$$\forall (n, k) \in [\![1, \mathrm{N}_m]\!] \times [\![0, \mathrm{O} - 1]\!], x^{cwt}(k, n) = w_{2^k s_0}(n)\,. \tag{9.20}$$

Un exemple d'une trajectoire de fondamental, en valeurs logarithmiques, standardisée et de ses composantes CWT associées sont visibles figure 9.4. Ces résultats ont été obtenus sur O = 10 octaves avec une ondelette mère de Ricker (familièrement appelée "chapeau mexicain") d'une largeur initiale de 10 ms.

À partir des composantes CWT, la reconstruction est effectuée en utilisant les équations suivantes [206] :

$$\forall n \in [\![1, \mathrm{N}_m]\!], \hat{x}(n) = \frac{1}{\mathrm{R}\psi_0(0)} \sum_{k=0}^{\mathrm{O}-1} \frac{x^{cwt}(k, n)}{\sqrt{2^k s_0}}, \tag{9.21}$$

avec $\psi_0(0)$ l'amplitude à l'origine de l'ondelette mère et R son facteur de reconstruction donné.

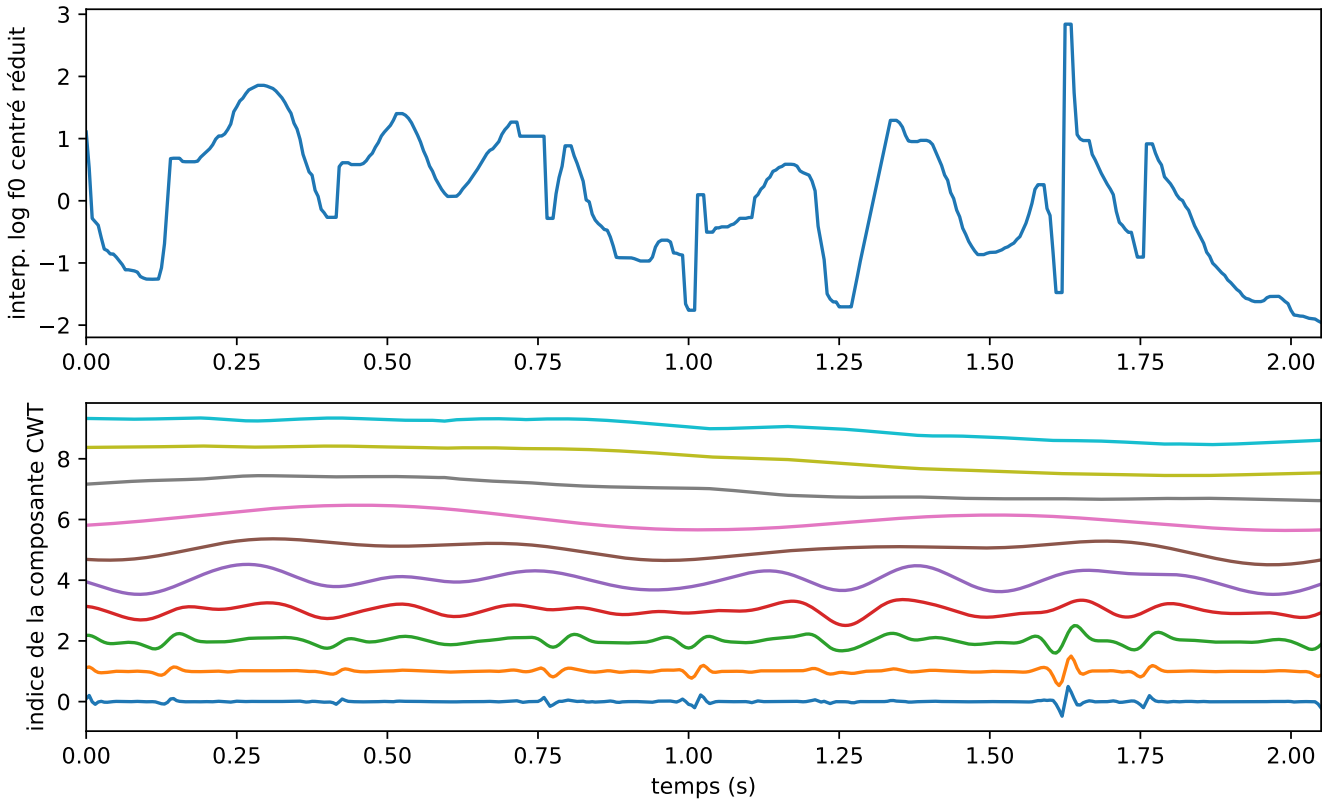


FIGURE 9.4 – Exemple d'une trajectoire centrée réduite du log fondamental, interpolée sur les silences et segments non-voisés, et de ses composantes CWT associées.

9.2 Modélisation et lissage des modifications temporelles

Comme introduit précédemment, les approches de conversion de la parole basent les modifications de débit sur de simples règles statistiques et conditionnelles. À notre connaissance, aucun travail ne propose d'intégrer les modifications du débit dans la fonction de conversion. Nous proposons donc une nouvelle façon d'aborder le problème en exploitant l'alignement temporel des signaux d'apprentissage afin d'extraire une caractéristique de modification temporelle directement intégrable dans la fonction de conversion.

Une hypothèse de travail facilitant le calcul et l'analyse des résultats présentés dans cette section est de supposer que les caractéristiques des signaux ont une période d'analyse fixe et identique. Ce n'est pas le cas, par exemple, pour les outils d'analyse qui fournissent des valeurs sur des fenêtres synchrones au fondamental, pour lesquels il faudra ajuster le raisonnement. Pour nos travaux, nous travaillons avec STRAIGHT qui fonctionne avec des valeurs synchrones au fondamental lors de l'analyse et de la synthèse, mais l'outil dont le principe est d'analyser les caractéristiques en haute-résolution fournit les valeurs de celles-ci à intervalles réguliers, de 5ms par défaut. Nous considérerons donc cette hypothèse de travail vérifiée dans toute la section.

9.2.1 Problèmes de modélisation des modifications temporelles

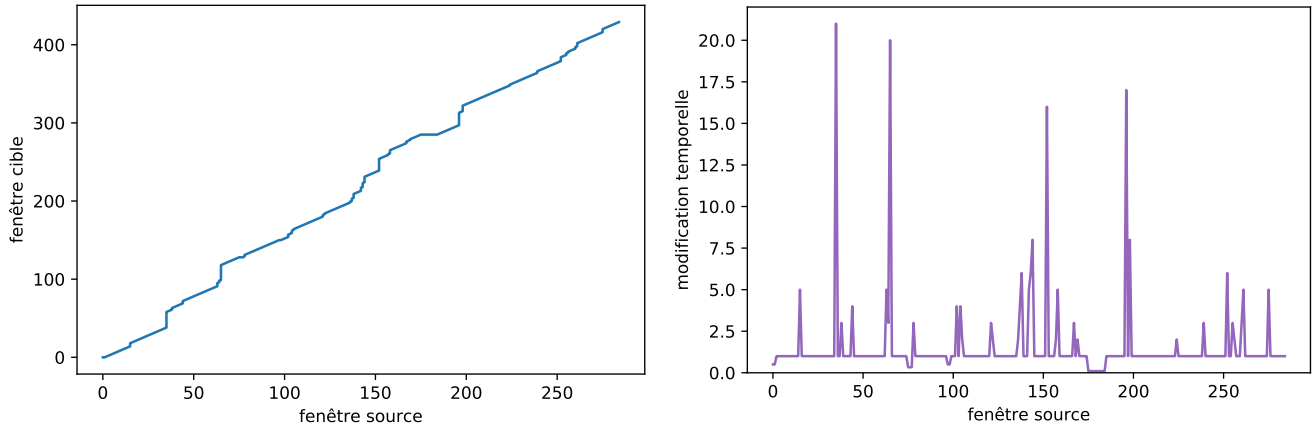
Lors d'un apprentissage parallèle en conversion de la parole, les caractéristiques acoustiques des deux signaux d'une paire sont alignées temporellement par un algorithme d'alignement temporel. On utilise classiquement la méthode de déformation temporelle dynamique, usuellement abrégé par l'acronyme anglais pour *Dynamic Time Warping* (DTW), consistant à calculer un chemin optimal entre deux séquences qui minimise une distance moyenne entre les fenêtres associées. À chaque fenêtre source est associé une ou plusieurs fenêtres cibles, et vice-versa. On peut alors observer le chemin d'alignement en traçant les associations entre les fenêtres sources et cibles. Un exemple de chemin d'alignement, entre deux signaux fenêtrés d'une paire source/cible, obtenu par DTW sur les 12 premiers coefficients MFCC est visible figure 9.5a. Le coefficient énergétique a été retiré et l'alignement est effectué avec la distance mel-cepstrale, usuellement abrégé par l'acronyme anglais pour *Mel-Cepstral Distance* (MCD). Du point de vue de la modification du débit de parole, le chemin d'alignement peut être interprété de la façon suivante.

- Lorsque plusieurs fenêtres cibles sont associées à une seul fenêtre source, observable par un segment vertical du chemin d'alignement, cela correspond à un ralentissement du débit. Le nombre de fenêtres cibles donne alors le facteur de modification temporel appliqué au niveau de la fenêtre source.
- Lorsqu'une seul fenêtre cible est associée à plusieurs fenêtres sources, observable par un segment horizontal du chemin d'alignement, cela correspond à une accélération du débit. L'inverse du nombre de fenêtres sources donne alors le facteur de modification temporel appliqué au niveau de ces fenêtres sources.
- Lorsqu'une seule fenêtre cible est associée à une seule fenêtre source, observable par un segment diagonal du chemin d'alignement, cela correspond à une conservation du débit. Le facteur de modification temporel appliqué au niveau de la fenêtre source est alors unitaire.

Un tracé de ces facteurs de modification temporel bruts est visible figure 9.5b

Mathématiquement, ce sont bien ces modifications temporelles qu'il faudrait appliquer à chaque fenêtre du signal de parole source afin de minimiser la MCD avec le signal de parole cible et il faudrait donc utiliser cette nouvelle caractéristique dans la fonction de conversion. Cependant, on remarque que tels quels les facteurs possèdent des valeurs très éparses ce qui soulève plusieurs problèmes majeurs. Premièrement, cela ne reflète pas des modifications naturelles de la parole qui devraient être plus lisses. Deuxièmement, la fonction de conversion aura des difficultés à apprendre une telle caractéristique disparate. Finalement, comme nous l'avons vu lors de l'introduction des différents vocodeurs chapitre 2, des valeurs trop extrêmes de modifications temporelles introduisent des dégradations importantes lors de la re-synthèse.

Une solution directe consiste à calculer les facteurs de modification non plus par fenêtre mais en regroupant les fenêtres correspondant à une même classe phonétique ou partageant un même trait phonétique comme le



(a) exemple d'un chemin d'alignement entre deux séquences

(b) facteurs de modification temporelle bruts

FIGURE 9.5 – Exemple de chemin d'alignement entre deux séquences temporelles obtenu par DTW et facteurs de modification temporelle bruts résultants.

voisement. Ainsi, en faisant le rapport entre le nombre de fenêtres cibles d'un certain type et le nombre de fenêtres sources de ce même type, on obtient un facteur de modification temporelle unique pour chaque type. C'est ce que propose les approches actuelles en calculant des modifications temporelles moyennes basées majoritairement sur le voisement. Dans le cadre de notre étude, nous avons calculé ces facteurs moyens en fonction du voisement sur les deux bases de données qui ont été introduites chapitre 3, celle de parole Lombard Lombard-GRID [6] et celle de parole claire LUCID [15]. Les valeurs sont visibles dans le tableau 9.1, les silences ont été pris en compte en dehors du voisement. On note des valeurs moyennes cohérentes avec les résultats des études antérieures sur ces styles de parole, avec un ralentissement des sons voisés ainsi qu'une légère accélération des sons non-voisés et des silences pour la parole Lombard, et un ralentissement important peu importe le voisement ainsi qu'un allongement des silences extrêmement marqué pour la parole claire. Cette simplification du calcul retire le besoin de fonction de conversion mais les modifications temporelles résultantes, qui sont maintenant représentés par des fonctions en escalier, ne sont toujours pas naturelles et ne prennent pas du tout en compte le contexte et les spécificité des phonèmes autre que leur voisement.

À notre connaissance, la seule étude ayant proposé de contextualiser les modifications temporelles est pour de la conversion d'émotion pour les voix de synthèse. Dans cette étude, INANOGLU et al. utilisent un arbre de décision pour choisir le facteur de modification temporelle à appliquer à un phonème en fonction de sa durée et de son contexte (position du phonème dans le mot, position du mot dans la phrase...) [80]. En rajoutant de nombreux paramètres dans le choix du facteur, cette approche procure une flexibilité des modifications aboutissant à un rendu plus naturel pour la conversion d'émotions qui se basent fortement sur le débit comme la colère. Cependant, cela nécessite un traitement spécifique basé sur une connaissance du contexte difficilement obtenable en traitement de signaux non-synthétiques. Nous proposons alors une nouvelle approche exploitant le chemin d'alignement temporel afin d'extraire une caractéristique de modification du débit naturelle, flexible et facilement intégrable dans une fonction de conversion.

base de données	voisé	non-voisé	silence
Lombard-GRID	1,14	0,94	0,90
LUCID (parole claire)	1,59	1,52	3,46

TABLEAU 9.1 – Facteurs de modification temporelle moyens en fonction du voisement calculés sur deux bases de donnée : Lombard-GRID (parole Lombard) et LUCID (parole claire).

9.2.2 Proposition de modélisation des modifications temporelles

Le principe de l'approche nouvelle proposée est d'approximer le chemin d'alignement temporel afin d'exploiter sa dérivée pour extraire une trajectoire de facteurs de modification temporelle lisse et continue et donc plus naturelle. Plusieurs contraintes émergent alors de cet énoncé :

1. l'approximation doit être dérivable afin de pouvoir calculer les facteurs de modification temporelle,
2. la dérivée de l'approximation doit elle aussi être dérivable afin d'assurer une trajectoire de facteurs continue et lisse,
3. l'approximation doit être strictement croissante afin que les facteurs soient strictement positifs.

La première étape consiste à choisir la méthode d'approximation du chemin d'alignement temporel. Une solution pratique est l'utilisation d'un modèle additif généralisé avec lissage par N splines cubiques. Les splines d'ordre p étant de classe C^{p-1}, prendre $p \geq 3$ permet de répondre aux contraintes 1 et 2. De plus, l'ajout d'une contrainte de stricte monotonicité permet de répondre à la contrainte 3. Des approximations du chemin d'alignement observé précédemment sont visibles figure 9.6, celles-ci ont été effectuées avec 20, 50 et 150 splines d'ordre 3 que l'on peut observer sur chaque sous-figure. On remarque que le chemin a bien été lissé et que la contrainte de stricte monotonicité est bien respectée. Naturellement, plus le nombre de splines est grand, plus l'approximation est proche du chemin d'alignement, il sera cependant intéressant d'étudier l'influence de ce paramètre dans la suite de notre étude.

La deuxième étape consiste simplement à calculer la dérivée de l'approximation au niveau de chaque fenêtre source afin d'obtenir les facteurs de modification temporelle. La trajectoire des modifications temporelles ainsi obtenu avec 20, 50 et 150 splines d'ordre 3 pour le chemin d'alignement observé précédemment sont visibles figure 9.6d. On remarque qu'en augmentant le nombre de splines, la trajectoire tend effectivement vers les modifications brutes observable figure 9.5b. Cependant les variations sont aussi plus brusques et les valeurs plus extrêmes, ce qui nous ramène aux problèmes décrits dans la section précédente. Choisir un nombre modéré de splines semble alors être intéressant afin d'atténuer les variations du facteur tout en conservant les tendances très localisées des modifications brutes.

D'autres exemples de modélisation des modifications temporelles obtenues à partir des bases de données Lombard-GRID et LUCID introduites précédemment sont visibles respectivement figure 9.7 et figure 9.8.

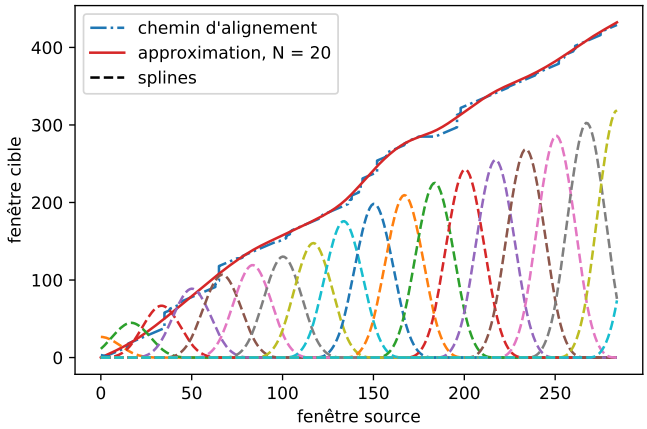
9.2.3 Performances objectives de la modélisation

Afin de mesurer l'intérêt de cette nouvelle modélisation, nous avons mis en place une mesure de distance temporelle entre deux signaux de parole basée sur le chemin d'alignement entre leurs séquences de coefficients MFCC. Ainsi en appliquant les facteurs de modifications temporelles de différentes modélisations sur les signaux normaux d'une base de données, la distance temporelle mesurée, entre le signal modifié et le signal du style visé, reflète les performances de chaque modélisation.

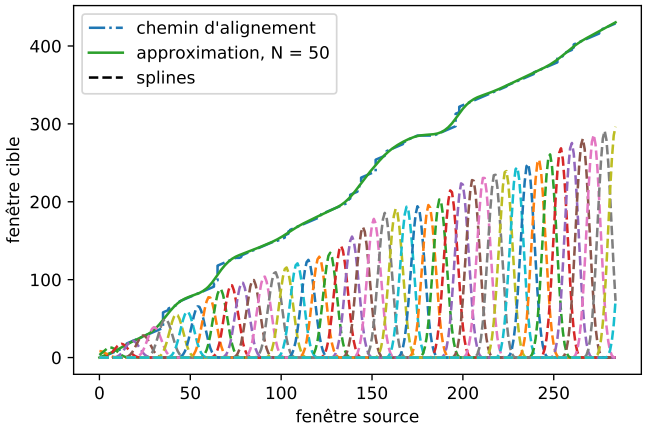
Distance temporelle proposée

La distance temporelle entre deux signaux de parole proposée consiste simplement à appliquer une DTW sur leurs séquences de coefficients MFCC et à compter le nombre de fenêtres qui ne sont pas alignées. Une façon graphique de se représenter cette distance étant qu'elle correspond à la somme de la longueur de tous les segments verticaux et horizontaux du chemin d'alignement entre les deux séquences. Par exemple si le chemin d'alignement est parfaitement diagonal à part en deux endroits où une fenêtre cible est associée à 3 fenêtres sources (segment horizontal i.e. accélération du débit) et où une fenêtre source est associée à 5 fenêtres cibles (segment vertical i.e. ralentissement du débit) alors la distance D_t entre les deux séquences est de $D_t = (3-1) + (5-1) = 6$.

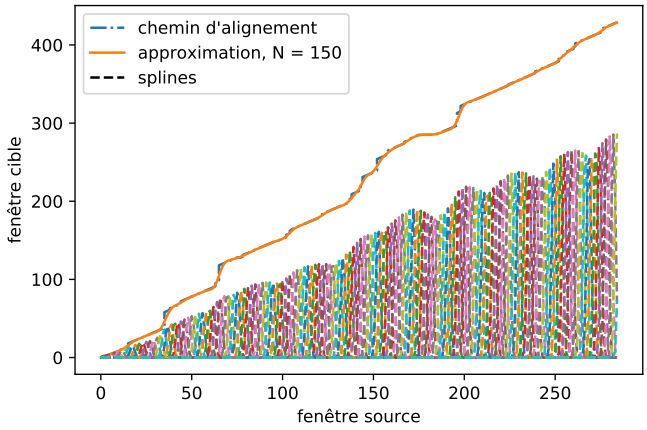
Nous appliquons donc les facteurs de modifications temporelles issus de différentes modélisations sur chaque signal de parole naturelle des bases de données Lombard-GRID et LUCID avec l'outil STRAIGHT. Les distances



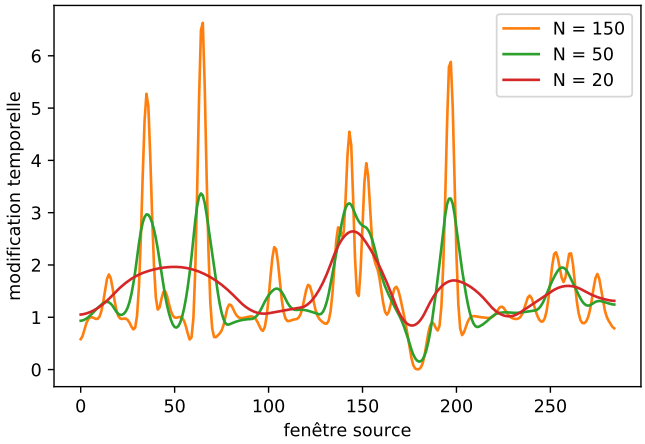
(a) N = 20



(b) N = 50



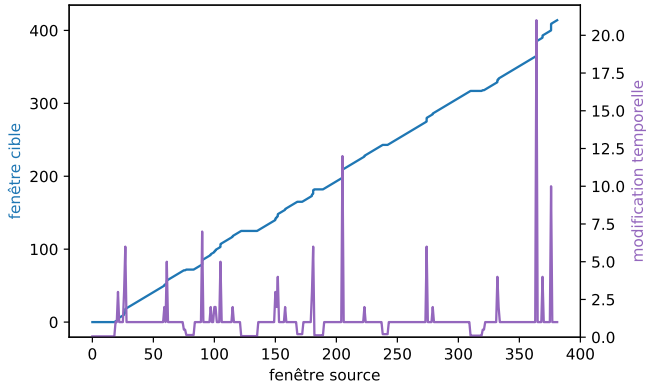
(c) N = 150



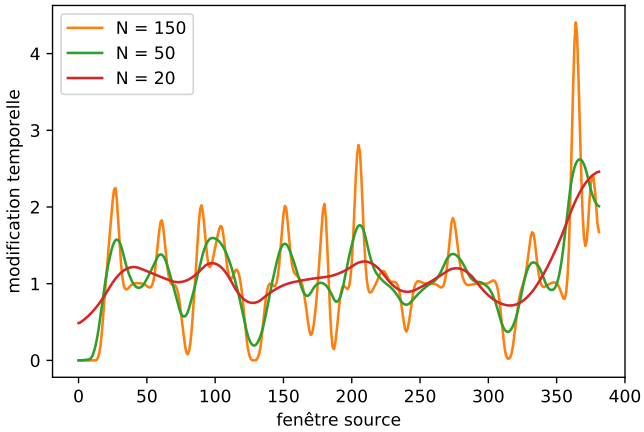
(d) Facteurs de modification temporelle lisses.

