Conversion de voix pour la synthèse de la parole

Taoufik En-Najjary

N° d'ordre : 3166

THÈSE

présentée devant

l'Université de Rennes I

pour obtenir le grade de:

DOCTEUR DE L'UNIVERSITE DE RENNES 1
Mention: Traitement du signal et Télécommunications

par

Taoufik En-najjary
École doctorale : MATISSE

Conversion de voix pour la synthèse de la parole

Soutenue le 8 Avril 2005 devant la commission d'examen

M. :	Gérard	FAUCON	Président
MM. :	Geneviève	BAUDOIN	Rapporteurs
	Yannis	STYLIANOU	
MM. :	Olivier	BOEFFARD	Examinateurs
	Olivier	ROSEC	
M. :	Thierry	CHONAVEL	Directeur

Table des matières

Table des figures

Liste des tableaux

Remerciements

Je tiens à remercier, en tout premier lieu, Mr Thierry Chonavel, Professeur à l'ENST de Bretagne, et Mr Olivier Rosec, Ingénieur de recherche et développement à la division R&D de France Télécom, qui ont dirigé cette thèse. Tout au long de ces trois années, Mr Olivier Rosec a su orienter mes recherches aux bons moments grâce à ses compétences scientifiques, tout en tirant partie de ma formation. Je voudrais le remercier aussi pour le temps et la patience qu'il m'a accordés tout au long de ces années. De plus, les conseils qu'il m'a divulgués tout au long de la rédaction, ont toujours été clair et succinct, me facilitant grandement la tâche et me permettant d'aboutir à la production de cette thèse. Merci à lui, ainsi qu'à Mr Thierry Chonavel, pour leurs précieux conseils. Mes plus sincères remerciements vont également à Mr Thierry Moudenc, responsable de l'unité de recherche et développement "Vocalisation Multimodale Innovante" à France Télécom, qui m'a chaleureusement accueilli dans son équipe. Ses conseils et ses commentaires ont été fort utiles. Je le remercie également pour avoir relu attentivement ce manuscrit.

Je remercie Mr Gérard Faucon, Professeur à l'Université de Rennes 1, qui m'a fait l'honneur de présider la commission d'examen. Je remercie tout particulièrement Mr Yannis Stylianou, professeur associé à l'Université de Crète, et Mme Genvième Baudoin, Professeur à l'ESIEE, d'avoir bien voulu accepter la charge de rapporteur. Je remercie également Mr Olivier Boëffard, Maître de conférence à l'ENSSAT, d'avoir bien voulu examiner ce travail.

Je souhaite aussi remercier mes amis, en particulier Meryem, Salma, Mehand, Yassine, Mohammed, Jamal , Soufiane et Damien, témoins de mes joies, de mes fatigues, de mes enthousiasmes et de mes hauts et bas, qui m'ont soutenu durant ces années.

Pour finir ("last but not least"), ma gratitude est addressée à ma famille dont les encouragements et la générosité sont inestimables. Mes plus chaleureux remerciements vont a Samar, mon épouse, qui a su partager ma joie lorsqu'un résultat s'est avéré bon et qui a su me réconforter dans le cas contraire.

Abréviations

AR :	Autoregressive
CE :	Conditional Expectation
DFW :	Dynamic Frequency Warping
DTW :	Dynamic Time Warping
EM :	Expectation Maximisation
GMM :	Gaussian Mixture Model
HMM :	Hidden Markov Model
HNM :	Harmonic plus Noise Model
LMR :	Linear Multivariate Regression
LPC :	Linear Prediction Coefficients
MOS :	Mean Opinion Score
PARCOR :	Partial Correlation
PSOLA :	Pitch-Synchronous Overlap-Add
TD-PSOLA :	Time Domaine Pitch-Synchronous Overlap-Add
TTS :	Text-To-Speech
VC :	Voice Conversion
VQ :	Vector Quantization

Résumé

Cette thèse s'inscrit dans le cadre des travaux de recherche entrepris par la division R&D de France Telecom dans le domaine de la synthèse de la parole à partir du texte. Elle concerne plus particulièrement le domaine de la conversion de voix, technologie visant à transformer le signal de parole d'un locuteur de référence dit locuteur source, de telle façon qu'il semble, à l'écoute, avoir été prononcé par un autre locuteur cible, identifié au préalable, dit locuteur cible. Le but de cette thèse est donc la diversification de voix de synthèse via la conception et le développement d'un système de conversion de voix de haute qualité.

Les approches étudiées dans cette thèse se basent sur des techniques de classification par GMM (Gaussian Mixture Model) et une modélisation du signal de parole par HNM (Harmonic plus Noise Model). Dans un premier temps, l'influence de la paramétrisation spectrale sur la performance de conversion de voix par GMM est analysée. Puis, la dépendance entre l'enveloppe spectrale et la fréquence fondamentale est mise en évidence. Deux méthodes de conversion exploitant cette dépendance sont alors proposées et évaluées favorablement par rapport à l'état de l'art existant.