FIGURE 9.6 – Approximations du chemin d'alignement et modifications temporelles lisses résultantes.

locuteur n°15, phrase : "*place red with L five now*"

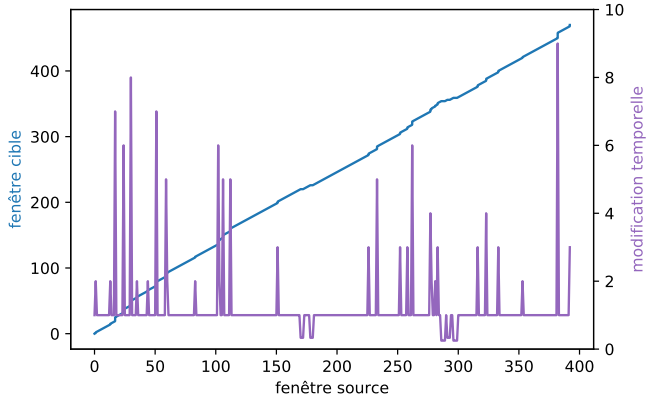


(a) chemin et facteurs bruts

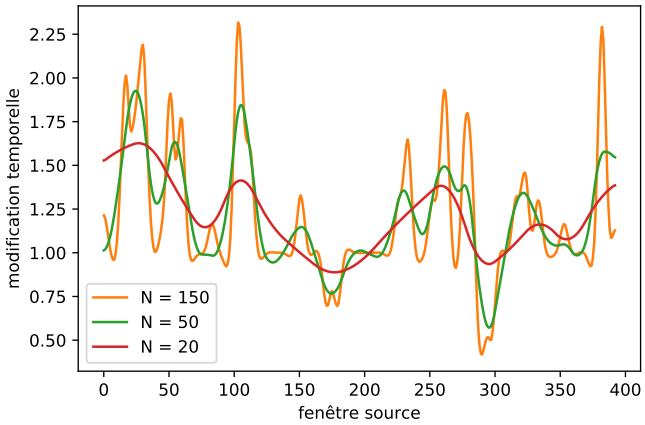


(b) facteurs lissés

locuteur n°33, phrase : "*lay green with U nine soon*"

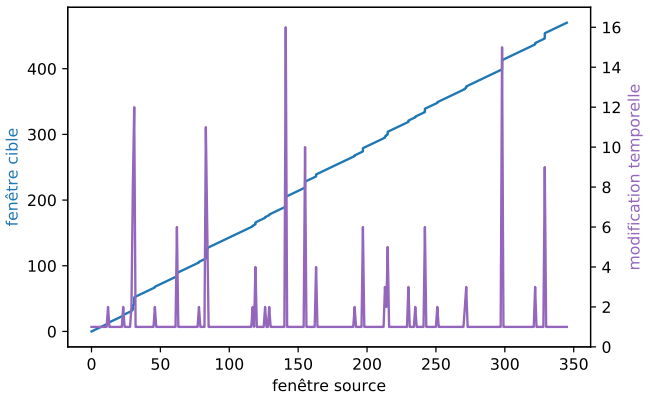


(c) chemin et facteurs bruts

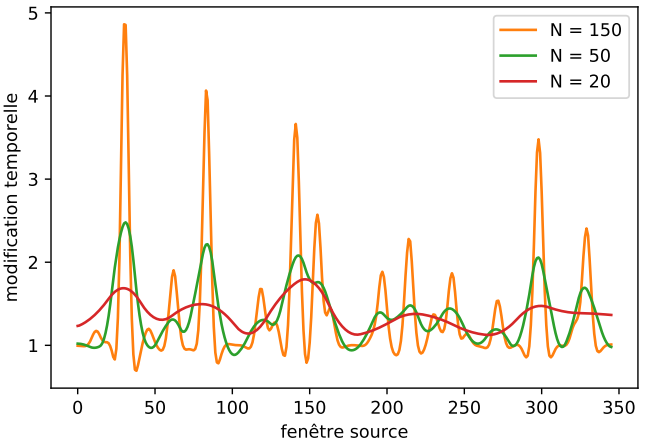


(d) facteurs lissés

locuteur n°50, phrase : "*lay green with N six please*"



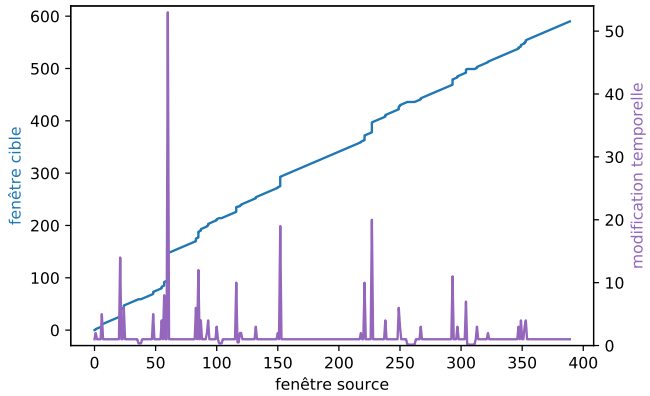
(e) chemin et facteurs bruts



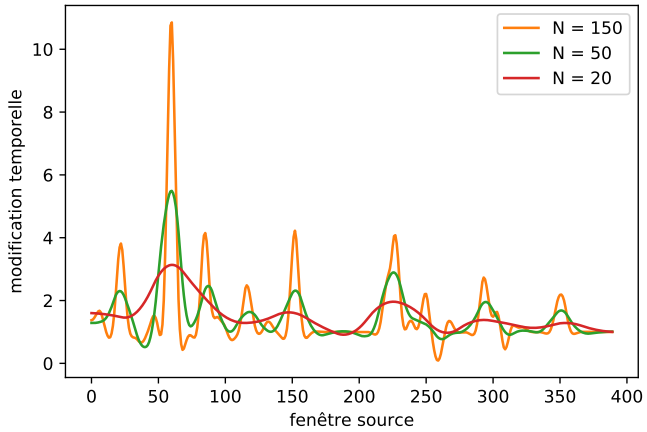
(f) facteurs lissés

FIGURE 9.7 – Exemples de modélisation des modifications temporelles sur la base de données Lombard-GRID

locuteur n°13, phrase : "*The shop was called Bear Essentials.*"

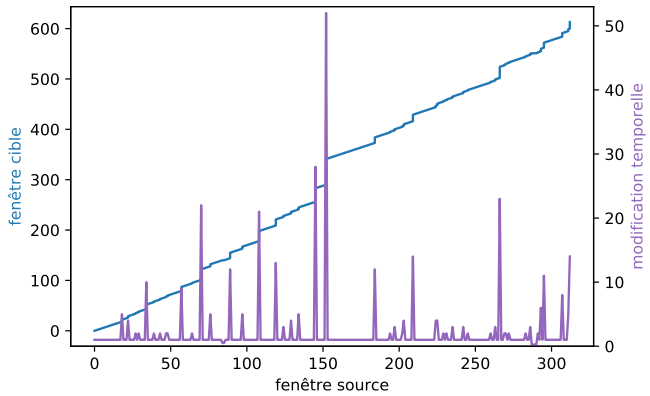


(a) chemin et facteurs bruts

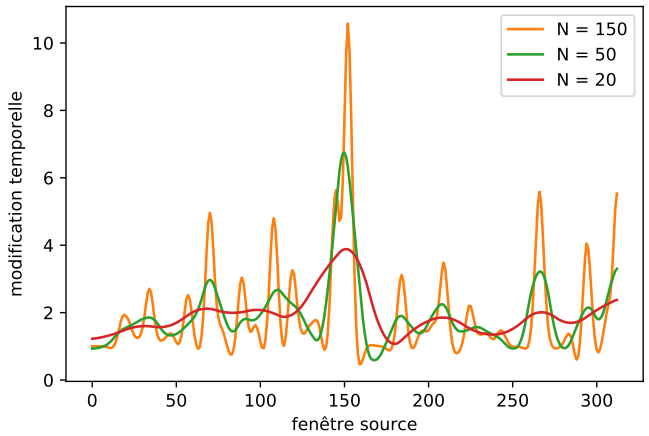


(b) facteurs lissés

locuteur n°25, phrase : "*Daisy the sheep was grazing.*"

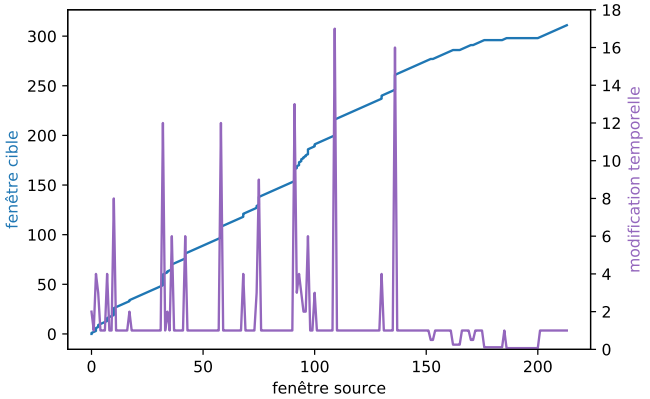


(c) chemin et facteurs bruts

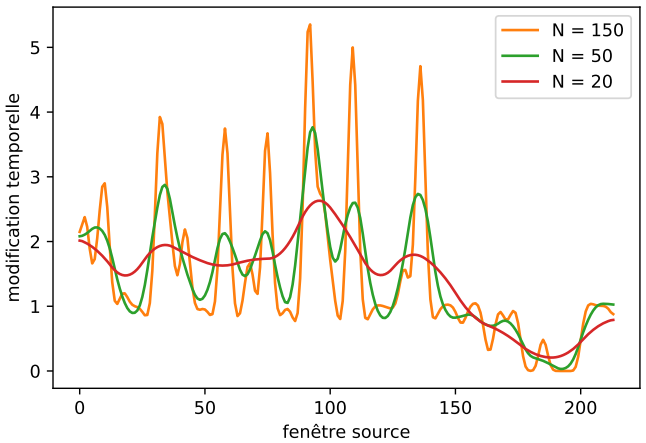


(d) facteurs lissés

locuteur n°55, phrase : "*All the pins were sharp.*"



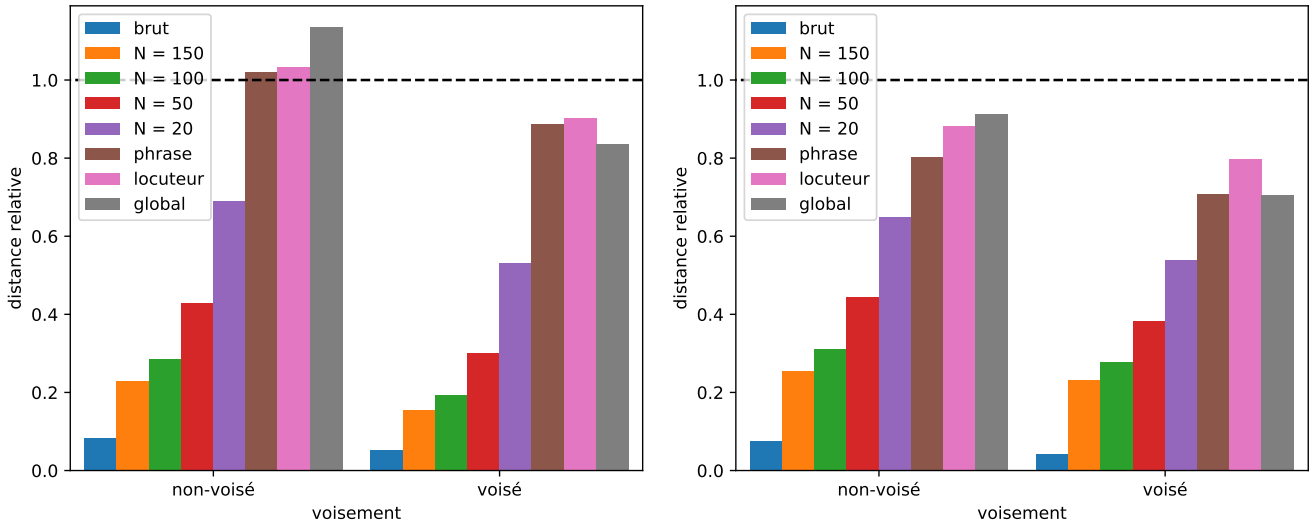
(e) chemin et facteurs bruts



(f) facteurs lissés

FIGURE 9.8 – Exemples de modélisation des modifications temporelles sur la base de données LUCID

temporelles entre les signaux traités et les signaux visés correspondants sont alors calculées. Les différentes modélisation des facteurs de modifications temporelles étudiées sont les facteurs bruts (qui devraient naturellement obtenir la distance la plus faible), ceux obtenus par la nouvelle modélisation proposée (pour différents nombres de splines) et ceux obtenus classiquement en moyennant sur le type de voisement (à l'échelle de la phrase, du locuteur, ou globale). Pour une bonne lecture et interprétation des résultats, nous avons aussi calculé une distance temporelle de référence entre les signaux de parole naturelle et les signaux du style visé correspondants. Cette distance permet de normaliser les distances obtenues pour chaque signaux afin d'obtenir une distance relative à une absence de modification. Les graphiques des résultats des distances relatives moyennes obtenues sur les deux bases de données, et pour chaque type de voisement, sont consultables figure 9.9.



(a) Lombard GRID

(b) LUCID

FIGURE 9.9 – Performances de la modélisation des modifications temporelles proposée sur deux bases de données.

Analyse des résultats

Premièrement, on remarque qu'en utilisant les facteurs bruts, la distance relative est bien minimale mais non-nulle. Cela s'explique par la re-synthèse des signaux avec les modifications temporelles suivie de la nouvelle analyse qui peuvent entraîner de légers changements dans le nouveau chemin d'alignement qui ne sera pas parfaitement diagonal.

Pour ce qui est des modélisations classiques basées sur des facteurs fixés, avec pour seul paramètre le voisement, on remarque des résultats aux comportements très similaires entre les deux bases de données. En effet, pour les modélisations moyennes :

- globales, dont les facteurs utilisés sont indiqués tableau 9.1, on note une distance relative intéressante pour les sons voisés, mais bien moindre pour les sons non-voisés,
- spécifiques aux locuteurs, l'écart de distance relative entre les deux types de voisement se resserre avec des performances moindres pour les sons voisés et plus intéressantes pour les sons non-voisés,
- spécifiques aux phrases, on retrouve une diminution de la distance relative pour les sons voisés et toujours une diminution pour les sons non-voisés.

Il est clair que les sons non-voisés profitent pleinement du changement d'échelle avec une évolution positive des performances lorsque les facteurs sont calculés sur des échelles de plus en plus proches de la phrase traitée. Plus surprenant, les performances pour les sons voisés semblent être dégradées par le changement d'échelle.

La différence majeure entre les résultats des modélisations classiques sur les deux bases de données est que les distances relatives de la base LUCID sont globalement plus faibles que celles de la base Lombard-GRID, surtout pour les sons non-voisés où les performances sur la parole Lombard sont en moyennes moins bonnes qu'en l'absence de modifications. Cela s'explique par des modifications temporelles naturelles plus importantes dans la parole claire que dans la parole Lombard. Comme nous pouvons le voir dans le tableau 9.1 pour les moyennes globales, mais c'est aussi le cas pour les autres échelles (locuteur et phrase), les facteurs moyens de la parole Lombard sont proches de l'unité, et donc de l'absence de modifications, surtout pour les sons voisés. Ainsi, en ne prenant pas en compte la variance, l'utilisation de ces facteurs moyens engendre des modifications locales souvent inadaptées qui ont de grandes chances d'éloigner temporellement le signal traité du signal visé. Ce dysfonctionnement est moins problématique pour les sons voisés de la parole Lombard car les facteurs ont des variances faibles permettant à leur valeurs moyennes de modéliser plus correctement des modifications temporelles rapprochant légèrement le signal traité du signal visé. Bien que les facteurs de modifications temporelles de la parole claire ont une grande variance, leurs valeurs moyennes sont bien plus éloignés de l'unité, ce style de parole possède donc une marge de manoeuvre plus importante et est bien moins sensible aux erreurs de modifications temporelles locales du signal.

Concernant les nouvelles modélisations, on note une diminution importante de la distance relative attestant de leurs bonnes performances, même avec un lissage important des facteurs. Naturellement, la distance diminue avec le nombre de splines utilisées car on se rapproche des valeurs des coefficients bruts. À l'inverse des modifications classiques moyennes, les performances de ces modélisations lisses semblent maintenant quasi-équivalentes pour les deux bases de données. Cela s'explique par le fait que notre modélisation ne dépend pas de la valeur moyenne des facteurs bruts mais directement de leur distribution temporelle : les spécificités locales des styles sont captés par la trajectoire lisse du modèle. On note même des distances légèrement plus faibles pour les sons voisés de la parole Lombard dont les facteurs à la variance très faible sont parfaitement adapté à un lissage, même avec peu de splines.

Synthèse

Pour synthétiser les résultats, la nouvelle modélisation proposée semble effectivement capable de capter les variations locales des facteurs de modifications temporelles, tout en proposant une trajectoire plus naturelle et lisse que les facteurs bruts obtenus mathématiquement par DTW. Ces trajectoires sont aussi parfaitement adaptées au modèles d'apprentissage dans lesquels elles peuvent être intégrées parmi les autres caractéristiques pour de la conversion de parole. Le nombre de splines utilisées peut être vu comme un paramètre de variance qui dépendra du style de parole étudié.

Dans cette section nous avons travaillé sur des styles de parole qui nous intéressent dans le cadre de notre étude sur le renforcement de la parole. Cependant, cette nouvelle modélisation lisse des facteurs de modifications temporelles peut être exploitée dans tous les domaines de la conversion de voix, comme la conversion d'identité ou la conversion d'émotions. Cette contribution, proposant une nouvelle modélisation des modifications temporelles extraites à partir de l'alignement entre les signaux et son intégration dans des modèles d'apprentissage pour de la conversion de voix, a fait l'objet d'un dépôt de brevet d'invention intitulé "Système de conversion de la parole par apprentissage statistique avec modélisation continue des modifications temporelles." soumis le 27/01/2020.

En revanche, actuellement rien ne nous indique encore que cette modélisation permette une meilleure conversion du style d'un point de vue perceptif. En effet, les modifications classiques moyennes ont déjà montré qu'elles permettaient d'atteindre un rendu perceptivement proche du style visé, et la modélisation proposée pour les modifications temporelles, censée être plus naturelle, pourrait ne pas être perceptivement plus performante.

En tout cas, cette étude préliminaire nous a permis de confirmer que les modifications classiques moyennes était objectivement très peu performantes pour aligner deux signaux de style différents et qu'elles ne captent pas la diversité importante des modifications temporelles locales en fonction du contexte des phonèmes. Des tests perceptifs mesurant les performances de la nouvelle modélisation à obtenir un rendu naturelle et proche des styles visés est alors prévue à l'avenir.

Conclusion du chapitre 9

Dans ce chapitre, nous avons proposé de multiples adaptations, dans les approches de renforcement par conversion du style de parole, permettant de mieux prendre en compte les aspects temporels lors de l'apprentissage. L'exploitation des RNN spécialisés dans le traitement des séquences et l'utilisation d'une transformée en ondelette pour obtenir une représentation de certaines caractéristiques sur différentes échelles temporelles, représentent une étape importante pour la prise en compte du contexte global des phonèmes lors de leur traitement.

Enfin, conscients du traitement trop primaire des modifications temporelles des systèmes de conversion de parole, nous avons proposé une nouvelle modélisation lisse et continue des modifications temporelles directement intégrable dans les fonctions de conversion. Une première analyse objective montre que la modélisation proposée représente les modifications temporelles plus naturellement et plus précisément que les approches moyennées actuelles.

Dans le chapitre suivant, nous mettons en place un système de conversion du style de parole neutre vers Lombard et étudions l'intégration de toutes ces propositions de changement sur les performances d'apprentissage du système.

Chapitre 10

Conversion du style de parole neutre vers parole Lombard avec améliorations du traitement des aspects temporels

Sommaire

Retour à la table des matières

Introduction du chapitre 10

Nous avons vu chapitre 8 que le renforcement paramétrique par conversion du style de parole avait beaucoup de potentiel mais que, pour l'instant, malgré des résultats perceptifs de similarité intéressants, le gain d'intelligibilité engendré n'est pas significatif. La raison principale évoquée porte sur les dégradations introduites par le vocodeur mais nous pensons que la simplicité de traitement des aspects temporels y participe grandement aussi. Ainsi, nous avons proposé chapitre 9 de multiples adaptations et nouveautés permettant de mieux prendre en compte le contexte temporel global des phonèmes traités. Cela passe par l'exploitation d'architectures de réseaux de neurones spécialisées dans le traitement des séquences temporelles longues, l'utilisation d'une transformée en ondelettes pour obtenir une représentation de certaines caractéristiques sur différentes échelles temporelles et la création d'une nouvelle modélisation des modifications temporelles directement intégrable dans les fonctions de conversion.

Dans ce chapitre, un système de conversion du style de parole neutre vers parole Lombard est mis en place et nous étudions l'intégration de toutes ces propositions de changement sur les performances d'apprentissage du système. Nous présenterons d'abord, dans la section 10.1, les caractéristiques de la parole converties par le système, ainsi que les fonctions de conversion utilisées. Puis, une évaluation objective de l'apprentissage de la tâche de conversion sera menée dans la section 10.2.

10.1 Caractéristiques et fonctions de conversion

10.1.1 Extraction des caractéristiques

Soit un ensemble de données parallèles de M paires de signaux source/cible. Nous utilisons le vocodeur STRAIGHT [94] permettant d'extraire le spectrogramme et la fréquence fondamentale sur des fenêtres espacées de 5 ms. Notre but étant de transformer la prosodie des signaux de parole, nous utilisons les caractéristiques classiques de conversion de parole à savoir les évolutions de la fréquence fondamentale, de l'énergie et des coefficients MFCC. Afin de manipuler des grandeurs perceptivement pertinentes, les valeurs de la fréquence fondamentale et de l'énergie sont projetées sur des échelles logarithmiques. Pour une bonne lecture et interprétation des résultats, nous avons choisi l'échelle des demi-tons (dt) pour la fréquence fondamentale, $\mathrm{dt} = 39{,}87\log(\mathrm{F}/50)$, et l'échelle des décibels pour l'énergie. Pour chaque paire de signaux alignés (x_m, y_m), les valeurs de la fréquence fondamentale en demi-tons sur chacune des N_m fenêtres, notées $(\mathbf{f0}_{[x_m]}, \mathbf{f0}_{[y_m]}) \in (\mathbb{R}_+^{\mathrm{N}_m \times 1})^2$ sont obtenues directement par l'analyse. Les valeurs de l'énergie instantanée de chaque fenêtre, notées $(\mathbf{e}_{[x_m]}, \mathbf{e}_{[y_m]}) \in (\mathbb{R}_+^{\mathrm{N}_m \times 1})^2$, sont calculés à partir des spectrogrammes, notés $(\left|\mathbf{S}_{[x_m]}\right|^2, \left|\mathbf{S}_{[y_m]}\right|^2) \in (\mathbb{R}_+^{\mathrm{N}_m \times \mathrm{F}})^2$, avec $\mathrm{F} = 1024$, de la façon suivante :

$$\forall n \in [[1, \mathrm{N}_m]], e(n) = 20\log\left(\sqrt{\sum_{f=0}^{\mathrm{F}-1}\left|\mathrm{S}(n,f)\right|^2}\right). \tag{10.1}$$

Les coefficients MFCC sont calculés classiquement en appliquant un banc de (Q+1) filtres triangulaires, espacés selon l'échelle de Mel, sur le spectre de puissance puis en prenant la transformée en cosinus discrète (TCD) du logarithme du mel-spectre :

$$|\mathbf{S}|^2 \in \mathbb{R}_+^{\mathrm{N}_m \times \mathrm{F}} \xrightarrow{\text{mel scale}} |\tilde{\mathbf{S}}|^2 \in \mathbb{R}_+^{\mathrm{N}_m \times \mathrm{Q}+1} \xrightarrow{\log + \mathrm{DCT}} \mathbf{mfcc} \in \mathbb{R}^{\mathrm{N}_m \times \mathrm{Q}+1}. \tag{10.2}$$

Nous choisissons classiquement $\mathrm{Q} = 24$ et ne retenons pas le premier coefficient qui est directement associé à l'énergie de la fenêtre, on note alors les coefficients utilisés $(\mathbf{mc}_{[x_m]}, \mathbf{mc}_{[y_m]}) \in (\mathbb{R}^{\mathrm{N}_m \times \mathrm{Q}})^2$. Notons tout de même que le nombre de filtres peut être différent du nombre de coefficients mel-cepstraux utilisés. Pour la reconstruction du spectre, nous utilisons simplement la TCD inverse pour récupérer le spectre de puissance à partir du mel-cepstre.

10.1.2 Transformée en ondelettes continue

Comme introduit section 9.1.2, l'utilisation d'une transformée en ondelettes continue, ou CWT, afin de représenter les caractéristiques sur différentes échelles temporelles peut améliorer grandement l'apprentissage de la dynamique des caractéristiques. Une CWT sur dix octaves peut alors être appliquée sur le fondamental et/ou l'énergie avec une ondelette mère de Ricker de taille $s_0 = 2$ couvrant alors une durée initiale de 10 ms. Les nouvelles caractéristiques résultantes, notées $\mathbf{x}^{cwt} \in \mathbb{R}^{\mathrm{N}_m \times 10}$, sont données par l'équation 9.20. Un exemple d'une trajectoire de fondamental et de ses composantes CWT associées avec ce paramétrage avait déjà été présenté section 9.1.2 sur la figure 9.4. Pour la phase de conversion, la reconstruction est effectuée en utilisant l'équation 9.21 avec $\psi_0(0) = 0,8673$ l'amplitude à l'origine de l'ondelette de Ricker et $\mathrm{R} = 3,541$ son facteur de reconstruction donné.

10.1.3 Caractéristiques dynamiques

Finalement, pour toute caractéristique de dimension L $\mathbf{x} \in \mathbb{R}^{\mathrm{N}_m \times \mathrm{L}}$, nous pouvons calculer ses caractéristiques *delta* notées $\mathbf{x}^{\delta} \in \mathbb{R}^{\mathrm{N}_m \times \mathrm{L}}$, et ses caractéristiques *delta-delta*, comme détaillé section 8.2.2. Pour la reconstruction à partir de ces caractéristiques dynamiques, l'algorithme MLPG [205] est utilisé afin d'estimer des trajectoires cohérentes :

$$\hat{\mathbf{x}} = \mathrm{MLPG}(\mathbf{x}, \mathbf{x}^{\delta}, \mathbf{x}^{\delta^2}) . \tag{10.3}$$

L'utilisation des caractéristiques *delta*, et *delta-delta*, est primordiale pour les modèles qui ne prennent pas en compte la dépendance temporelle entre les échantillons des signaux, comme pour le modèle GMM ou le modèle FFNN. En revanche, l'utilisation de ces pré-traitements pour les modèles RNN est plus discutable car le principe même de leur architecture est de prendre en compte cette dépendance. Des tests préliminaires nous ont montré que l'utilisation des caractéristiques *delta*, et *delta-delta*, pour toutes les caractéristiques améliorait systématiquement l'apprentissage, pour tous les modèles, et ce malgré la multiplication par trois du nombre de dimensions des caractéristiques.

10.1.4 Entrées/sorties des fonctions de conversion

Les caractéristiques d'entrées $\mathbf{X}_m$ et de sorties $\mathbf{Y}_m$ utilisées pour entraîner la fonction de conversion sont alors regroupées en concaténant les caractéristiques avec les traitements choisis. Pour les entrées si, par exemple, nous utilisons les coefficients de CWT du fondamental et de l'énergie, les caractéristiques regroupées s'expriment alors :

$$\mathbf{X}_m = [\mathbf{f0}^{cwt}_{[x_m]}, (\mathbf{f0}^{cwt}_{[x_m]})^{\delta}, (\mathbf{f0}^{cwt}_{[x_m]})^{\delta^2}, \mathbf{e}^{cwt}_{[x_m]}, (\mathbf{e}^{cwt}_{[x_m]})^{\delta}, (\mathbf{e}^{cwt}_{[x_m]})^{\delta^2}, \mathbf{mc}_{[x_m]}, (\mathbf{mc}_{[x_m]})^{\delta}, (\mathbf{mc}_{[x_m]})^{\delta^2}] \in \mathbb{R}^{\mathrm{N}_m \times \mathrm{C}_x} , \tag{10.4}$$

avec $\mathrm{C}_x = (3 * 10) + (3 * 10) + (3 * 24) = 132$. Et pour les sorties si, par exemple, nous utilisons seulement les coefficients de CWT du fondamental, elles s'expriment alors :

$$\mathbf{Y}_m = [\mathbf{f0}^{cwt}_{[y_m]}, (\mathbf{f0}^{cwt}_{[y_m]})^{\delta}, (\mathbf{f0}^{cwt}_{[y_m]})^{\delta^2}, \mathbf{e}_{[y_m]}, (\mathbf{e}_{[y_m]})^{\delta}, (\mathbf{e}_{[y_m]})^{\delta^2}, \mathbf{mc}_{[y_m]}, (\mathbf{mc}_{[y_m]})^{\delta}, (\mathbf{mc}_{[y_m]})^{\delta^2}] \in \mathbb{R}^{\mathrm{N}_m \times \mathrm{C}_y} , \tag{10.5}$$

avec $\mathrm{C}_y = (3 * 10) + (3 * 1) + (3 * 24) = 105$. La plupart du temps les caractéristiques d'entrée et de sortie utilisées seront identiques mais ce n'est pas forcément le cas comme nous le verrons par la suite.

10.1.5 Fonctions de conversion

La fonction de conversion a une importance majeure vis-à-vis des performances du système. Nous proposons alors d'étudier l'influence de différents types de fonctions pour la conversion de parole neutre à la parole Lombard. Comme référence nous utilisons des fonctions de conversion classiques déjà utilisées pour cette tâche dans des études antérieures à savoir le modèle GMM, présenté dans la section 8.2.2, et plus récemment le modèle FFNN, présenté dans la section 8.2.3. Nous proposons aussi l'utilisation d'architectures de réseaux de neurones

(ANN) plus adaptées au traitement de séquences temporelles à savoir les modèles RNN et leurs variantes. Plus précisément nous étudions les performances des modèles RNN classiques complètement récurrents, usuellement abrégé par l'acronyme anglais pour *Fully Recurrent* (FR), de l'utilisation de cellules GRU et de cellules LSTM, ces notions ont été présentées dans la section 9.1.1. Nous étudions aussi l'influence de la directionnalité des modèles RNN qui, pour rappel, sont unidirectionnels (UD), s'ils prédisent chaque échantillon de la séquence en se basant sur les échantillons passés, ou bidirectionnels (BD), s'ils les prédisent en en se basant sur les échantillons passés et futurs.

Le choix des hyper-paramètres est crucial vis-à-vis des performances du modèle d'apprentissage. Pour le modèle GMM, nous utilisons la variante VBGMM avec un nombre important de composantes dans le mélange (200) afin d'exploiter leur capacité naturelle à choisir automatiquement le nombre de composante adapté. Pour les ANN, les hyper-paramètres sont choisis précautionneusement en utilisant une approche appelée *Hyperband* [110]. C'est une méthode adaptative récente de recherche aléatoire qui est parvenue à accélérer grandement la recherche d'hyper-paramètres pour une variété de problème d'apprentissage profond.

Les ANN entraînés utilisent l'algorithme d'optimisation Adam et les fonctions d'activation des unités cachés du modèle FFNN sont des ReLU. Des tests préliminaires nous ont alors permis de choisir des intervalles pour les distributions de probabilités utilisées pour la recherche d'hyper-paramètres par *Hyperband*, tout cela est visible dans le tableau 10.1.

			FFNN		RNN	
hyper-paramètre	type	distribution	min	max	min	max
nombre de couches cachées	entier	uniforme	2	6	2	4
nombre d'unités cachées	entier	log-uniforme	64	512	32	512
taux d'abandon	réel	uniforme	0	0,4	0	0,4
taille des mini-lots	entier	log-uniforme	128	1024	1	32
taux d'apprentissage initial	réel	log-uniforme	10^{-5}	10^{-3}	10^{-5}	10^{-3}
taux d'affaiblissement des poids	réel	log-uniform	10^{-9}	10^{-5}	10^{-9}	10^{-5}

TABLEAU 10.1 – Distributions de probabilité des hyper-paramètres pour les ANN, obtenues par des tests préliminaires, utilisées par l'algorithme de recherche *Hyperband*.

10.2 Évaluation objective

Le test est effectué sur la base de données "Lombard GRID" [6] introduite section 3.1. Le corpus est divisé en un ensemble d'entraînement (80%), un ensemble de développement (10%) et un ensemble de test (10%) où chaque locuteur apparaît proportionnellement. L'ensemble d'entraînement rassemble les données utilisées pour l'optimisation des poids des modèles. L'ensemble de développement est utilisé par *Hyperband* afin de sélectionner les hyper-paramètres pour chaque modèle. Finalement, l'ensemble de test ne sert qu'une seule fois pour mesurer les performances finales de chaque modèle. L'objectif étant de mesurer les performances d'apprentissage du système, et que chaque locuteur peut utiliser une stratégie Lombard différente, il est important d'avoir les mêmes locuteurs dans les différents ensembles. Pour une évaluation subjective, où l'on chercherait plutôt à évaluer les propriétés de généralisation du système, des locuteurs différents devraient alors être placés dans l'ensemble d'entraînement et celui de test.

Enfin, les caractéristiques sont standardisées (centrées, réduites) non pas globalement mais fonction du locuteur afin de supprimer leurs attributs individuels et de conserver uniquement ceux du style de parole [175]. En effet, une standardisation moyenne conserverait les différences inter-locuteur de ces statistiques et complexifierait la représentation des données dans l'espace des caractéristiques, pénalisant alors l'apprentissage des modifications spécifiques au style de parole. Par contre, l'apprentissage se faisant sur des caractéristiques standardisées, le modèle n'apprend pas les changements des statistiques globales. Ainsi les modifications des moyennes et variances des caractéristiques sont inférées par interpolation linéaire sur les valeurs de la base d'entraînement.

10.2.1 Mesures de performances

Une fois les modèles entraînés sur les ensembles d'entraînement et de développement, les caractéristiques des signaux de l'ensemble de test sont converties puis comparées aux caractéristiques Lombard naturelles en calculant classiquement la *Root Mean Squared Error* (RMSE) pour le fondamental (dt) et l'énergie (dB), et la MCD moyenne pour les coefficients MFCC (dB) afin de mesurer la distorsion spectrale :

$$\mathrm{MCD}(\hat{\mathbf{mc}}_i, \mathbf{mc}_i) = \frac{10}{\log 10}\sqrt{\sum_{m=1}^{24}(\hat{\mathbf{mc}}_{m,i} - \mathbf{mc}_{m,i})^2}\,. \tag{10.6}$$

Cependant, nous avons vu juste précédemment que les modèles apprennent et agissent sur les caractéristiques standardisées, puis les statistiques globales (moyennes et variances) sont inférées par interpolation. Les calculs de la RMSE du fondamental et de l'énergie étant très sensibles aux écarts des statistiques globales, il sera difficile d'analyser les performances de l'apprentissage uniquement sur ces mesures. Ainsi, pour ces deux caractéristiques nous calculons aussi le coefficient de corrélation r, entre la caractéristique convertie $\hat{\mathbf{y}} \in \mathbb{R}^{\mathrm{N}_m \times 1}$ et la caractéristique correspondante $\mathbf{y} \in \mathbb{R}^{\mathrm{N}_m \times 1}$ du Lombard naturel :

$$r(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\mathrm{Cov}(\hat{\mathbf{y}}, \mathbf{y})}{\sigma_{\hat{\mathbf{y}}} \cdot \sigma_{\mathbf{y}}} = \frac{\sum_{n=1}^{\mathrm{N}_m}(\hat{y}(n) - \overline{\hat{y}})(y(n) - \overline{y})}{\sqrt{\sum_{n=1}^{\mathrm{N}_m}(\hat{y}(n) - \overline{\hat{y}})^2 \sum_i (y(n) - \overline{y})^2}}\,, \tag{10.7}$$

avec $\overline{y} = \frac{1}{\mathrm{N}_m}\sum_{n=1}^{\mathrm{N}_m} y(n)$ et $\overline{\hat{y}} = \frac{1}{\mathrm{N}_m}\sum_{n=1}^{\mathrm{N}_m} \hat{y}(n)$.

En appliquant ces calculs aux caractéristique de la parole neutre nous obtenons des valeurs de référence correspondant à une absence de modification. Toutes les configurations ont été entraînées en utilisant *Hyperband* et les résultats de performances obtenues sont affichées dans le tableau 10.2 pour :

- un apprentissage sans CWT,
- un apprentissage avec les coefficients de CWT du fondamental et de l'énergie,
- un apprentissage avec les coefficients de CWT du fondamental seulement.

Les résultats présentés ici sont légèrement différents de ceux de notre publication [65] car les nouveaux calculs ont été effectués seulement sur les fenêtres d'intérêts en fonction des caractéristiques. Ainsi les mesures objectives pour le fondamental ont été prises seulement sur les fenêtres voisés et celles pour l'énergie et les coefficients MFCC ont été prises sur les fenêtres de parole i.e. sans les silences.

10.2.2 Analyse des résultats de l'apprentissage

Apprentissage sans CWT

En comparant les résultats obtenus avec toutes les caractéristiques simples, sans CWT, nous observons, dans le tableau 10.2, que tous les modèles de conversion ont efficacement appris les transformations et toutes les améliorations des mesures par rapport à la référence sont statistiquement hautement significatives ($p << 10^{-3}$). En revanche, les performances entre les modèles sont plus mitigées.

- Concernant les deux mesures du fondamental, pas de différences significatives ne sont relevées entre les modèles ($p > 0,1$), à part pour le LSTM BD qui obtient un score de corrélation significativement supérieur au modèle GMM ($p = 0,05$).
- Concernant l'énergie, tous les ANN obtiennent des performances statistiquement supérieures au modèle GMM pour les deux mesures ($p < 10^{-3}$). Parmi les ANN, les performances sur la corrélation de l'énergie sont statistiquement équivalentes ($p > 0,1$) mais l'intérêt des architectures récurrentes commence à se

modèle	fondamental		énergie		MFCC
	RMSE (dt)	corr. (%)	RMSE (dB)	corr. (%)	MCD (dB)
référence	3,96	54,7	7,75	86,8	6,00
	simple		simple		simple
GMM	2,18	66,3	4,63	84,5	4,73
FFNN	2,11	67,8	4,10	89,5	4,66
FR UD	2,10	67,8	4,00	89,6	4,69
FR BD	2,10	68,1	3,91	89,8	4,62
GRU UD	2,14	68,5	3,94	89,7	4,62
GRU BD	2,14	69,0	**3,83**	**90,1**	4,59
LSTM UD	2,13	67,7	4,21	88,9	4,62
LSTM BD	**2,08**	**70,1**	4,11	89,3	**4,53**
	CWT		CWT		simple
GMM	2,33	62,3	4,70	85,7	5,15
FFNN	2,05	70,1	3,83	89,4	4,80
FR UD	2,06	69,8	3,82	89,6	4,75
FR BD	2,04	70,6	3,82	89,9	4,68
GRU UD	2,08	71,1	3,85	89,5	4,73
GRU BD	2,07	**72,5**	**3,79**	**90,2**	4,62
LSTM UD	2,06	71,5	3,85	89,5	4,75
LSTM BD	**2,02**	72,3	3,90	89,6	**4,61**
	CWT		simple		simple
FFNN	2,02	70,4	3,92	89,7	4,73
GRU BD	**1,98**	**73,5**	**3,44**	**90,7**	**4,56**
LSTM UD	2,00	72,2	4,31	89,0	4,70
LSTM BD	2,01	73,3	4,10	89,4	4,59

TABLEAU 10.2 – Synthèse des performances moyennes pour chaque caractéristique obtenue.

manifester sur la RMSE, notamment avec l'utilisation des cellules GRU. On remarque que, dans l'ordre, FR UD, FR BD, GRU UD, puis GRU BD, présentent des RMSE de plus en plus faibles atteignant une amélioration significative de -0,27 dB ($p < 10^{-3}$) par rapport au modèle FFNN. L'utilisation des cellules LSTM quant à elle présente des performances bien moins intéressantes sur l'énergie, statistiquement équivalentes au modèle FFNN ($p > 0,1$) mais significativement moins bonnes que les autres architectures récurrentes ($p < 0,1$).