Les problèmes liés à la mise en oeuvre de la conversion de voix sont également abordés. Le premier problème est la complexité élevée du processus de conversion par rapport au processus de synthèse lui-même (entre 1,5 et 2 fois le coût de calcul de la synthèse elle-même). Pour cela, une technique de conversion a été développée et conduit à une réduction de la complexité d'un facteur compris entre 45 et 130. Le deuxième problème concerne la mise en oeuvre de la conversion de voix lorsque les corpus d'apprentissage source et cible sont différents. Une méthodologie a ainsi été

proposée rendant possible l'apprentissage de la fonction de transformation à partir d'enregistrements quelconques.

Abstract

This thesis lies within the scope of the research tasks undertaken by division R&D of France Telecom in the text-to-speech synthesis field. More particularly, it relates to the field of voice conversion, a technology aiming at modifying a source speaker's speech so that it is perceived as another speaker had uttered it. The aim of this thesis is thus the diversification of synthesis voice via the design and the development of a high quality voice conversion system.

The approaches studied in this thesis are based on GMM classification techniques and HNM modeling of speech signal. First, the influence of the spectral features coding on the GMM-based voice conversion performance is analyzed. Then, the dependence between the spectral envelope and the fundamental frequency is highlighted. Two voice conversion methods exploiting this dependence are proposed and then evaluated favorably compared to the existing state of the art.

Problems related to the implementation of the voice conversion system are also tackled. The first problem is the high complexity of the voice conversion process compared to the synthesis process itself (the conversion task costs between 1.5 and 2 times more than the synthesis task itself). For that, a simplified GMM-based voice conversion procedure was presented, which enables reducing the conversion complexity by a factor between 45 and 130. The second problem relates to the learning of voice conversion function when the source and target training corpus are different. A method making possible the training of the transformation function using unspecified recordings was thus proposed.

Introduction

Les recherches entreprises ces dernières années ont permis une amélioration sensible de la qualité de la parole synthétique. De ce fait, on peut considérer que la brique technologique de synthèse de la parole a atteint un degré de maturité suffisant pour pouvoir être intégrée dans de nombreux services. Cependant, son déploiement à grande échelle exige une diversification des voix. Le but serait donc de pouvoir proposer un catalogue de voix suffisamment riche pour personnaliser les services vocaux.

Dans le même temps, la création des nouvelles voix pour un système de synthèse de la parole est une opération coûteuse. En effet, les systèmes TTS [1] actuels sont basés sur la concaténation de segments de parole sélectionnés dans une grande base de parole, dite dictionnaire acoustique. La création d'un tel dictionnaire demande un travail considérable. Dans un premier temps, il s'agit d'effectuer l'enregistrement de la base acoustique. Pour cela, un locuteur est invité à parler pendant plusieurs heures et suivant un style d'élocution bien défini. Les non-homogénéités du timbre résultant de la fatigue du locuteur ou des variations des conditions d'enregistrement lorsque plusieurs sessions d'enregistrement sont nécessaires conduisent à un style peu naturel de la parole synthétique. Puis, il s'ensuit un lourd traitement de ces enregistrements avant d'aboutir à un dictionnaire utilisable par le système de synthèse (Étiquetages phonétiques, segmentation en phonèmes et en diphones[2] etc ...) [Nef04]. Si ce traitement est en partie automatisé, une lourde phase de vérification manuelle reste absolument nécessaire, ce

[1] Text-To-Speech

[2] Le diphone peut être défini comme le segment qui s'étend de la zone stable d'une réalisation phonétique à la zone stable de la réalisation suivante et qui protège en son centre toute la zone de transition [Eme77]

qui augmente considérablement le temps de création d'une voix. Actuellement, le temps nécessaire pour la création d'une nouvelle voix est estimé à deux mois. L'enregistrement d'un nouveau corpus de parole chaque fois qu'on veut changer la voix du système est une solution irréaliste. Une autre solution est d'avoir recours à la conversion du voix.

La conversion de voix est une technique qui consiste à modifier le signal de parole d'un locuteur de référence appelé aussi locuteur source, d'une façon telle qu'il semble, à l'écoute, être prononcé par le locuteur désiré, dénommé locuteur cible. Pour cela, un apprentissage est mené sur un enregistrement restreint des locuteurs source et cible afin de déterminer une fonction de transformation qui sera ensuite appliquée au locuteur de référence et réalisera ainsi la conversion de voix [KM98a, SCM98, EnRC04a, EnRC04b]. Cette technologie, appliquée dans le cadre de la synthèse de la parole par concaténation, offre un moyen simple et rapide de diversifier les voix de synthèse en limitant les opérations d'enregistrement et surtout de vérification d'un corpus de parole entier.

La technologie de conversion de voix peut trouver d'autres applications dans de nombreux contextes tels que ceux de doublage sonore de films, de traduction automatique destinée à permettre les conversations téléphoniques entre locuteurs ne parlant pas la même langue (application de téléphonie interprétée) [YS98, ASK90] ou dans le codage de la parole à très bas débit [SNB96]. La conversion de voix pourrait également être utilisée pour évaluer la robustesse des systèmes de reconnaissance de locuteur.