- Concernant les coefficients MFCC, tous les modèles ANN obtiennent des performances significativement très supérieures au modèle GMM ($p < 10^{-3}$) à part pour le modèle FFNN pour qui elles sont moins significatives ($p = 0,04$) et le modèle FR UD pour qui elles ne le sont plus du tout ($p = 0,26$). On note aussi que la version BD de chaque modèle RNN surpasse systématiquement sa version UD, ce qui était déjà le cas sur les autres caractéristiques mais cette différence est cette fois statistiquement significative pour le modèle FR ($p = 0,03$) et le modèle LSTM ($p = 0,02$) bien que toujours non significative pour le modèle GRU ($p = 0,45$).

Ces premières observations mettent en lumière certains points intéressants :

- le modèle GMM obtient de très bonnes performances, notamment pour le fondamental, mais l'utilisation de modèles ANN semble tout de même plus adaptée en proposant des performances globales significativement supérieures,
- parmi les modèles ANN, l'utilisation d'architectures récurrentes proposée montre des résultats objectifs légèrement plus performants que l'utilisation classique d'un modèle FFNN, et plus particulièrement avec l'utilisation de cellules GRU/LSTM,
- la tendance logique qu'ont les versions BD des RNN à surpasser leurs versions UD est observable mais rarement significative, cela peut s'expliquer par le fait qu'un RNN BD utilise deux fois plus de paramètres, ce qui demanderait alors plus de données pour exploiter pleinement l'intérêt de ces variantes et obtenir, potentiellement, des améliorations plus significatives,
- le fait que le LSTM soit objectivement le modèle le plus performant concernant la conversion du fondamental et des coefficients MFCC, et plus du tout concernant l'énergie, peut aussi s'expliquer par le nombre de paramètres important dans les cellules LSTM qui rendent l'optimisation plus difficile favorisant certaines caractéristiques et débouchant alors sur des performances variables.

Apprentissage avec CWT

Intéressons nous maintenant aux résultats obtenus avec l'utilisation de la CWT consistant à remplacer les valeurs simples uni-dimensionnels du fondamental et l'énergie par leur 10 coefficients de CWT. En comparant les résultats obtenus uniquement avec la CWT, toujours consultables tableau 10.2, nous observons, encore une fois, que tous les modèles de conversion ont efficacement appris les transformations des caractéristiques et toutes les améliorations des mesures par rapport à la référence sont toujours statistiquement hautement significatives ($p << 10^{-3}$). En revanche, les performances du modèle GMM vis-à-vis du fondamental sont significativement moins intéressantes que celles des modèles ANN ($p < 10^{-2}$), alors qu'elles étaient équivalentes sans CWT. À part ce détail sur lequel nous reviendrons ci-après, toutes les conclusions faites lors de la comparaison des résultats obtenus sans la CWT entre les différents modèles de conversion s'appliquent ici.

Le tableau 10.3 donne les statistiques des différences de moyenne sur toutes les mesures entre l'utilisation des valeurs simples et l'utilisation des coefficients CWT sur les deux caractéristiques pour chacun des modèles. Ces statistiques ont été obtenues par des tests-t appariés avec la correction de Bonferroni. Cela semble améliorer globalement les performances de tous les modèles sur ces deux caractéristiques, à part pour le modèle GMM qui ne parvient pas du tout à exploiter cette nouvelle représentation et obtient des performances dégradées sur le fondamental. En revanche, les performances sur les coefficients MFCC diminuent légèrement pour tous les modèles, ce qui peut s'expliquer par l'augmentation du nombre de dimensions des autres caractéristiques rendant l'apprentissage des coefficients MFCC, avec la même quantité de données, plus complexe. Pour chaque modèle on trouve les résultats détaillés suivants :

- Le modèle GMM perd effectivement grandement en performance avec une augmentation significative de la RMSE de 0,15 dB pour le fondamental ($p = 0,08$), une diminution significative de sa corrélation de 4,06 p.p. ($p = 0,04$) et une augmentation significative de la MCD de 0,42 dB ($p < 10^{-2}$ pour les coefficients MFCC. L'énergie est moins impactée avec, au contraire, une RMSE équivalente et une légère augmentation significative de la corrélation de 1,2 p.p. ($p = 0,02$).
- Les modèles FFNN, FR UD, FR BD, GRU UD et LSTM BD, obtiennent des améliorations globalement similaires. Une légère amélioration de la RMSE pour l'énergie, ainsi qu'une légère dégradation de la MCD pour les coefficients MFCC, pas toujours significatives. Concernant le fondamental, les améliorations sont très légères ($\approx 0,05$ dt) et non significatives pour la RMSE ($p > 0,1$) alors qu'elles sont assez importantes pour la corrélation (entre 2,0 et 2,7 p.p.) mais toujours non significatives à cause d'une variance trop importante.
- Les modèles GRU BD, resp. LSTM UD, présentent des améliorations quasi-similaires si ce n'est que cette fois la corrélation obtient un gain encore plus important de 3,54 p.p., resp. 3,78 p.p., et cette amélioration est cette fois significative ($p = 0,05$ resp. $p = 0,04$).

Ainsi, la CWT procure globalement un meilleur apprentissage de la modification de la trajectoire du fondamental mais l'ajout de nombreuses dimensions supplémentaires complexifie l'apprentissage, en augmentant le nombre de paramètres à optimiser, et ne convient pas à tous les modèles, surtout au modèle GMM. Pour rappel, les cellules GRU possèdent une architecture très similaire aux cellules LSTM mais avec moins de paramètres et nous avons vu que cette différence permettait déjà au modèle GRU d'être plus stable lors de l'apprentissage sans CWT. Avec l'augmentation du nombre de dimension, l'écart de performance entre ces deux modèles s'amenuise et, bien que ce ne soit pas significatif, le modèle GRU semble mieux maîtriser cette nouvelle représentation et obtient des performances équivalentes (UD), voir légèrement supérieures (BD). Concernant l'énergie, l'utilisation de la CWT tend à améliorer légèrement la conversion de cette caractéristique mais obtient des gains de performance très faibles et rarement significatifs. À partir de ces observations, nous pouvons conclure que l'énergie ne profite pas autant que le fondamental de l'utilisation de la CWT et va, au contraire, complexifier l'optimisation pour les autres caractéristiques en augmentant le nombre de paramètres.

Apprentissage avec CWT seulement sur le fondamental

C'est à partir de ces dernières observations que nous avons proposé d'entraîner les modèles qui semblent le plus profiter de l'utilisation des coefficients de la CWT sur le fondamental, mais en conservant les valeurs simples de l'énergie. Ces modèles sont le GRU BD, le LSTM UD et le LSTM BD, nous avons aussi entraîné un modèle FFNN en guise de comparaison. Les résultats, toujours consultables dans le tableau 10.2 pour les performances et tableau 10.3 pour leurs améliorations, confirment alors bien nos suppositions.

- Pour le modèle FFNN, pas de changements comparé à l'utilisation de la CWT sur les deux caractéristiques avec toujours une légère amélioration de la RMSE de l'énergie de -0,18 dB ($p = 0,02$), une légère dégradation de la MCD de 0,07 dB ($p = 0,07$) et pas de différences significatives sur le reste des performances.
- Le LSTM BD profite de ce changement de façon mitigée avec une nette amélioration de la corrélation du fondamental de 3,20 p.p. qui devient significative ($p = 0,07$) et une stabilisation de la MCD, cependant le gain sur la RMSE de l'énergie n'est plus significatif.
- Les modèles GRU BD, resp. LSTM UD, profitent pleinement de ce changement avec une nette amélioration de la corrélation du fondamental de 4,47 p.p. qui est encore plus significative ($p = 0,01$). On observe aussi une stabilisation de la MCD, bien qu'elle soit toujours légèrement dégradée de 0,08 dB pour le LSTM UD ($p = 0,08$). Finalement, un inversement s'opère sur la RMSE de l'énergie avec une amélioration qui devient significative pour le GRU BD, avec -0,39 dB ($p < 0,01$), et qui ne l'est plus pour le LSTM, avec -0,01 dB ($p = 0,90$).

Ainsi, l'utilisation des coefficients CWT seulement sur le fondamental permet d'améliorer significativement l'apprentissage de la trajectoire de celui-ci pour les RNN. En revanche, les modèles LSTM présentent toujours les

modèle	fondamental				énergie				MFCC	
	RMSE		corr.		RMSE		corr.		MCD	
	$\overline{\text{diff.}}$(dt)	p	$\overline{\text{diff.}}$(p.p.)	p	$\overline{\text{diff.}}$(dB)	p	$\overline{\text{diff.}}$(p.p.)	p	$\overline{\text{diff.}}$(dB)	p
	simple → CWT				simple → CWT				simple → simple	
GMM	0,15	**0,08**	-4,06	**0,04**	0,07	0,53	1,20	**0,02**	0,42	$<$ **10^{-2}**
FFNN	-0,06	0,45	2,28	0,23	-0,28	$<$ **10^{-2}**	-0,10	0,82	0,14	$<$ **10^{-2}**
FR UD	-0,03	0,69	2,04	0,29	-0,18	**0,02**	0,05	0,90	0,06	**0,10**
FR BD	-0,07	0,41	2,50	0,19	-0,09	0,26	0,08	0,85	0,06	0,10
GRU UD	-0,05	0,53	2,64	0,16	-0,09	0,25	-0,13	0,77	0,10	$<$ **10^{-2}**
GRU BD	-0,07	0,42	3,54	**0,05**	-0,05	0,57	0,09	0,83	0,02	0,53
LSTM UD	-0,07	0,37	3,78	**0,04**	-0,35	$<$ **10^{-2}**	0,57	0,17	0,13	$<$ **10^{-2}**
LSTM BD	-0,06	0,43	2,26	0,20	-0,20	**0,01**	0,34	0,43	0,08	**0,03**
	simple → CWT				simple → simple				simple → simple	
FFNN	-0,10	0,25	2,63	0,17	-0,18	**0,02**	0,16	0,70	0,07	**0,07**
GRU BD	-0,15	**0,06**	4,47	**0,01**	-0,39	$<$ **0,01**	0,62	0,12	-0,03	0,35
LSTM UD	-0,14	**0,09**	4,47	**0,01**	0,11	0,16	0,04	0,90	0,08	**0,02**
LSTM BD	-0,07	0,37	3,20	**0,07**	-0,01	0,90	0,08	0,86	0,06	0,12

TABLEAU 10.3 – Synthèse des statistiques liées à l'amélioration des performances de toutes les caractéristiques par l'utilisation des coefficients CWT sur le fondamental et/ou l'énergie. $\overline{\text{diff.}}$ correspond aux différences moyennes des performances et p aux p-$values$ associées.

mêmes défauts d'instabilité vis-à-vis des autres caractéristiques, probablement liés à leur grand nombre de paramètres à optimiser. Au contraire, le modèle GRU BD parvient à conserver des performances équivalentes, voir légèrement meilleures, pour les autres caractéristiques.

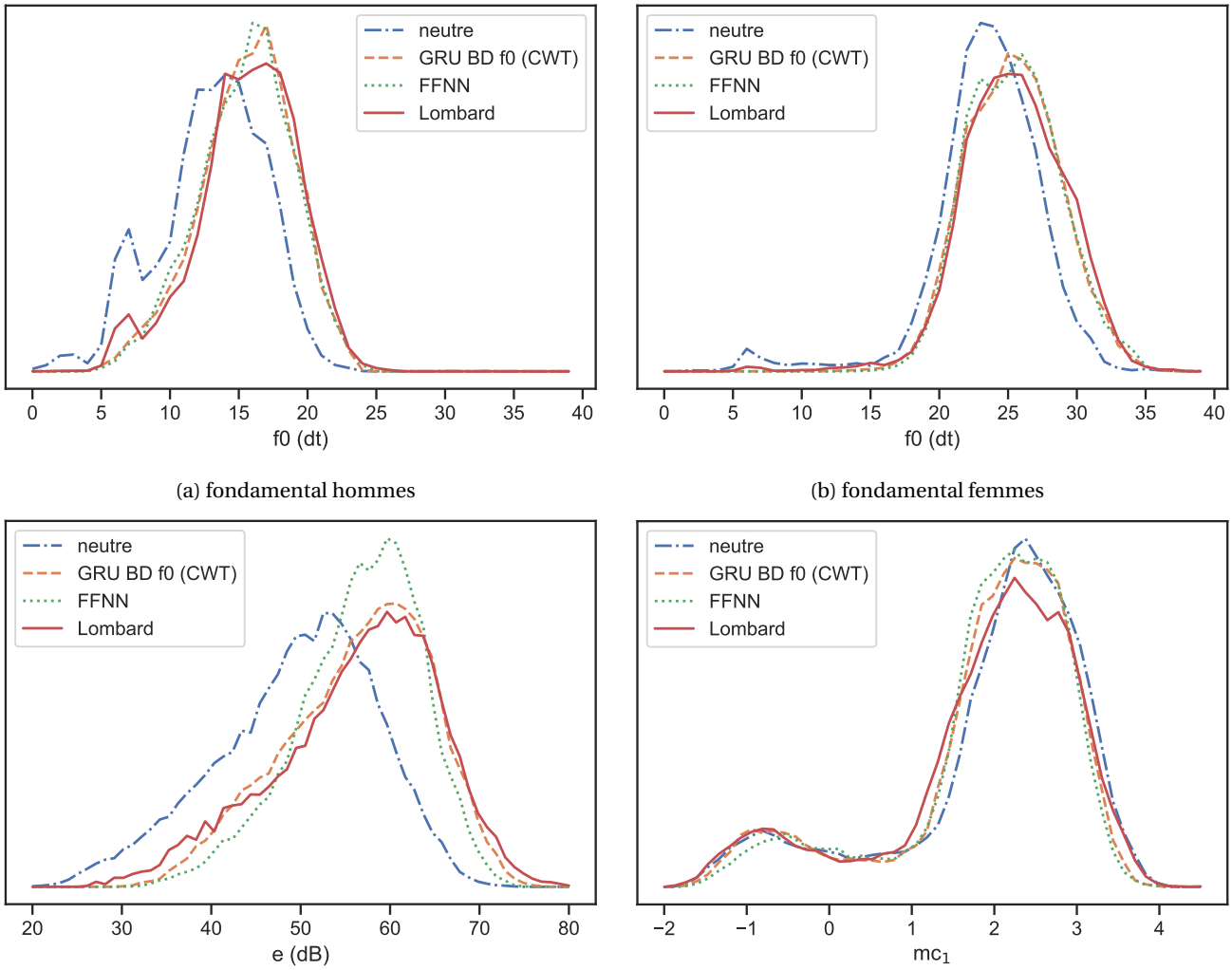
10.2.3 Distribution des caractéristiques converties

L'analyse précédente se concentrait sur les performances liées à l'apprentissage des transformations locales en comparant les caractéristiques converties fenêtre par fenêtre. Il est aussi intéressant d'observer comment se comportent les conversions d'un point de vue global en visualisant les distributions de certaines caractéristiques. La figure 10.1 montre les histogrammes de trois caractéristiques prosodiques principales, à savoir le fondamental, l'énergie instantanée et la pente spectrale ($\mathbf{mc}_1$), obtenus sur l'ensemble de test, pour la parole neutre, la parole Lombard naturelle et deux paroles neutres converties en Lombard. Les deux modèles choisis pour les paroles converties sont le FFNN sans CWT, la base de référence actuelle, et le GRU BD avec CWT sur le fondamental, notre modèle qui détient les meilleurs performances objectives d'apprentissage. Enfin, vu que la distribution du fondamental diffère grandement entre les hommes et les femmes, nous traitons les genres séparément. Nous pouvons clairement voir que pour les deux modèles, les distributions des caractéristiques converties tendent vers les distributions de la parole Lombard naturelle. Cela semble encore plus prononcé pour celles obtenues avec le modèle GRU BD, et surtout pour l'énergie avec une redistribution étalée très proche de celle de la parole Lombard pour ce dernier alors que le modèle FFNN a eu tendance à plus concentrer l'énergie autour de 60 dB.

Pour mieux quantifier ces observations, et comparer les distributions des autres modèles, nous calculons la divergence entre les distributions des caractéristiques converties P_c et celles des caractéristiques naturelles Lombard P_L. Le calcul est effectué à partir de la divergence Jenson-Shannon (JSD), une version symétrique de la divergence Kullback-Leibler (KLD) de la façon suivante :

$$\mathrm{JSD}(\mathrm{P}_c||\mathrm{P}_\mathrm{L}) = \frac{1}{2}\mathrm{KLD}((\mathrm{P}_c||\frac{\mathrm{P}_c+\mathrm{P}_\mathrm{L}}{2}) + \frac{1}{2}\mathrm{KLD}((\mathrm{P}_\mathrm{L}||\frac{\mathrm{P}_c+\mathrm{P}_\mathrm{L}}{2}). \tag{10.8}$$

Pour la nouvelle distribution d'une caractéristique convertie d'un modèle, plus sa divergence est faible, plus cette distribution est proche de celle de la parole Lombard naturelle. D'un point de vue perceptif, il est aussi



(a) fondamental hommes

(b) fondamental femmes

(c) énergie

(d) pente spectrale

FIGURE 10.1 – Histogrammes des distributions des caractéristiques prosodiques principales. Quatre distributions sont représentées sur chaque sous-figure avec une pour la parole neutre, une pour la parole Lombard naturelle et deux pour les paroles converties par les modèles : FFNN sans l'utilisation de la CWT et GRU BD avec l'utilisation des coefficients CWT uniquement du fondamental.

intéressant que cette nouvelle distribution s'éloigne de celle de la parole neutre naturelle P_n. C'est pourquoi nous calculons aussi la divergence nette d_{net} qui quantifie la distance relative entre une nouvelle distribution et celles des deux paroles naturelles, neutre et Lombard. Elle se calcule en prenant la différence entre la divergence par rapport à la parole neutre et la divergence par rapport à la parole Lombard :

$$d_{net} = \mathrm{JSD}(P_c||P_n) - \mathrm{JSD}(P_c||P_L)\,. \tag{10.9}$$

Plus la divergence nette est faible, pour la nouvelle distribution d'une caractéristique convertie d'un modèle, plus cette distribution est relativement proche de celle de la parole neutre naturelle et éloignée de celle de la parole Lombard naturelle. Au contraire, plus cette mesure est élevée, plus cette distribution est relativement proche de celle de la parole Lombard naturelle et éloignée de celle de la parole neutre. Les divergences et divergences nettes pour toutes les distributions des trois caractéristiques obtenus avec les différents modèles sont donnés dans le tableau 10.4.

On peut voir que les distributions obtenues par conversion avec les modèles sont toujours plus proches de la parole Lombard que de la parole neutre ($d_{net} > 0$). Dans la continuité, les distributions du fondamental et de l'énergie sont toujours très proches de celles de la parole Lombard ($\mathrm{JSD}(P_c||P_L) << \mathrm{JSD}(P_n||P_L)$), or ce n'est pas le cas pour la pente spectrale dont les divergences avec la parole Lombard avoisinent celles de la parole neutre ($\mathrm{JSD}(P_c||P_L) \approx \mathrm{JSD}(P_n||P_L)$). Cela peut s'expliquer par une proximité importante entre la distribution de la pente spectrale de la parole neutre naturelle et celle de la parole Lombard naturelle comme on peut le voir par la faible valeur de $\mathrm{JSD}(P_n||P_L) = 0,57$, et plus visuellement sur la figure 10.1d. Ainsi, bien que tous les modèles produisent des distributions plus proches de celle de la parole Lombard que de la parole neutre ($d_{net} > 0$), certains modèles ne parviennent pas à capter les tendances globales de la parole Lombard naturelle sur cette caractéristique. En revanche, les modèles RNN avec les cellules GRU et LSTM y parviennent et plus particulièrement les architectures BD. Notons aussi que le modèle GMM obtient de très bonnes performances sur le fondamental et surtout sur l'énergie avec la divergence simple la plus faible. La structure même du modèle GMM est, en effet, parfaitement adaptée pour convertir des distributions globales, mais ses performances médiocres sur les transformations locales, détaillées section 8.2.2, soulignent les limites de ce modèle pour de la conversion de parole.

modèle	fondamental				énergie		pente spectrale	
	hommes		femmes					
	JSD	d_{net}	JSD	d_{net}	JSD	d_{net}	JSD	d_{net}
neutre	6,35	−6,35	3,81	−3,81	7,88	−7,88	0,57	−0,57
	simple				simple		simple	
GMM	0,73	4,59	0,51	3,56	**1,20**	6,16	0,62	−0,01
FFNN	0,65	5,50	0,61	3,48	2,12	7,11	0,86	0,24
FR (UD)	0,99	5,06	0,75	3,11	1,79	7,64	0,92	0,19
FR (BD)	0,87	5,23	0,59	3,28	1,50	8,04	0,61	0,26
GRU (UD)	0,52	5,46	0,57	3,59	1,58	7,99	0,53	**0,30**
GRU (BD)	0,48	5,41	0,54	3,30	1,40	**8,34**	**0,40**	0,22
LSTM (UD)	**0,47**	5,50	0,55	**3,83**	2,26	7,90	0,61	0,22
LSTM (BD)	0,51	**5,52**	**0,49**	3,81	2,10	8,10	0,47	0,25
	CWT				simple		simple	
FFNN	0,87	5,45	0,66	3,03	1,60	7,48	1,04	0,12
GRU (BD)	0,53	5,57	0,45	3,28	**0,66**	7,65	0,48	**0,19**
LSTM (UD)	0,59	5,65	**0,40**	3,49	2,68	**7,73**	0,86	0,19
LSTM (BD)	**0,37**	**5,91**	0,40	**3,52**	2,19	7,68	**0,48**	0,17

TABLEAU 10.4 – Divergences et divergences nettes calculées pour toutes les distributions des trois caractéristiques obtenus avec les différents modèles. Toutes les valeurs ont été multipliées par un facteur 100 pour une meilleure lecture des résultats.

Pour ce qui est de l'utilisation de la CWT, nous pouvons toujours voir l'incapacité du modèle FFNN à exploiter les coefficients du fondamental avec lesquels il obtient une divergence simple plus élevée et une divergence nette plus faible. Pour les modèles RNN les effets sont logiquement moins visibles et plus mitigés à l'échelle globale qu'à l'échelle locale. On note cependant que la divergence simple et la divergence nette augmentent globalement pour les hommes, ce qui veut dire que la distribution s'éloigne de celle de la parole Lombard mais s'éloigne d'autant plus de la parole neutre. Au contraire, pour les femmes les deux divergences diminuent, ce qui veut dire que la distribution se rapproche de celle de la parole Lombard mais se rapproche d'autant plus de la parole neutre. Ces changements surprenants peuvent s'expliquer par la présence de plusieurs locuteurs aux stratégies Lombard différentes dans la base de données, l'espace sur lequel les caractéristiques sont converties est alors plus complexe qu'un simple chemin entre une parole neutre et une parole Lombard moyennées. On notera tout de même que le modèle GRU BD fait encore preuve d'une très grande stabilité et on observe même une diminution importante de la divergence simple pour l'énergie, faisant écho à l'amélioration de la RMSE vue section 10.2.2 dans ce cas de figure. Au contraire, les LSTM ont des résultats parfois très performants mais extrêmement variables entre les caractéristiques.

10.2.4 Introduction de la nouvelle modélisation des modifications temporelles

Une étape restante est l'introduction de la nouvelle modélisation des modifications temporelles proposée section 9.2.2 dans les modèles d'apprentissages. Pour cela nous avons calculé les trajectoires des facteurs de conversion pour chaque paire de phrases de la base donnée avec un nombre de 50 splines de lissage. Des tests préliminaires nous ont montré l'intérêt d'utiliser l'analyse CWT pour cette caractéristique, nous utilisons donc les coefficients comme caractéristique d'apprentissage. Les résultats objectifs précédents nous ont aussi permis de choisir les modèles d'apprentissage qui semblent les plus adaptés à savoir le GRU BD, le LSTM BD, ainsi que le FFNN toujours en guise de comparaison. Toute la procédure de préparation des données, de paramétrage et d'optimisation des modèles présentée dans la section précédente est identique. Le seul changement est l'ajout de la caractéristique du débit de parole dans les caractéristiques de sortie.

Les résultats de l'apprentissage sont consultables dans le tableau 10.5 au sein duquel nous faisons aussi apparaître les résultats de l'apprentissage sans les facteurs de modification temporelle afin d'étudier la potentielle influence que peut avoir cet ajout sur les autres caractéristiques. Nous observons d'ailleurs qu'il ne semble pas y en avoir : les performances pour les trois caractéristiques classiques (fondamental, énergie et coefficients MFCC) sont totalement équivalente à l'apprentissage sans les facteurs ($p > 0,1$). L'ajout des facteurs de modification temporelle n'influence donc pas l'apprentissage des autres caractéristiques et ce résultat nous montre aussi l'impressionnante stabilité que procure *Hyperband* dans le choix des hyper-paramètres. Concernant les performances d'apprentissage des facteurs de modification temporelle, pour la référence et les modèles entraînés sans facteurs nous utilisons des facteurs unitaires correspondant à une absence de modification temporelle. On voit alors que l'apprentissage des facteurs de la nouvelle modélisation a été efficace avec un gain notable très significatif ($p << 10^{-3}$) pour tous les modèles qui obtiennent un score équivalent ($p << 10^{-3}$).

Voyons maintenant ce que donne l'application des facteurs de modification temporelle estimés sur les signaux de parole neutre vis-à-vis de l'alignement temporel avec les signaux de parole Lombard. Pour cela nous utilisons la distance temporelle D_t, introduite section 9.2.3, afin de mesurer les performances de l'alignement avec les facteurs estimés en comparaison aux facteurs moyens obtenus à différentes échelles (globale, locuteur, phrase). Les graphiques des distances temporelles moyennes relatives (normalisés par la distance temporelle obtenue en l'absence de modifications), pour chaque type de voisement, sont consultables figure 10.2. Parmi les trois modèles d'apprentissage utilisés, les RNN sont équivalents et légèrement plus performants que le modèle FFNN. Pour les segments voisés, les facteurs issus des modèles RNN sont aussi légèrement plus performants que l'utilisation de facteurs moyens globaux, bien que ce ne soit pas statistiquement significatif. En revanche, pour les segments non-voisés, les facteurs estimés par les modèles d'apprentissage diminuent significativement la distance relative en comparaison à l'utilisation des facteurs moyens globaux, des facteurs moyens spécifiques au locuteur et même des facteurs moyens spécifiques à la phrase.

modèle	fondamental		énergie		MFCC	fact. temp.
	RMSE (dt)	corr. (%)	RMSE (dB)	corr. (%)	MCD (dB)	RMSE (s.u.)
référence	3,96	54,7	7,75	86,8	6,00	0,254
	CWT		simple		simple	aucun
FFNN	2,02	70,4	3,92	89,7	4,73	0,254
GRU BD	**1,98**	**73,5**	**3,44**	**90,7**	**4,56**	0,254
LSTM BD	2,01	73,3	4,10	89,4	4,59	0,254
	CWT		simple		simple	CWT
FFNN	2,00	70,3	3,97	89,6	4,73	0,202
GRU BD	**1,99**	**73,1**	**3,49**	**90,3**	**4,61**	**0,199**
LSTM BD	2,02	72,8	4,05	89,5	4,63	0,203

TABLEAU 10.5 – Synthèse des performances moyennes pour chaque caractéristique obtenue.

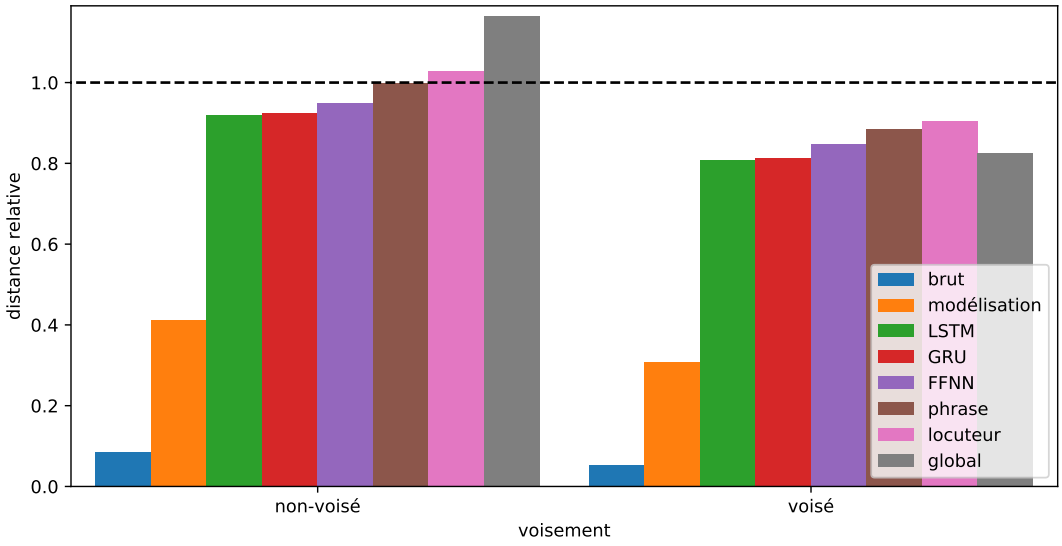


FIGURE 10.2 – Performances de la modélisation des modifications temporelles proposée .

Conclusion du chapitre 10

Dans ce chapitre, nous avons introduit les multiples adaptations des aspects temporels, proposées chapitre 9, dans un système de conversion de parole neutre vers parole Lombard.

L'utilisation d'un modèle RNN, et surtout de ses extensions LSTM et GRU, permet un apprentissage légèrement plus performant qu'avec un modèle FFNN sur toutes les caractéristiques acoustiques. De plus, la trajectoire du fondamental profite grandement de la représentation sur différentes échelles temporelles, par l'utilisation de la CWT, en obtenant une amélioration très significative de l'apprentissage sur cette caractéristique. Et pourtant, ces deux améliorations complexifient grandement l'apprentissage en introduisant de nombreux paramètres supplémentaires à optimiser. Pour exploiter pleinement le potentiel de ces approches, il faudrait donc travailler sur une base de données plus conséquente, comme SpeechDatCar introduite chapitre 3, sur laquelle on pourrait s'attendre à un apprentissage encore bien plus performant.

De plus, l'intégration de la nouvelle modélisation des modifications temporelles directement dans les fonctions de conversion permet d'obtenir des résultats aussi performants, pour les segments voisés, voir plus performants, pour les segments non-voisés, que les approches actuelles basées sur l'utilisation de facteurs moyens. Des tests subjectifs seront tout de même nécessaires pour attester de l'intérêt perceptifs de cette nouvelle approche. Il serait aussi intéressant d'étudier l'utilité de la nouvelle modélisation pour d'autres styles de parole présentant des modifications temporelles plus marquées comme la parole claire.

Finalement, nos travaux proposent donc une architecture d'apprentissage des modifications de la parole objectivement plus performante, avec une conversion du style théoriquement plus naturelle, que les architectures actuelles. En revanche, des écoutes préliminaires sur l'application directes de ces modifications sur des signaux de parole avec le vocodeur STRAIGHT laisse entendre de nombreux artefacts de synthèse. Cela aura plutôt tendance à dégrader l'intelligibilité comme nous avons pu l'observer avec notre participation au challenge *Hurricane*, qui a été présenté chapitre 6, avec une diminution de l'intelligibilité par notre système de conversion. L'amplitude importante des modifications des caractéristiques acoustiques, permettant de passer d'un style de parole à un autre, complexifient grandement l'obtention d'une synthèse de qualité. C'est possiblement pour cette raison que nous étions la seule équipe à proposer une approche paramétrique pour le challenge. Une attention particulière devra alors être portée sur l'exploitation des paramètres convertis pour proposer une synthèse de meilleure qualité avant d'espérer obtenir des gains d'intelligibilité similaires aux styles naturels que sont la parole claire et la parole Lombard.

Conclusion

Apports

Les travaux menés au cours de cette thèse ont permis d'apporter des analyses, approfondissements et nouveautés dans les deux grands domaines du renforcement de la parole à savoir le renforcement direct et paramétrique. En effet, en étudiant les méthodes actuelles de renforcement de la parole, et en cherchant à les appliquer dans un cadre applicatif concret, des adaptations d'approches bien installées ont été proposées et des contributions sur des méthodes encore peu développées ont été faites.

Nouvelle contrainte perceptive

D'abord de façon général, nous avons mis en évidence une limitation des contraintes énergétiques classiques lors des mesures de performance des traitements de renforcement de la parole. En effet, dans notre contexte d'étude automobile, et dans toute application réelle où l'utilisateur a accès au niveau de présentation du signal de parole, contraindre l'énergie moyenne du signal ne s'avère pas une solution viable. Les performances alors estimées avec cette contrainte classique ne sont plus du tout représentatives en conditions réelles, surtout pour les méthodes directes de ré-allocation spectro-temporelle de l'énergie. Ainsi, nous avons proposé une nouvelle contrainte énergétique perceptive basée sur une pondération fréquentielle en dB(A) prenant en compte la sensibilité de l'oreille.

Maximisation exacte d'un critère d'intelligibilité sous la nouvelle contrainte perceptive

Nous avons ensuite montré que, parmi les méthodes de renforcement direct, celles étant basées sur la maximisation du SII sont fortement sensibles à cette nouvelle contrainte perceptive. Cette méthode populaire est très performante pour améliorer l'intelligibilité dans de multiple bruits avec une contrainte énergétique classique et, ne nécessitant que le signal de parole et une estimation spectrale du bruit ambiant, elle est parfaitement adaptée à notre contexte bruité automobile. En revanche, sa tendance à concentrer l'énergie spectrale dans les zones sensibles de l'oreille est indéniablement un facteur important pour le gain d'intelligibilité recensé et ce comportement est nécessairement atténué par la nouvelle contrainte perceptive proposée. Après avoir remarqué que les approches de maximisation du SII existantes étaient toutes basées sur des approximations du critère, nous avons proposé une procédure d'optimisation exacte, basée sur une résolution en sous-problèmes convexes. Cette procédure nous permet, d'une part, de nous assurer que les traitements utilisés maximisent exactement le SII, et, d'une autre part, d'étudier l'optimalité des résultats obtenus par les approximations. La procédure proposée prend en compte n'importe quelle pondération pour la contrainte énergétique et nous avons adapté les approches par approximation pour que ce soit aussi le cas à des fins de comparaison. Un examen approfondi des spectres obtenus sur des bruits classiques a alors été mené permettant de comprendre et évaluer chaque approche. Dans les bruits classiques considérés, les méthodes par approximation du SII proposent des spectres quasi-optimaux mais chacune dans des plages de RSB différents, nous avons alors proposé une extension couplant ces méthodes améliorant les résultats et approchant d'autant plus les spectres optimaux sur l'ensemble des RSB. Finalement, l'influence de la nouvelle contrainte énergétique perceptive provoque l'effet attendu sur les spectres optimaux, rendant la distribution plus étalée en tempérant la concentration de l'énergie dans les zones sensibles de l'oreille.

Des tests perceptifs d'intelligibilité ont ensuite été menés afin d'attester des performances de l'approche avec l'exactitude de l'optimisation et la nouvelle contrainte proposée. D'abord nous avons participé à la deuxième édition du challenge *Hurricane* regroupant de nombreux algorithmes actuels de renforcement direct de la parole et proposant de les évaluer pour différentes langues dans différentes conditions de réverbération. La contrainte énergétique classique était imposée mais cela restait une bonne opportunité de tester les performances de l'optimisation exacte du SII et étudier sa robustesse à la réverbération. Les résultats obtenus sont très satisfaisants, du même ordre de grandeur que la majorité des approches, et notre méthode propose même les meilleures performances à faible RSB, réverbération proche, pour la langue anglaise. D'autant plus que les méthodes qui performent le mieux sont toutes équipées d'un module de compression dynamique supplémentaire, connu pour

améliorer les performances mais rendant l'analyse des contributions des algorithmes impossible sans référence. Enfin, des modifications inopinées des niveaux de présentation annoncés lors des tests ont rendu caduque l'analyse détaillée des effets de la réverbération sur notre méthode qui est très sensible au RSB. Bien que le paramétrage de la méthode était inadapté, des gains d'intelligibilité toujours équivalent aux autres méthodes ont été relevés supposant alors une certaine robustesse de la méthode à la réverbération.

Nous avons aussi mis en place des tests perceptifs afin d'étudier l'influence de la nouvelle contrainte énergétique perceptive sur la méthode de maximisation exacte du SII dans différentes situations automobiles. Trois bruits ont été choisis pour leur diversité énergétique et spectrale : un bruit basse vitesse (basse fréquence, peu énergétique), un bruit haute vitesse (basse fréquence, très énergétique) et un bruit basse vitesse en présence de pluie (large bande, modérément énergétique). Pour les bruits basse fréquence les gains d'intelligibilité sont très significatifs bien qu'ils soient légèrement plus important dans le bruit haute vitesse que le bruit basse vitesse. En présence de pluie, le bruit présente un spectre plus aplati, les gains d'intelligibilité sont alors extrêmement réduits et ne sont même plus significatifs. Ainsi, même avec la nouvelle contrainte énergétique perceptive, la méthode de maximisation exacte du SII est toujours très performante pour des bruits au spectre localisé car une ré-allocation énergétique astucieuse dans les zones spectrales avec peu de bruit permet d'atteindre des gains d'intelligibilité très intéressants. En revanche, pour un spectre de bruit large bande, il n'y a plus de bandes où le bruit est peu présent et la contrainte énergétique empêche une concentration trop importante de l'énergie dans les zones sensibles de l'oreille, ce qui provoque une redistribution de l'énergie plus étalée n'engendrant pas de gain d'intelligibilité significatif.

Améliorations des aspects temporels pour le renforcement paramétrique par conversion de voix

Concernant le renforcement paramétrique de la parole, nous avons relevé que de nombreuses limitations des méthodes de renforcement par modification de la parole sont supposées, par la littérature, provenir du manque de prise en compte du contexte des phonèmes dans les transformations empiriques proposées. C'est pourquoi nous nous sommes intéressés aux approches par conversion du style de parole cherchant à apprendre automatiquement des transformations, bien plus contextuelles et cohérentes, à appliquer aux paramètres de la parole afin de passer d'une parole neutre à une parole plus intelligible comme la parole Lombard ou la parole claire. Les résultats préliminaires de ce domaine, encore peu étudié, attestent de très bon scores de similarité mais de gains d'intelligibilité encore très éloignés de ceux des paroles naturelles. Les artefacts de synthèse introduits par les vocodeurs sont supposés, en grande partie, responsables de cette différence d'intelligibilité mais nous pensons que la négligence des aspects temporels y joue un rôle aussi très important. Premièrement, sur l'aspect dynamique des caractéristiques, l'utilisation classique de modèles GMM ou FFNN considère chaque échantillon comme étant indépendant et, malgré l'utilisation des *delta*/ *delta-delta features*, c'est un contexte très localisé qui est pris en compte. Nous proposons alors l'utilisation de modèles spécifiquement adaptés au traitement de séquences temporelles, que sont les RNN, permettant une prise en compte du contexte à des échelles bien plus grandes e.g. phonème, syllabe, mot et même phrase. De plus, l'utilisation d'une transformée en ondelettes est proposée afin d'aider à la représentation des caractéristiques sur différentes échelles temporelles. Deuxièmement, les aspects temporels liés à la manipulation du débit se sont toujours contentés d'appliquer des facteurs de modifications binaires basé sur le voisement des segments de parole. Nous avons alors proposé une nouvelle modélisation lisse des modifications temporelles directement intégrable dans les algorithmes d'apprentissage automatique et ayant fait l'objet d'un dépôt de brevet d'invention. Cette nouvelle représentation des modifications temporelles permet un apprentissage contextuel beaucoup plus poussé, en espérant aboutir à des transformations du débit plus naturel voir même plus efficace du point de vue de l'intelligibilité.

L'utilisation d'architectures récurrentes RNN, pour une tâche de conversion de parole neutre vers parole Lombard, s'est soldée par une tendance à l'amélioration des performances d'apprentissage mais non significative à cause d'une variance trop importante. En revanche, ces architectures parviennent à exploiter pleinement la représentation du fondamental sur de multiples échelles temporelles par la transformée en ondelettes en proposant alors des performances d'apprentissage de cette caractéristique significativement supérieures au modèle FFNN classique. L'introduction de la nouvelle modélisation des modifications temporelles dans le système de

conversion permet un apprentissage objectivement supérieur des modifications temporelles comparé à l'utilisation des facteurs binaires basés sur le voisement et plus particulièrement pour les sons non-voisés qui profitent pleinement de cette nouvelle représentation des facteurs de modification du débit de parole.

Limites et perspectives

Les interprétations liées à nos contributions ont aussi mis en évidence des limites permettant de contraster les résultats obtenus et laissant place à des perspectives riches d'approfondissement des problématiques introduites.

Légitimité de la nouvelle contrainte perceptive

En commençant par la contrainte perceptive introduite basée sur une échelle en dB(A) qui peut être critiquée pour mesurer le niveau perçu d'un signal de parole dans des bruits aux contenus spectraux et aux intensités variables. À travers des écoutes officieuses, nous avons relevé que la maximisation du SII sous la contrainte énergétique classique entraîne une augmentation évidente du niveau perçu et que sous la contrainte en dB(A) cet effet est largement minimisé. En revanche, bien que le niveau perçu soit mieux préservé, nous ne pouvons pas affirmer qu'il soit parfaitement maintenu. Une étude permettant de confirmer, voir ajuster, la pondération dans les bruits considérés, ne peut être que recommandée afin d'éviter tout ajustement du volume par l'utilisateur lors d'une écoute réelle.

Regard critique sur l'hypothèse de la méthode de maximisation du SII

Concernant la méthode de maximisation exacte du SII, on peut souligner qu'elle se base sur une hypothèse très forte supposant qu'une maximisation aveugle de ce critère serait optimal d'un point de vu de l'intelligibilité. Lors de notre participation au challenge *Hurricane*, le changement imprévu des niveaux de présentation annoncés a engendré des résultats inadaptés, nous empêchant d'étudier l'influence de la réverbération mais nous permettant aussi de remettre en question l'hypothèse précédente. En effet, il semblerait qu'un paramétrage plus doux de l'égalisation fréquentielle pourrait engendrer des gains d'intelligibilité supérieurs bien que sous-optimaux vis-à-vis du SII. Une proposition préliminaire d'adaptation du critère afin de générer des égalisations plus douces et contrastées a été faite mais nécessiterait un travail de paramétrage développé avec une forte dépendance au domaine d'écoute considéré. Cependant, on peut imaginer une adaptation automatique du critère en fonction des caractéristiques spectrales du bruit.

Conformité du protocole de test

Les résultats obtenus en contextes bruités automobiles, pour cette méthode de maximisation du SII, peuvent aussi être discutés vis-à-vis du protocole suivi. En effet, les tests d'intelligibilité ont été effectués avec une écoute de bruits automobiles à travers un casque audio mais le facteur important lié à la charge cognitive de conduite n'a pas été pris en compte. La mise en place de tests d'intelligibilité en simulateur de conduite est alors envisagée afin d'étudier les gains d'intelligibilité en situation d'écoute automobile réelle. D'autant plus que les RSB atteints lors des tests subjectifs sont extrêmement bas et demandent une concentration très importante pour répéter les stimuli verbaux. On peut donc s'attendre à des niveaux d'écoute plus élevés en situation réelle et à des gains d'intelligibilité moins importants car, la méthode de maximisation du SII étant très sensible au niveau de présentation, les gains d'intelligibilité théoriques à des niveaux plus importants sont bien plus modérés.

Qualité de synthèse pour la conversion du style de parole

Le renforcement paramétrique de la parole étant un domaine encore très jeune, le travail préliminaire proposé est encourageant mais demande de nombreux approfondissements avant d'être opérationnel. D'abord,

comme nous l'avons vu avec notre deuxième participation au challenge *Hurricane*, un effort important doit être porté sur la régularisation des paramètres convertis et sur la qualité de la synthèse des signaux de parole résultants au risque de plus dégrader l'intelligibilité que de l'améliorer.

Adaptation de la base de données

D'après les résultats préliminaires obtenus sur la base de données Lombard-GRID, un apprentissage très performant des modifications naturelles des paramètres de parole est fortement envisageable. Mais cet apprentissage nécessite une quantité de données plus importante pour pouvoir exploiter pleinement le traitement de séquences temporelles par les outils adaptés qui augmente significativement le nombre de paramètres complexifiant grandement l'optimisation des modèles. De plus, les stratégies Lombard sur lesquelles le système de conversion apprend sont produites dans un bruit SSN avec des phrases dénuées de sens et ces paramètres ne sont certainement pas parfaitement adaptés à notre domaine d'étude automobile. Ces deux limites peuvent être surmontées en choisissant une base de données plus adaptée comme SpeechDatCAR qui est produite en contexte bruité automobile sur des phrases sensées à partir de nombreux locuteurs et avec une très bonne qualité d'enregistrement. Dans cette base de données, la présence de multiples environnement automobile, et de langues différentes, permettrait aussi de rajouter des paramètres supplémentaire dans l'apprentissage et d'obtenir un système de conversion Lombard adaptatif aux conditions de conduite et à la langue des signaux de parole traités. Au regard des coûts d'obtention d'une telle base de donnée, une étude préliminaire comme celle que nous avons menée était indispensable avant de pouvoir considérer une telle acquisition.

Approfondissement de l'intérêt de la nouvelle modélisation des modifications temporelles

Pour ce qui est de la nouvelle modélisation des modifications du débit de parole, les résultats objectifs obtenus sont très encourageants mais des études perceptives permettant de mettre en évidence l'intérêt de cette nouvelle caractéristique sont indispensables. D'une part, d'un point de vue de l'intelligibilité, avec une synthèse de meilleure qualité, nous pouvons nous attendre à ce que des modifications complexes du débit couplées à la transformation des autres paramètres de parole, avec une prise en compte importante du contexte des phonèmes, engendre un système de conversion très performant approchant les gains d'intelligibilité procurés par la parole Lombard naturelle. D'une autre part, d'un point de vue similarité, les modifications du débit proposées, plus naturelles que les modifications classiques binaires, pourrait proposer une conversion du style se rapprochant d'une parole Lombard plus authentique.

Exploitation de la parole claire

Dans notre étude, nous nous sommes concentrés sur de la conversion vers la parole Lombard, déjà abordée dans d'autres travaux de conversion de parole, mais la parole claire est aussi un style procurant un gain d'intelligibilité important dans le bruit, jamais abordé en conversion de parole : il serait alors intéressant de mener une étude similaire sur ce style de parole. En effet, il a été relevé par d'autres travaux que les modifications spectrales de la parole claire, et notamment l'élargissement de l'espace vocalique qu'on ne retrouve pas dans la parole Lombard, engendraient des gains d'intelligibilité très importants dans le bruit. Pourtant, les tentatives d'introduire numériquement ces modifications n'ont pas engendré de gain d'intelligibilité et la raison soupçonnée concerne le manque de prise en compte du contexte et des spécificités des phonèmes traités. Notre système de conversion serait alors capable de prendre en compte ces particularités et de proposer un traitement plus prometteur concernant l'amélioration de l'intelligibilité de la parole. Concernant les modifications temporelles de la parole claire, l'apport à l'intelligibilité est plus discuté mais un aspect reconnu est que si elles y participent, alors une modélisation complexe est nécessaire pour pouvoir l'exploiter et c'est justement ce que nous proposons. Les modifications temporelles naturelles de la parole claire sont beaucoup plus présentes et marquées que celles de la parole Lombard, il serait alors très intéressant d'étudier les résultats perceptifs de notre modélisation sur ce mode de parole.

Potentielle portée de nos contributions dans d'autres domaines d'études

Enfin, la modélisation lisse des modifications temporelles proposée peut être exploitée en dehors des problématiques d'intelligibilité comme pour de la conversion d'émotion, ou de la conversion d'identité. En capturant les modifications temporelles de manière bien plus complexe et contextuelle que les méthodes classiques, notre nouvelle modélisation pourrait améliorer les performances de conversion objective, et subjective, dans d'autres domaine d'études. Plus particulièrement lorsque le débit de parole joue un rôle perceptif très important, par exemple pour de la conversion de parole neutre vers parole énervée, un traitement plus naturel et mieux adapté qu'une simple décision binaire semble prometteur du point de vue de l'authenticité du résultat.

Bibliographie

[1] M. ABE, S. NAKAMURA, K. SHIKANO et H. KUWABARA. « Voice conversion through vector quantization ». In : *Journal of the Acoustical Society of Japan (E)* 11.2 (1990), p. 71-76.

[2] T. ABE, T. KOBAYASHI et S. IMAI. « Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency ». In : *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. T. 2. IEEE. 1996, p. 1277-1280.

[3] T. ABE, T. KOBAYASHI et S. IMAI. « The IF spectrogram: a new spectral representation ». In : *Proc. ASVA* 97 (1997), p. 423-430.

[4] E. M. ADAMS et R. E. MOORE. « Effects of speech rate, background noise, and simulated hearing loss on speech rate judgment and speech intelligibility in young listeners ». In : *Journal of the American Academy of Audiology* 20.1 (2009), p. 28-39.

[5] R. AIHARA, R. TAKASHIMA, T. TAKIGUCHI et Y. ARIKI. « GMM-based emotional voice conversion using spectrum and prosody features ». In : *American Journal of Signal Processing* 2.5 (2012), p. 134-138.

[6] N. ALGHAMDI, S. MADDOCK, R. MARXER, J. BARKER et G. J. BROWN. « A corpus of audio-visual Lombard speech with frontal and profile views ». In : *The Journal of the Acoustical Society of America* 143.6 (2018), EL523-EL529.

[7] ANSI. « S3. 5-1997, Methods for the calculation of the speech intelligibility index ». In : *New York: American National Standards Institute* 19 (1997), p. 90-119.

[8] Y. ARAI et K. UKENA. « Influences on Virtual Car Driving while Paying Attention to Listening to Speech ». In : *Proceedings of the 13th ITS World Congress, London, 8-12 October 2006*. 2006.

[9] B. S. ATAL et S. L. HANAUER. « Speech analysis and synthesis by linear prediction of the speech wave ». In : *The journal of the acoustical society of America* 50.2B (1971), p. 637-655.

[10] V. AUBANEL et M. COOKE. « Information-preserving temporal reallocation of speech in the presence of fluctuating maskers. » In : *Interspeech*. 2013, p. 3592-3596.

[11] V. AUBANEL et M. COOKE. « Strategies adopted by talkers faced with fluctuating and competing-speech maskers ». In : *The Journal of the Acoustical Society of America* 134.4 (2013), p. 2884-2894.

[12] British Society of AUDIOLOGY. *Recommended Procedure for Pure-tone air-conduction and bone-conduction threshold audiometry with and without masking*. http://www.thebsa.org.uk, dernière visite le 9 décembre 2020.

[13] E. AXMEAR, J. REICHLE, M. ALAMSAPUTRA, K. KOHNERT, K. DRAGER et K. SELLNOW. « Synthesized speech intelligibility in sentences ». In : *Language, Speech, and Hearing Services in Schools* (2005).

[14] R. BADEAU et B. DAVID. « Weighted maximum likelihood autoregressive and moving average spectrum modeling ». In : *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2008, p. 3761-3764.

[15] R. BAKER et V. HAZAN. « LUCID: a corpus of spontaneous and read clear speech in British English ». In : *DiSS-LPSS Joint Workshop 2010*. 2010.

[16] H. BANNO, H. HATA, M. MORISE, T. TAKAHASHI, T. IRINO et H. KAWAHARA. « Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation ». In : *Acoustical science and technology* 28.3 (2007), p. 140-146.

[17] F. BEDERNA, H. SCHEPKER, C. ROLLWAGE, S. DOCLO, A. PUSCH, J. BITZER et J. RENNIES. « Adaptive compressive onset-enhancement for improved speech intelligibility in noise and reverberation ». In : *Proceedings of Interspeech*. 2020.

[18] J. BENCH et J. BAMFORD. *Speech-hearing tests and the spoken language of hearing-impaired children*. Academic Press, 1979.

[19] J. BENESTY, M. SONDHI et Y. HUANG. *Springer Handbook of Speech Processing*. Springer, 2007.

[20] T. BENT et A. R. BRADLOW. « The interlanguage speech intelligibility benefit ». In : *The Journal of the Acoustical Society of America* 114.3 (2003), p. 1600-1610.

[21] Z. S. BOND et T. J. MOORE. « A note on loud and Lombard speech ». In : *First International Conference on Spoken Language Processing*. 1990.

[22] Z. S. BOND et T. J. MOORE. « A note on the acoustic-phonetic characteristics of inadvertently clear speech ». In : *Speech communication* 14.4 (1994), p. 325-337.

[23] Z. S. BOND, T. J. MOORE et B. GABLE. « Acoustic–phonetic characteristics of speech produced in noise and while wearing an oxygen mask ». In : *The Journal of the Acoustical Society of America* 85.2 (1989), p. 907-912.

[24] H. BORIL et P. POLLÁK. « Design and collection of Czech Lombard speech database ». In : *Ninth European Conference on Speech Communication and Technology*. 2005.

[25] A. R. BRADLOW, M. BLASINGAME et K. LEE. « Language-independent talker-specificity in bilingual speech intelligibility: Individual traits persist across first-language and second-language speech ». In : *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 9.1 (2018).

[26] A. R. BRADLOW, N. KRAUS et E. HAYES. « Speaking clearly for children with learning disabilities ». In : *Journal of Speech, Language, and Hearing Research* (2003).

[27] A. R. BRADLOW, G. M. TORRETTA et D. B. PISONI. « Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics ». In : *Speech communication* 20.3 (1996), p. 255.

[28] T. BRAND et B. KOLLMEIER. « Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests ». In : *The Journal of the Acoustical Society of America* 111.6 (2002), p. 2801-2810.

[29] K. M. BRETTHAUER et B. SHETTY. « The nonlinear knapsack problem–algorithms and applications ». In : *European Journal of Operational Research* 138.3 (2002), p. 459-472.

[30] H. BROUCKXON, W. VERHELST et B. D. SCHUYMER. « Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments ». In : *Ninth Annual Conference of the International Speech Communication Association*. 2008.

[31] D. BYRD. « Relations of sex and dialect to reduction ». In : *Speech Communication* 15.1-2 (1994), p. 39-54.

[32] S. CAMERON et H. DILLON. « Development of the listening in spatialized noise-sentences test (LISN-S) ». In : *Ear and hearing* 28.2 (2007), p. 196-211.

[33] O. CAPPÉ, J. LAROCHE et E. MOULINES. « Regularized estimation of cepstrum envelope from discrete frequency points ». In : *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE. 1995, p. 213-216.

[34] A. CASTELLANOS, J. M. BENEDI et F. CASACUBERTA. « An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect ». In : *Speech Communication* 20.1-2 (1996), p. 23-35.

[35] P. S. CHANDA et S. PARK. « Speech intelligibility enhancement using tunable equalization filter ». In : *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. T. 4. IEEE. 2007, p. IV-613.

[36] F. CHARPENTIER. « Pitch detection using the short-term phase spectrum ». In : *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*. T. 11. IEEE. 1986, p. 113-116.

[37] F. CHARPENTIER et M. STELLA. « Diphone synthesis using an overlap-add technique for speech waveforms concatenation ». In : *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*. T. 11. IEEE. 1986, p. 2015-2018.

[38] F. R. CHEN. « Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level ». Thèse de doct. Massachusetts Institute of Technology, 1980.

[39] C. CHERMAZ et S. KING. « A sound engineering approach to near end listening enhancement ». In : *Proceedings of Interspeech*. 2020.

[40] M. COOKE. « A glimpsing model of speech perception in noise ». In : *The Journal of the Acoustical Society of America* 119.3 (2006), p. 1562-1573.

[41] M. COOKE et V. AUBANEL. « Effects of linear and nonlinear speech rate changes on speech intelligibility in stationary and fluctuating maskers ». In : *The Journal of the Acoustical Society of America* 141.6 (2017), p. 4126-4135.

[42] M. COOKE, J. BARKER, S. CUNNINGHAM et X. SHAO. « An audio-visual corpus for speech perception and automatic speech recognition ». In : *The Journal of the Acoustical Society of America* 120.5 (2006), p. 2421-2424.

[43] M. COOKE, S. KING, M. GARNIER et V. AUBANEL. « The listening talker: A review of human and algorithmic context-induced modifications of speech ». In : *Computer Speech & Language* 28.2 (2014), p. 543-571.

[44] M. COOKE et Y. LU. « Spectral and temporal changes to speech produced in the presence of energetic and informational maskers ». In : *The Journal of the Acoustical Society of America* 128.4 (2010), p. 2059-2069.

[45] M. COOKE, C. MAYO et C. VALENTINI-BOTINHAO. « Intelligibility enhancing speech modifications: the hurricane challenge ». In : *Interspeech*. 2013, p. 3552-3556.

[46] M. COOKE, C. MAYO et J. VILLEGAS. « The contribution of durational and spectral changes to the Lombard speech intelligibility benefit ». In : *The Journal of the Acoustical Society of America* 135.2 (2014), p. 874-883.

[47] C. D'ALESSANDRO, V. DARSINOS et B. YEGNANARAYANA. « Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources ». In : *IEEE Transactions on Speech and Audio processing* 6.1 (1998), p. 12-23.

[48] Collège National D'AUDIOPROTHÈSE. *CD d'audiométrie vocale*. https://www.college-nat-audio.fr/cd/coffret-de-5-cd-audiometrie-vocale, dernière visite le 9 décembre 2020.

[49] S. DESAI, E. V. RAGHAVENDRA, B. YEGNANARAYANA, A. W. BLACK et K. PRAHALLAD. « Voice conversion using artificial neural networks ». In : *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE. 2009, p. 3893-3896.

[50] G. L. EBBITT et N. A. RAUF. *Acoustic Absorption in Vehicles and the Measurement of Short Reverberation Times*. Rapp. tech. SAE Technical Paper, 1997.

[51] G. L. EBBITT et T. M. REMTEMA. « Automotive Speech Intelligibility Measurements ». In : *Sound & Vibration* (2017), p. 7.

[52] J. J. EGAN. « Psychoacoustics of the Lombard voice response. » In : *Journal of Auditory Research* (1972).

[53] D. ERRO, A. MORENO et A. BONAFONTE. « Voice conversion based on weighted frequency warping ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 18.5 (2009), p. 922-931.

[54] S. H. FERGUSON. « Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners ». In : *The Journal of the Acoustical Society of America* 116.4 (2004), p. 2365-2373.

[55] S. H. FERGUSON et D. KEWLEY-PORT. « Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners ». In : *The Journal of the Acoustical Society of America* 112.1 (2002), p. 259-271.

[56] M. FITZPATRICK, J. KIM et C. DAVIS. « The effect of seeing the interlocutor on auditory and visual speech production in noise ». In : *Auditory-Visual Speech Processing 2011*. 2011.

[57] J. L. FLANAGAN et R. M. GOLDEN. « Phase vocoder ». In : *Bell System Technical Journal* 45.9 (1966), p. 1493-1509.

[58] J. E. FOURNIER. *Audiométrie vocale: les épreuves d'intelligibilité et leurs applications au diagnostic, à l'expertise et à la correction prothétique des surdités*. Maloine, 1951.

[59] M. GARNIER. « Communiquer en environnement bruyant: de l'adaptation jusqu'au forçage vocal ». Thèse de doct. Université Pierre et Marie Curie-Paris VI, 2007.

[60] M. GARNIER, L. BAILLY, M. DOHEN, P. WELBY et H. LŒVENBRUCK. « An acoustic and articulatory study of Lombard speech: Global effects on the utterance ». In : *Ninth International Conference on Spoken Language Processing*. 2006.

[61] M. GARNIER, M. DOHEN, H. LOEVENBRUCK, P. WELBY et L. BAILLY. « The Lombard Effect: a physiological reflex or a controlled intelligibility enhancement? » In : *7th International Seminar on Speech Production*. 2006, p. 255-262.

[62] M. GARNIER et N. HENRICH. « Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? » In : *Computer Speech & Language* 28.2 (2014), p. 580-597.

[63] M. GARNIER, N. HENRICH et D. DUBOIS. « Influence of sound immersion and communicative interaction on the Lombard effect ». In : *Journal of Speech, Language, and Hearing Research* 53.3 (2010), p. 588-608.

[64] M. GARNIER, L. MÉNARD et G. RICHARD. « Effect of being seen on the production of visible speech cues. A pilot study on Lombard speech ». In : *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[65] E. GENTET, B. DAVID, S. DENJEAN, G. RICHARD et V. ROUSSARIE. « Neutral to Lombard Speech Conversion with Deep Learning ». In : *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, p. 7739-7743.

[66] E. GENTET, B. DAVID, S. DENJEAN, G. RICHARD et V. ROUSSARIE. « Optimisation d'un critère d'Intelligibilité de la Parole dans un Contexte Bruité Automobile ». In : *CFA 2018*. 2018.

[67] E. GENTET, B. DAVID, S. DENJEAN, G. RICHARD et V. ROUSSARIE. « Speech Intelligibility Enhancement by Equalization for in-Car Applications ». In : *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, p. 6934-6938.

[68] E. GENTET, B. DAVID, S. DENJEAN, G. RICHARD et V. ROUSSARIE. *Système de conversion de la parole par apprentissage statistique avec modélisation continue des modifications temporelles*. demande de brevet FR déposée le 27 janvier 2020, numéro de dossier 2000753.

[69] E. GODOY, M. KOUTSOGIANNAKI et Y. STYLIANOU. « Approaching speech intelligibility enhancement with inspiration from Lombard and Clear speaking styles ». In : *Computer Speech & Language* 28.2 (2014), p. 629-647.

[70] R. L. GOLDSWORTHY et J. E. GREENBERG. « Analysis of speech-based speech transmission index methods with implications for nonlinear operations ». In : *The Journal of the Acoustical Society of America* 116.6 (2004), p. 3679-3689.

[71] D. W. GRIFFIN et J. S. LIM. « Multiband excitation vocoder ». In : *IEEE Transactions on acoustics, speech, and signal processing* 36.8 (1988), p. 1223-1235.

[72] J. GROSSE et S. van de PAR. « A speech preprocessing method based on overlap-masking reduction to increase intelligibility in reverberant environments ». In : *Journal of the Audio Engineering Society* 65.1/2 (2017), p. 31-41.

[73] B. HAGERMAN et C. KINNEFORS. « Efficient adaptive methods for measuring speech reception threshold in quiet and in noise ». In : *Scandinavian audiology* 24.1 (1995), p. 71-77.

[74] J. L. HALL et J. L. FLANAGAN. « Intelligibility and listener preference of telephone speech in the presence of babble noise ». In : *The Journal of the Acoustical Society of America* 127.1 (2010), p. 280-285.

[75] V. HAZAN, J. GRYNPAS et R. BAKER. « Is clear speech tailored to counter the effect of specific adverse listening conditions? » In : *The Journal of the Acoustical Society of America* 132.5 (2012), EL371-EL377.

[76] V. HAZAN et D. MARKHAM. « Acoustic-phonetic correlates of talker intelligibility for adults and children ». In : *The Journal of the Acoustical Society of America* 116.5 (2004), p. 3108-3118.

[77] V. HOHMANN et B. KOLLMEIER. « The effect of multichannel dynamic compression on speech intelligibility ». In : *The Journal of the Acoustical Society of America* 97.2 (1995), p. 1191-1195.

[78] B. W. HORNSBY. « The Speech Intelligibility Index: What is it and what's it good for? » In : *The Hearing Journal* 57.10 (2004), p. 10-17.

[79] D. Y. HUANG, S. RAHARDJA et E. P. ONG. « Lombard effect mimicking ». In : *Seventh ISCA Workshop on Speech Synthesis*. 2010.

[80] Z. INANOGLU et S. YOUNG. « Data-driven emotion conversion in spoken English ». In : *Speech Communication* 51.3 (2009), p. 268-283.

[81] F. ITAKURA. « Analysis synthesis telephony based on the maximum likelihood method ». In : *The 6th international congress on acoustics, 1968*. 1968, p. 280-292.

[82] E. JOKINEN, P. ALKU et M. VAINIO. « Comparison of post-filtering methods for intelligibility enhancement of telephone speech ». In : *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE. 2012, p. 2333-2337.

[83] J. C. JUNQUA. « The Lombard reflex and its role on human listeners and automatic speech recognizers ». In : *The Journal of the Acoustical Society of America* 93.1 (1993), p. 510-524.

[84] J. C. JUNQUA, S. FINCKE et K. FIELD. « Influence of the speaking style and the noise spectral tilt on the Lombard reflex and automatic speech recognition ». In : *Fifth International Conference on Spoken Language Processing*. 1998.

[85] N. KADIRI. « Conséquences d'un environnement bruité sur la production de la parole ». Thèse de doct. Toulouse 3, 1998.

[86] A. KAIN et M. W. MACON. « Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction ». In : *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. T. 2. IEEE. 2001, p. 813-816.

[87] A. KAIN et M. W. MACON. « Spectral voice conversion for text-to-speech synthesis ». In : *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*. T. 1. IEEE. 1998, p. 285-288.

[88] J. M. KATES. « Understanding compression: Modeling the effects of dynamic-range compression in hearing aids ». In : *International journal of audiology* 49.6 (2010), p. 395-409.

[89] J. M. KATES et K. H. AREHART. « Coherence and the speech intelligibility index ». In : *The journal of the acoustical society of America* 117.4 (2005), p. 2224-2237.

[90] J. M. KATES et K. H. AREHART. « The hearing-aid speech perception index (HASPI) ». In : *Speech Communication* 65 (2014), p. 75-93.

[91] H. KAWAHARA. « Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited ». In : *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*. T. 2. IEEE. 1997, p. 1303-1306.

[92] H. KAWAHARA, H. KATAYOSE, A. DE CHEVEIGNÉ et R. D. PATTERSON. « Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity ». In : *Sixth european conference on speech communication and technology*. 1999.

[93] H. KAWAHARA, I. MASUDA-KATSUSE et A. DE CHEVEIGNE. « Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds1 ». In : *Speech communication* 27.3-4 (1999), p. 187-207.

[94] H. KAWAHARA, M. MORISE, T. TAKAHASHI, R. NISIMURA, T. IRINO et H. BANNO. « Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation ». In : *Acoustics, Speech and Signal Processing, ICASSP. IEEE International Conference on*. IEEE. 2008, p. 3933-3936.

[95] H. KAWANAMI, Y. IWAMI, T. TODA, H. SARUWATARI et K. SHIKANO. « GMM-based voice conversion applied to emotional speech synthesis ». In : *Eighth European Conference on Speech Communication and Technology (Eurospeech)*. 2003.

[96] M. C. KILLION et S. FIKRET-PASA. « The 3 types of sensorineural hearing loss: Loudness and intelligibility considerations ». In : *Hearing journal* 46 (1993), p. 31-31.

[97] M. C. KILLION, P. A. NIQUETTE, G. I. GUDMUNDSEN, L. J. REVIT et S. BANERJEE. « Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners ». In : *The Journal of the Acoustical Society of America* 116.4 (2004), p. 2395-2405.

[98] J. C. KRAUSE et L. D. BRAIDA. « Acoustic properties of naturally produced clear speech at normal speaking rates ». In : *The Journal of the Acoustical Society of America* 115.1 (2004), p. 362-378.

[99] J. C. KRAUSE et L. D. BRAIDA. « Evaluating the role of spectral and envelope characteristics in the intelligibility advantage of clear speech ». In : *The Journal of the Acoustical Society of America* 125.5 (2009), p. 3346-3357.

[100] J. C. KRAUSE et L. D. BRAIDA. « Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility ». In : *The Journal of the Acoustical Society of America* 112.5 (2002), p. 2165-2172.

[101] E. A. KRETSINGER et N. B. YOUNG. « The use of fast limiting to improve the intelligibility of speech in noise ». In : *Communications Monographs* 27.1 (1960), p. 63-69.

[102] K. D. KRYTER, J. C. R. LICKLIDER et S. S. STEVENS. « Premodulation clipping in AM voice communication ». In : *The Journal of the Acoustical Society of America* 19.1 (1947), p. 125-131.

[103] H. LANE, B. TRANEL et C. SISSON. « Regulation of voice communication by sensory dynamics ». In : *The Journal of the Acoustical Society of America* 47.2B (1970), p. 618-624.

[104] J. LAROCHE et M. DOLSON. « Improved phase vocoder time-scale modification of audio ». In : *IEEE Transactions on Speech and Audio processing* 7.3 (1999), p. 323-332.

[105] J. LAROCHE, Y. STYLIANOU et E. MOULINES. « HNM: A simple, efficient harmonic+ noise model for speech ». In : *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE. 1993, p. 169-172.

[106] R. H. LASKAR, D. CHAKRABARTY, F. A. TALUKDAR, K. S. RAO et K. BANERJEE. « Comparing ANN and GMM in a voice conversion framework ». In : *Applied Soft Computing* 12.11 (2012), p. 3332-3342.

[107] B. LEE, M. HASEGAWA-JOHNSON, C. GOUDESEUNE, S. KAMDAR, S. BORYS, M. LIU et T. HUANG. « AVICAR: Audio-visual speech corpus in a car environment ». In : *Eighth International Conference on Spoken Language Processing*. 2004.

[108] G. LI, X. WANG, R. HU, H. ZHANG et S. KE. « Normal-To-Lombard Speech Conversion by LSTM Network and BGMM for Intelligibility Enhancement of Telephone Speech ». In : *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2020, p. 1-6.

[109] H. LI, S. W. FU, Y. TSAO et J. YAMAGISHI. « iMetricGAN: Intelligibility Enhancement for Speech-in-Noise using Generative Adversarial Network-based Metric Learning ». In : *arXiv preprint arXiv:2004.00932* (2020).

[110] L. LI, K. JAMIESON, G. DESALVO, A. ROSTAMIZADEH et A. TALWALKAR. « Hyperband: A novel bandit-based approach to hyperparameter optimization ». In : *arXiv preprint arXiv:1603.06560* (2016).

[111] J. S. LIÉNARD et M. G. DI BENEDETTO. « Effect of vocal effort on spectral properties of vowels ». In : *The Journal of the Acoustical Society of America* 106.1 (1999), p. 411-422.

[112] P. C. LOIZOU. *Speech enhancement: theory and practice*. CRC press, 2013.

[113] E. LOMBARD. « Le signe de l'elevation de la voix ». In : *Ann. Mal. de L'Oreille et du Larynx* (1911), p. 101-119.

[114] A. R. LÓPEZ, S. SESHADRI, L. JUVELA, O. RÄSÄNEN et P. ALKU. « Speaking style conversion from normal to Lombard speech using a glottal vocoder and Bayesian GMMs ». In : *Proc. Interspeech 2017* (2017), p. 1363-1367.

[115] Y. LU et M. COOKE. « Speech production modifications produced by competing talkers, babble, and stationary noise ». In : *The Journal of the Acoustical Society of America* 124.5 (2008), p. 3261-3275.

[116] Y. LU et M. COOKE. « Speech production modifications produced in the presence of low-pass and high-pass filtered noise ». In : *The Journal of the Acoustical Society of America* 126.3 (2009), p. 1495-1499.

[117] Y. LU et M. COOKE. « The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise ». In : *Speech Communication* 51.12 (2009), p. 1253-1262.

[118] J. MA, Y. HU et P. C. LOIZOU. « Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions ». In : *The Journal of the Acoustical Society of America* 125.5 (2009), p. 3387-3405.

[119] C. MACKENZIE et J. GREEN. « Cognitive–linguistic deficit and speech intelligibility in chronic progressive multiple sclerosis ». In : *International journal of language & communication disorders* 44.4 (2009), p. 401-420.

[120] K. C. MACKIE et P. J. DERMODY. *Word intelligibility tests in audiology for the assessment of communication adequacy*. Australian Government Publishing Service, 1982.

[121] C. MAYO, V. AUBANEL et M. COOKE. « Effect of prosodic changes on speech intelligibility ». In : *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[122] R. MCAULAY et T. QUATIERI. « Speech analysis/synthesis based on a sinusoidal representation ». In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34.4 (1986), p. 744-754.

[123] I.V. MCLOUGHLIN et R.J. CHANCE. « LSP-based speech modification for intelligibility enhancement ». In : *Digital Signal Processing Proceedings, 1997. DSP 97., 1997 13th International Conference on*. T. 2. IEEE. 1997, p. 591-594.

[124] S. E. MILLER, R. S. SCHLAUCH et P. J. WATSON. « The effects of fundamental frequency contour manipulations on speech intelligibility in background noise ». In : *The Journal of the Acoustical Society of America* 128.1 (2010), p. 435-443.

[125] H. MING, D. HUANG, M. DONG, H. LI, L. XIE et S. ZHANG. « Fundamental frequency modeling using wavelets for emotional voice conversion ». In : *International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2015, p. 804-809.

[126] H. MING, D. HUANG, L. XIE, J. WU, M. DONG et H. LI. « Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion ». In : (2016).

[127] C. MOKBEL. « Reconnaissance de la parole dans le bruit: bruitage/débruitage ». Thèse de doct. 1992.

[128] R. B. MONSEN. « The oral speech intelligibility of hearing-impaired talkers ». In : *Journal of Speech and Hearing Disorders* 48.3 (1983), p. 286-296.

[129] B. C. MOORE et B. R. GLASBERG. « Modeling binaural loudness ». In : *The Journal of the Acoustical Society of America* 121.3 (2007), p. 1604-1612.

[130] B. C. MOORE et B. R. GLASBERG. « Suggested formulae for calculating auditory-filter bandwidths and excitation patterns ». In : *The journal of the acoustical society of America* 74.3 (1983), p. 750-753.

[131] B. C. MOORE, B. R. GLASBERG et T. BAER. « A model for the prediction of thresholds, loudness, and partial loudness ». In : *Journal of the Audio Engineering Society* 45.4 (1997), p. 224-240.

[132] A. MORENO, B. LINDBERG, C. DRAXLER, G. RICHARD, K. CHOUKRI, S. EULER et J. ALLEN. « SPEECHDAT-CAR. a large speech database for automotive environments. » In : *LREC*. 2000.

[133] M. MORISE. « CheapTrick, a spectral envelope estimator for high-quality speech synthesis ». In : *Speech Communication* 67 (2015), p. 1-7.

[134] M. MORISE. « D4C, a band-aperiodicity estimator for high-quality speech synthesis ». In : *Speech Communication* 84 (2016), p. 57-65.

[135] M. MORISE. « Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals. » In : *INTERSPEECH*. 2017, p. 2321-2325.

[136] M. MORISE. « PLATINUM: A method to extract excitation signals for voice synthesis system ». In : *Acoustical Science and Technology* 33.2 (2012), p. 123-125.

[137] M. MORISE, H. KAWAHARA et T. NISHIURA. « Rapid F0 estimation for high-SNR speech based on fundamental component extraction ». In : *Trans. IEICEJ* 93 (2010), p. 109-117.

[138] M. MORISE et Y. WATANABE. « Sound quality comparison among high-quality vocoders by using resynthesized speech ». In : *Acoustical Science and Technology* 39.3 (2018), p. 263-265.

[139] M. MORISE, F. YOKOMORI et K. OZAWA. « WORLD: a vocoder-based high-quality speech synthesis system for real-time applications ». In : *IEICE TRANSACTIONS on Information and Systems* 99.7 (2016), p. 1877-1884.

[140] A. MOUCHTARIS, J. VAN DER SPIEGEL et P. MUELLER. « Nonparallel training for voice conversion based on a parameter adaptation approach ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 14.3 (2006), p. 952-963.

[141] E. MOULINES et F. CHARPENTIER. « Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones ». In : *Speech communication* 9.5-6 (1990), p. 453-467.

[142] I. R. MURRAY et J. L. ARNOTT. « Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion ». In : *The Journal of the Acoustical Society of America* 93.2 (1993), p. 1097-1108.

[143] M. NARENDRANATH, H. A. MURTHY, S. RAJENDRAN et B. YEGNANARAYANA. « Transformation of formants for voice conversion using artificial neural networks ». In : *Speech communication* 16.2 (1995), p. 207-216.

[144] K. NATHWANI, M. DANIEL, G. RICHARD, B. DAVID et V. ROUSSARIE. « Formant shifting for speech intelligibility improvement in car noise environment ». In : *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, p. 5375-5379.

[145] K. NATHWANI, G. RICHARD, B. DAVID, P. PRABLANC et V. ROUSSARIE. « Speech intelligibility improvement in car noise environment by voice transformation ». In : *Speech Communication* 91 (2017), p. 17-27.

[146] A. T. NEEL. « Intelligibility of normal speakers: Error analysis ». In : *The Journal of the Acoustical Society of America* 98.5 (1995), p. 2982-2983.

[147] R. NIEDERJOHN et J. GROTELUESCHEN. « The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression ». In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24.4 (1976), p. 277-282.

[148] J. B. NIELSEN et T. DAU. « Development of a Danish speech intelligibility test ». In : *International journal of audiology* 48.10 (2009), p. 729-741.

[149] J. B. NIELSEN et T. DAU. « The Danish hearing in noise test ». In : *International journal of audiology* 50.3 (2011), p. 202-208.

[150] M. NILSSON, S. D. SOLI et J. A. SULLIVAN. « Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise ». In : *The Journal of the Acoustical Society of America* 95.2 (1994), p. 1085-1099.

[151] R. PATEL, M. EVERETT et E. SADIKOV. « Loudmouth:: modifying text-to-speech synthesis in noise ». In : *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. ACM. 2006, p. 227-228.

[152] R. PATEL et K. W. SCHELL. « The influence of linguistic content on the Lombard effect ». In : *Journal of Speech, Language, and Hearing Research* 51.1 (2008), p. 209-220.

[153] K. L. PAYTON, R. M. UCHANSKI et L. D. BRAIDA. « Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing ». In : *The Journal of the Acoustical Society of America* 95.3 (1994), p. 1581-1592.

[154] M. A. PICHENY, N. I. DURLACH et L. D. BRAIDA. « Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech ». In : *Journal of Speech, Language, and Hearing Research* 29.4 (1986), p. 434-446.

[155] M. A. PICHENY, N. I. DURLACH et Louis D. BRAIDA. « Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech ». In : *Journal of Speech, Language, and Hearing Research* 32.3 (1989), p. 600-603.

[156] D. PISONI, R. BERNACKI, H. NUSBAUM et M. YUCHTMAN. « Some acoustic-phonetic correlates of speech produced in noise ». In : *ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing*. T. 10. IEEE. 1985, p. 1581-1584.

[157] A. L. PITTMAN et T. L. WILEY. « Recognition of speech produced in noise ». In : *Journal of Speech, Language, and Hearing Research* 44.3 (2001), p. 487-496.

[158] M. PUCKETTE. « Phase-locked vocoder ». In : *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*. IEEE. 1995, p. 222-225.

[159] T. F. QUATIERI et R. J. MCAULAY. « Peak-to-RMS reduction of speech based on a sinusoidal model ». In : *IEEE Transactions on Signal Processing* 39.2 (1991), p. 273-288.

[160] C. M. RANKOVIC. « An application of the articulation index to hearing aid fitting ». In : *Journal of Speech, Language, and Hearing Research* 34.2 (1991), p. 391-402.

[161] D. M. RASETSHWANE. « Enhancement of speech intelligibility using speech transients extracted by a wavelet packet-based real-time algorithm ». Thèse de doct. University of Pittsburgh, 2009.

[162] J. RENNIES, H. SCHEPKER, C. VALENTINI-BOTINHAO et M. COOKE. « Intelligibility enhancing speech modifications : the hurricane challenge 2.0 ». In : *Proc. Interspeech, Shanghai, China* (2020).

[163] K. S. RHEBERGEN, N. J. VERSFELD et W. A. DRESCHLER. « Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise ». In : *The Journal of the Acoustical Society of America* 120.6 (2006), p. 3988-3997.

[164] M. S. RIBEIRO et R. A. CLARK. « A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2015, p. 4909-4913.

[165] G. RICHARD et C. D'ALESSANDRO. « Analysis/synthesis and modification of the speech aperiodic component ». In : *Speech Communication* 19.3 (1996), p. 221-244.

[166] A. RÖBEL et X. RODET. « Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation ». In : 2005.

[167] J. ROUCH et E. PARIZET. « Speech Modifications to Increase the Intelligibility of Vocal Messages Broadcast by Driving Assistance Systems Intended For Hearing-Impaired Drivers ». In : *Acta Acustica united with Acustica* 104.4 (2018), p. 668-677.

[168] S. ROUCOS et A. WILGUS. « High quality time-scale modification for speech ». In : *ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing*. T. 10. IEEE. 1985, p. 493-496.

[169] B. SAUERT, G. ENZNER et P. VARY. « Near end listening enhancement with strict loudspeaker output power constraining ». In : *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*. Citeseer. 2006.

[170] B. SAUERT et P. VARY. « Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations ». In : *Signal Processing Conference, 2010 18th European*. IEEE. 2010, p. 1919-1923.

[171] B. SAUERT et P. VARY. « Near end listening enhancement: Speech intelligibility improvement in noisy environments ». In : *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. T. 1. IEEE. 2006, p. I-I.

[172] B. SAUERT et P. VARY. *Near-end listening enhancement: Theory and application*. Mainz GmbH, 2014.

[173] M. R. SCHÄDLER. « Optimization and evaluation of an intelligibility improving signal processing approach (IISPA) for the Hurricane Challenge 2.0 with FADE ». In : *Proceedings of Interspeech*. 2020.

[174] H. SCHEPKER, J. RENNIES et S. DOCLO. « Improving speech intelligibility in noise by SII-dependent preprocessing using frequency-dependent amplification and dynamic range compression. » In : *INTERSPEECH*. 2013, p. 3577-3581.

[175] S. SESHADRI, L. JUVELA, O. RÄSÄNEN et P. ALKU. « Vocal effort based speaking style conversion using vocoder features and parallel learning ». In : *IEEE Access* 7 (2019), p. 17230-17246.

[176] S. SESHADRI, L. JUVELA, J. YAMAGISHI, O. RÄSÄNEN et P. ALKU. « Cycle-consistent adversarial networks for non-parallel vocal effort based speaking style conversion ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, p. 6835-6839.

[177] S. SHARMA, R. TRIPATHY et U. SAXENA. « Critical appraisal of speech in noise tests: a systematic review and survey ». In : *International Journal of Research in Medical Sciences* 5.1 (2016), p. 13-21.

[178] H. S. SHIN, M. S. CHOI, T. KIM et H. G. KANG. « Binaural loudness based speech reinforcement with a closed-form solution ». In : *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2010, p. 4274-4277.

[179] J. W. SHIN et N. S. KIM. « Perceptual reinforcement of speech signal based on partial specific loudness ». In : *IEEE signal processing letters* 14.11 (2007), p. 887-890.

[180] S. SINGH. *Measurement procedures in speech, hearing, and language*. University Park Press, 1975.

[181] M.D. SKOWRONSKI et J.G. HARRIS. « Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments ». In : *Speech Communication* 48.5 (2006), p. 549-558.

[182] B. J. STANTON, L. H. JAMIESON et G. D. ALLEN. « Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions ». In : *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. IEEE. 1988, p. 331-334.

[183] R. STANTON, N. D. GAUBITCH, P. NAYLOR et M. BROOKES. « A Differentiable Approximation to Speech Intelligibility Index with Applications to Listening Enhancement ». In : *Audio Engineering Society Conference: 54th International Conference: Audio Forensics*. Audio Engineering Society. 2014.

[184] H. J. M. STEENEKEN et T. HOUTGAST. « A physical method for measuring speech-transmission quality ». In : *The Journal of the Acoustical Society of America* 67.1 (1980), p. 318-326.

[185] W. STYLER. « Using Praat for linguistic research ». In : *University of Colorado at Boulder Phonetics Lab* (2013).

[186] Y. STYLIANOU. « Voice transformation: a survey ». In : *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2009, p. 3585-3588.

[187] Y. STYLIANOU, O. CAPPÉ et E. MOULINES. « Continuous probabilistic transform for voice conversion ». In : *IEEE Transactions on speech and audio processing* 6.2 (1998), p. 131-142.

[188] W. V. SUMMERS, D. B. PISONI, R. H. BERNACKI, R. I. PEDLOW et M. A. STOKES. « Effects of noise on speech production: Acoustic and perceptual analyses ». In : *The Journal of the Acoustical Society of America* 84.3 (1988), p. 917-928.

[189] L. SUN, S. KANG, K. LI et H. MENG. « Voice conversion using deep bidirectional long short-term memory based recurrent neural networks ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2015, p. 4869-4873.

[190] J. SUNDBERG et M. NORDENBERG. « Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech ». In : *The Journal of the Acoustical Society of America* 120.1 (2006), p. 453-457.

[191] SYNCHROARTS. *Vocalign project*. http://www.synchroarts.com, dernière visite le 9 décembre 2020.

[192] C. H. TAAL, R. C. HENDRIKS et R. HEUSDENS. « A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure ». In : *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE. 2012, p. 4061-4064.

[193] C. H. TAAL, J. JENSEN et A. LEIJON. « On optimal linear filtering of speech for near-end listening enhancement ». In : *IEEE Signal Processing Letters* 20.3 (2013), p. 225-228.

[194] Y. TANG et M. COOKE. « Energy reallocation strategies for speech enhancement in known noise conditions ». In : *Eleventh Annual Conference of the International Speech Communication Association*. 2010.

[195] Y. TANG et M. COOKE. « Optimised spectral weightings for noise-dependent speech intelligibility enhancement ». In : *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[196] Y. TANG et M. COOKE. « Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints ». In : *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

[197] Y. TANG, M. COOKE et al. « Glimpse-Based Metrics for Predicting Speech Intelligibility in Additive Noise Conditions. » In : *Interspeech*. 2016, p. 2488-2492.

[198] C. TANTIBUNDHIT, J. R. BOSTON, C. LI, J. D. DURRANT, S. SHAIMAN, K. KOVACYK et A. EL-JAROUDI. « New signal decomposition method based speech enhancement ». In : *Signal Processing* 87.11 (2007), p. 2607-2628.

[199] J. TAO, Y. KANG et A. LI. « Prosody conversion from neutral speech to emotional speech ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (2006), p. 1145-1154.

[200] V. C. TARTTER, H. GOMES et E. LITWIN. « Some acoustic effects of listening to noise on speech production ». In : *The Journal of the Acoustical Society of America* 94.4 (1993), p. 2437-2440.

[201] M. THEUNISSEN, D. W. SWANEPOEL et J. HANEKOM. « Sentence recognition in noise: Variables in compilation and interpretation of tests ». In : *International journal of audiology* 48.11 (2009), p. 743-757.

[202] I. B. THOMAS et R. J. NIEDERJOHN. « Enhancement of speech intelligibility at high noise levels by filtering and clipping ». In : *Journal of the Audio Engineering Society* 16.4 (1968), p. 412-415.

[203] I. R. TITZE. « On the relation between subglottal pressure and fundamental frequency in phonation ». In : *The Journal of the Acoustical Society of America* 85.2 (1989), p. 901-906.

[204] T. TODA, Alan W. BLACK et K. TOKUDA. « Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 15.8 (2007), p. 2222-2235.

[205] K. TOKUDA, T. YOSHIMURA, T. MASUKO, T. KOBAYASHI et T. KITAMURA. « Speech parameter generation algorithms for HMM-based speech synthesis ». In : *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*. T. 3. IEEE. 2000, p. 1315-1318.

[206] C. TORRENCE et G. P. COMPO. « A practical guide to wavelet analysis ». In : *Bulletin of the American Meteorological society* 79.1 (1998), p. 61-78.

[207] J. P. TUBACH. *La parole et son traitement automatique*. Rapp. tech. 1989.

[208] R. M. UCHANSKI, S. S. CHOI, L. D. BRAIDA, C. M. REED et N. I. DURLACH. « Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate ». In : *Journal of Speech, Language, and Hearing Research* 39.3 (1996), p. 494-509.

[209] V. VAILLANCOURT, C. LAROCHE, C. MAYER, C. BASQUE, M. NALI, A. ERIKS-BROPHY, S. D. SOLI et C. GIGUÈRE. « Adaptation of the hint (hearing in noise test) for adult canadian francophone populations ». In : *International Journal of Audiology* 44.6 (2005), p. 358-361.

[210] H. VALBRET, E. MOULINES et J. P. TUBACH. « Voice transformation using PSOLA technique ». In : *Speech communication* 11.2-3 (1992), p. 175-187.

[211] C. VALENTINI-BOTINHAO, J. YAMAGISHI et S. KING. « Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise? » In : *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

[212] W. VERHELST et M. ROELANDS. « An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech ». In : *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*. T. 2. IEEE. 1993, p. 554-557.

[213] M. VIKTOROVITCH. *Implementation of a new metric for assessing and optimizing the speech intelligibility inside cars*. Rapp. tech. SAE Technical Paper, 2005.

[214] F. VILLAVICENCIO, A. ROBEL et X. RODET. « Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation ». In : *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. T. 1. IEEE. 2006, p. I-I.

[215] J. VILLEGAS et M. COOKE. « Maximising objective speech intelligibility by local f0 modulation ». In : *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[216] J. C. WEBSTER et R. G. KLUMPP. « Effects of Ambient Noise and Nearby Talkers on a Face-to-Face Communication Task ». In : *The Journal of the Acoustical Society of America* 34.7 (1962), p. 936-941.

[217] R. H. WILSON. « Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance ». In : *Journal of the American Academy of Audiology* 14.9 (2003), p. 453-470.

[218] B. YEGNANARAYANA, C. D'ALESSANDRO et V. DARSINOS. « An iterative algorithm for decomposition of speech signals into periodic and aperiodic components ». In : *IEEE Transactions on Speech and Audio processing* 6.1 (1998), p. 1-11.

[219] S. E. YOHO, S. A. BORRIE, T. S. BARRETT et D. B. WHITTAKER. « Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology ». In : *Attention, Perception, & Psychophysics* 81.2 (2019), p. 558-570.

[220] S.D. YOO, J.R. BOSTON, A. EL-JAROUDI, CC. LI, J.D. DURRANT, K. KOVACYK et S. SHAIMAN. « Speech signal modification to increase intelligibility in noisy environments ». In : *The Journal of the Acoustical Society of America* 122.2 (2007), p. 1138-1149.

[221] TC. ZORILA, V. KANDIA et Y. STYLIANOU. « Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression ». In : *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[222] E. ZWICKER et E. TERHARDT. « Analytical expressions for critical-band rate and critical bandwidth as a function of frequency ». In : *The Journal of the Acoustical Society of America* 68.5 (1980), p. 1523-1525.

Annexes

Annexe A

Listes des stimuli verbaux équilibrés

Liste 1 (entraînement)	Ajust. (dB)
Ce casse-tête est difficile.	-7,4
La statue s'élève sur la place.	-4,9
Elle a compté jusqu'à dix.	-4,1
Le charretier fouette son cheval.	-3,9
La soupe était délicieuse.	-3,6
Le petit garçon chante bien.	-3,0
La cage contient un oiseau.	-2,7
L'éléphant a une longue trompe.	2,6
Elle joue avec ma poupée.	2,6
Les grenouilles sont vertes.	2,9
Maman achète du pain.	2,9
Le bouffon amuse le roi.	3,0
Le hameau est loin du village.	3,2
Les grenouilles plongent dans l'eau.	3,6
Le bouillon est servi sur la table.	3,6
Les flocons de neige sont blancs.	3,7
Le clown est vraiment drôle.	4,0
Le marchand vend des bonbons.	4,6
Mes frères jouent au base-ball.	4,7
L'ours trouve du miel.	4,9

Liste 2 (entraînement)	Ajust. (dB)
Cette histoire est triste.	-6,2
L'avion a traversé le ciel.	-4,7
Ce musicien joue du piano.	-4,0
L'escroc est recherché par la police.	-3,9
Le cocher fouette son cheval.	-3,6
La soirée se passe en famille.	-2,8
Le sculpteur taille la pierre.	-2,7
Le maçon a terminé le mur.	2,6
Il m'a lancé la balle.	2,7
Il joue aux billes avec moi.	2,9
La fille lave ses mains.	2,9
Le veau grossit vite.	3,1
L'oiseau s'envole du nid.	3,4
Le canon tonne à la frontière.	3,6
L'îlot est au milieu du lac.	3,7
Il loue un film d'horreur.	3,7
Le mulot vit dans les champs.	4,6
Maman épluche une orange.	4,6
Son veston est troué.	4,7
Ce petit canard apprend à nager.	3,7

Liste 3	Ajust. (dB)
Le coq réveille le village.	0,0
Le chien dormait dehors.	2,4
II vit dans la jungle.	-0,1
Il doit prendre ses vitamines.	1,1
Les enfants courent dehors.	1,9
Le camion est rouge.	2,1
La salle était vide.	-1,7
Ce garçon pédale très vite.	-1,1
L'arbre est bien décoré.	-0,7
Il mange avec une fourchette.	0,9
L'oiseau est sur une branche.	1,5
Tous les chats sont gris.	-1,5
Elle va perdre son temps.	1,1
Le sac est plein de billes.	-0,6
La partition est sur le pupitre.	-2,5
Le berger garde les moutons.	-2,2
L'oiseau s'est enfui de sa cage.	-1,7
Le nageur a regagné la rive.	-1,0
Le pêcheur lance sa ligne.	-0,8
La prairie s'étend jusqu'au bois.	-0,7

Liste 4	Ajust. (dB)
La fenêtre est ouverte.	-1,1
Tout le monde est en classe.	0,9
Le serveur apporte la crème.	-1,4
Ils vont à la plage.	1,4
Il nage dans la rivière.	0,7
J'aime les couchers de soleil.	-0,9
Papa tirait le chariot.	0,7
Les enfants jouent dans le sable.	0,1
Il s'est perdu dans la ville.	-1,3
Elle écoute la radio.	-2,4
Elle était très patiente.	-2,1
Son sac était très lourd.	-2,0
La charrue creuse le sillon.	0,6
La banlieue entoure la ville.	0,7
Le témoin s'est rétracté devant le juge.	0,9
Le budget est en équilibre.	1,1
Le bonbon contient du sucre.	1,5
Le tombeau est surmonté d'une croix.	1,6
Le rideau est baissé à l'entracte.	1,7
Le fantôme hante le château.	1,8

Liste 5	Ajust. (dB)
L'écureuil grimpe dans l'arbre.	0,2
La marmotte creuse un trou.	0,2
Elle achète des légumes frais.	-0,3
L'homme est très poli.	1,8
Cette femme joue du piano.	-0,6
Ses cheveux sont blonds.	1,4
Son cerf-volant est jaune.	2,4
Ils étaient très malades.	-0,8
Le chien ramène le jouet.	-0,2
La souris mange du fromage.	1,0
Elle boit du jus d'orange.	-0,2
Les batteries ne fonctionnent plus.	0,6
Il mange de la crème glacée.	-0,6
Il a caché la plume.	0,4
Elle lui tire les cheveux.	0,4
Tu as caché mon jouet.	2,2
La mère berce son enfant.	-0,6
La dame a perdu son sac.	-2,5
Le champion a gagné la course.	-2,4
L'enfant fatigue ses parents.	-2,1

Liste 6	Ajust. (dB)
Elle prend un bain chaud.	-0,3
Cette église est très vieille.	-2,3
Les dragons crachent du feu.	-0,8
Ton jus est sur la table.	2,2
Ils prennent une marche.	1,6
Elle prend soin de sa mère.	0,6
Ils ont marché sur le pont.	1,8
Il mange sa soupe.	-0,4
La jeune fille se brosse les dents.	-1,2
Ils ont cassé tous les œufs.	-2,4
Le vent fait bouger les feuilles.	-0,2
Elle saute sur le trampoline.	-1,4
Ce bonbon est très sucré.	-0,1
Ils vont jouer au parc.	1,3
Elle a fait fondre de la glace.	0,9
J'ai un livre à colorier.	1,9
La fenêtre donne sur la cour.	-2,1
Le refrain est repris en chœur.	-1,4
L'écrin contient des bijoux.	-0,8
Le lapin mange de la salade.	-0,7

Liste 7	Ajust. (dB)
Il n'aime pas le brocoli.	2,1
Ils regardent le spectacle.	-1,0
Les roses blanches sont belles.	-0,3
La souris est un rongeur.	0,8
Les vacances sont finies.	-2,5
Le souper était chaud.	-1,2
J'ai peur des crocodiles.	-0,2
Notre fille se marie demain.	1,6
Le chat regarde l'oiseau.	0,2
Elle a perdu sa valise.	-0,9
Elle a fait son lit.	0,8
Il ne faut pas manger vite.	1,2
Elle porte des boucles d'oreille.	-0,3
Le groupe marchait vers le parc.	1,8
J'ai sali ma blouse.	1,1
Ma tante fait de la couture.	0,2
La route est indiquée sur la carte.	-2,3
Le bateau vogue sur la mer.	-1,8
Le docteur a ordonné un médicament.	-1,7
Le passeport n'a pas de visa.	-0,8

Liste 8	Ajust. (dB)
Le gamin est parti à l'école.	-2,2
Le départ est prévu pour demain.	-0,5
Le taureau entre dans l'arène.	0,8
Le sentier mène au bois.	1,8
Le train est entré en gare.	-0,8
La bague scintille au doigt.	-0,1
Le jardin entoure la maison.	0,1
Le soulier n'a plus de talon.	2,1
Le clairon réveille les soldats.	-0,2
Le défunt a laissé un testament.	-0,9
Le portrait est exposé au salon.	-1,6
Le coussin est sur le fauteuil.	-1,4
L'armée défend la nation.	-1,4
L'athlète entre dans le stade.	-0,5
Le charbon est extrait de la mine.	-1,4
Le boucher n'a plus de viande.	1,0
La vaisselle est sur l'évier.	-2,2
L'exploit mérite une récompense.	-0,6
La pipe est bourrée de tabac.	-0,5
La bicyclette n'a plus de roues.	-0,3

Liste 9	Ajust. (dB)
Le rabais est consenti aux acheteurs.	0,9
Le stylo est sur l'encrier.	-0,3
L'immeuble a trois étages.	-0,5
Le volcan est en éruption.	-1,4
La bouteille est à la cave.	-2,1
Le crédit est consenti par la banque.	-0,7
Le forçat s'est évadé du bagne.	-1,0
La couverture est sur le lit.	-1,4
Le jury a acquitté l'accusé.	-2,2
Le curé sort de l'église.	-1,4
L'aveu entraîne le pardon.	1,7
Le canot a chaviré dans l'estuaire.	0,2
Le maître a fini la leçon.	-1,0
La vache a regagné l'étable.	-1,3
La voiture est en panne.	-1,4
L'addition est de quatre chiffres.	-1,4
Le vaisseau glisse sur les flots.	0,2
La passion aveugle les violents.	2,0
L'heure sonne à l'horloge.	2,1
Le colonel commande le régiment.	2,5

Liste 10	Ajust. (dB)
Le flacon contient un parfum.	-1,9
Le serpent fuit sous les pierres.	-1,9
Le balcon surplombe la terrasse.	0,0
L'impôt est dû par chacun.	0,5
Le bouchon flotte sur l'eau.	1,7
La rançon est exigée par les bandits.	0,1
Le délit est passible de prison.	-1,1
La lampe est suspendue au plafond.	-1,0
Le gérant a fermé son magasin.	-2,0
Le courrier est arrivé en retard.	-1,0
Le forfait mérite un châtiment.	-1,2
Le pont est sur la rivière.	-0,7
La comédie est en deux actes.	-1,7
L'annonce est parue au journal.	2,1
L'enfant dort dans son berceau.	-0,7
Le parrain embrasse son filleul.	0,0
Le récit amuse le lecteur.	-0,7
La lettre est mise à la boîte.	0,4
Le téléphone est sur le bureau.	0,6
Le crémier vend du fromage.	0,6

ECOLE DOCTORALE

Titre: Amélioration de l'intelligibilité des signaux audio de parole en contexte bruité automobile

Mots clés: intelligibilité de la parole, parole Lombard, parole claire, renforcement de la parole, conversion de voix, apprentissage machine

Résumé: La quantité de diffusion de signaux de parole dans les habitacles automobiles est de plus en plus importante : télécommunications, radio, système de navigation... Cependant, malgré les efforts et les avancées mécaniques, beaucoup de bruits persistent au sein de l'habitacle dégradant fortement l'intelligibilité de ces signaux de parole. L'objectif de cette thèse est alors de développer des outils de renforcement de la parole visant à traiter les signaux avant leur dégradation afin d'assurer une bonne intelligibilité dans le bruit des habitacles automobiles.
Une approche de renforcement de la parole très performante consiste à utiliser un égaliseur fréquentiel afin d'optimiser un critère d'intelligibilité : le *Speech Intelligibility Index* (SII). Pour faciliter l'optimisation, les méthodes actuelles se basent sur des approximations du critère. De plus, en concentrant l'énergie spectrale du signal dans des zones où l'oreille est plus sensible, ces méthodes augmentent le volume perçu ce qui peut détériorer l'expérience utilisateur. Ainsi, en plus de proposer une méthode de résolution exacte du problème de maximisation du SII, nos travaux proposent d'introduire et étudier l'influence d'une nouvelle contrainte perceptive maintenant les signaux à leur niveau perçu.
La popularisation des approches d'apprentissage automatique pousse à apprendre les traitements de renforcement de la parole à partir d'exemples naturellement produits dans le bruit (parole Lombard), ou en sur-articulant (parole claire). Les travaux actuels ne parviennent pas à obtenir des gains d'intelligibilité aussi significatifs qu'avec les modifications naturelles et nous pensons que la négligence de nombreux aspects temporels pourrait en être partiellement responsable. Nos travaux proposent donc d'approfondir ces approches en exploitant des modèles d'apprentissage et des pré-traitements adaptés aux séquences temporelles longues. Nous proposons aussi une nouvelle modélisation des modifications du débit de la parole directement intégrable dans l'apprentissage machine ce qui n'avait jamais été fait auparavant.

Title: Speech intelligibility enhancement for in-car applications

Keywords: speech intelligibility, Lombard speech, clear speech, speech reinforcement, near-end listening enhancement, voice conversion, machine learning

Abstract: Speech is nowadays present in a number of in-car applications ranging from hands-free communications, radio programs to speech synthesis messages from the various car devices. However, despite the steady car manufacturing progress, significant noise still remains in the car interior that leads to a loss of intelligibility of speech signals. The PhD work aims at developping speech reinforcement tools in order to process the signals before they are played in a noisy in-car environment.
A highly effective speech reinforcement approach is to use a frequency equalizer to optimize an intelligibility criterion : the Speech Intelligibility Index (SII). To facilitate optimization, current methods are based on approximations of the criterion. In addition, by concentrating the spectral energy of the signal in areas where the ear is more sensitive, these methods increase the perceived volume which can deteriorate the user experience. Thus, in addition to proposing an exact method of solving the SII maximization problem, our work proposes to introduce and study the influence of a new perceptual constraint in order to maintain the signals at their perceived level.
The popularization of machine learning approaches pushes to learn speech reinforcement processings from examples naturally produced in noise (Lombard speech), or by over-articulation (clear speech). Current work fails to achieve intelligibility gains as significant as with natural modification, and we believe that the many temporal aspects neglect may be partially responsible. Our work therefore proposes to deepen these approaches by exploiting learning models and pre-processings adapted to long duration sequences. We also propose a new modeling of the speech rate modifications that directly fits in the machine learning model which had never been done before.

Institut Polytechnique de Paris
91120 Palaiseau, France