Les systèmes de conversion de voix

La conversion de voix est une application qui devient envisageable du fait de la disponibilité de systèmes d'analyse et de synthèse de signaux de parole de haute qualité. En effet, elle implique des modifications fines des caractéristiques du signal de parole. Pour cela, il est nécessaire d'extraire du signal des paramètres permettant d'avoir accès aux caractéristiques spectrales et prosodiques. Inversement, il faut être capable de restituer un signal de parole de bonne qualité à partir des paramètres modifiés. Trois problèmes interdépendants principaux doivent être abordés avant de concevoir un système de conversion de voix. Premièrement, des paramètres acoustiques caractérisant l'identité du locuteur doivent être déterminés. Deuxièmement, un modèle de parole est nécessaire

pour estimer ces paramètres et générer un signal de parole à partir des paramètres transformés. Troisièmement, le type de fonction de conversion, la méthode d'apprentissage et le mode d'application de la fonction de conversion doivent être décidés.

Dans la littérature, plusieurs techniques de conversion ont été proposées, parmi lesquelles, la quantification vectorielle [Abe92], la regression linéaire multiple [Val92], les réseaux de neurones [NMRY95a], l'interpolation de locuteurs [IS94], les modèles de mélange de gaussiennes [KM98a, Sty96a] ou des technique d'adaptation au locuteur via des modèles de Markov cachés (HMM pour Hidden Markov Models) [YTM+03].

Toutes ces techniques partagent la même stratégie : estimer une fonction de transformation entre une voix source et une voix cible. Cette fonction est appliquée par la suite à la voix source pour générer de la parole convertie. L'estimation d'une telle fonction de conversion nécessite deux corpus comprenant le même contenu phonétique. L'acquisition de telles bases est très délicate dans la mesure où il faut que les deux locuteurs prononcent le même texte et que les chaînes phonétiques réalisées soient identiques. Afin de remédier à ce problème, deux solutions se présentent : reformuler le problème de conversion de voix de telle façon qu'il puisse être résolu en utilisant des corpus non parallèles, ou rendre artificiellement parallèles deux corpus par de la synthèse de la parole.

En outre, la plupart de ces méthodes traitent essentiellement de la modification de l'enveloppe spectrale. Celles-ci sont bien entendu utiles, dans la mesure où elles permettent d'approcher le timbre du locuteur cible. Mais pour simuler correctement la voix d'un locuteur, cela reste insuffisant [KS95]. D'autres paramètres jugés cruciaux sur le plan de la perception doivent être pris en compte. Parmi eux, les paramètres prosodiques sont d'une importance toute particulière. Relativement peu de travaux de recherche ont abordé ce sujet délicat. La plupart du temps, les systèmes de conversion se contentent d'adjoindre à la transformation du timbre une mise à l'échelle du pitch. Plus précisément, cela revient à respecter à la fois le pitch moyen et la dynamique du pitch du locuteur cible.

L'objectif de cette thèse est la conception d'un système de conversion de voix permettant une transformation de l'enveloppe spectrale et de la fréquence fondamentale de

haute qualité. Nos principales contributions sur cet axe de recherche sont les suivantes :

- Etude de l'influence de la paramétrisation spectrale sur les performances d'un système de conversion de voix ;
- Etude de la corrélation entre la fréquence fondamentale et l'enveloppe spectrale, et traitement conjoint de ces informations lors de la conversion ;
- Réduction de la complexité de la conversion par GMM[3] afin de pouvoir l'intégrer dans le synthétiseur sans augmenter la complexité du système ;
- Adaptation d'un système de conversion de voix dans le cas de corpus non parallèles.

Organisation du document

Le suite du document est organisée comme suit. Dans le premier chapitre, nous introduisons les propriétés fondamentales du signal de parole, ainsi que les mécanismes de production de la parole. Nous considérons également les paramètres physiques et acoustiques caractéristiques du locuteur. Le deuxième chapitre introduit les principes d'un système de conversion de voix, et donne une description des méthodes de conversion présentes dans la littérature.

Dans le troisième chapitre, nous étudions l'influence de la modélisation spectrale sur les performances d'un système de conversion de voix. Ainsi, les paramétrisations de l'enveloppe spectrale par coefficients cepstraux et LSF[4] sont étudiées et leurs performances dans un contexte de conversion de voix sont analysées.

Dans le quatrième chapitre, nous présentons deux nouvelles techniques de transformation de la fréquence fondamentale. La première est basée sur la prédiction de la fréquence fondamentale à partir de l'enveloppe spectrale. La deuxième technique est une conversion conjointe de la fréquence fondamentale et de l'enveloppe spectrale.

Dans le cinquième chapitre, nous traitons des problèmes liés à la mise en oeuvre de la conversion de voix. Dans un premier temps, nous présentons une technique de réduction de complexité de la conversion par GMM permettant d'intégrer la fonction de

[3] Gaussian Mixture Model

[4] Line Spectral Frequency

conversion dans un système de synthèse sans augmenter la complexité du synthétiseur. Dans la deuxième partie de ce chapitre, nous présentons une technique permettant d'éviter la contrainte de parallélisme des corpus d'apprentissage via l'utilisation d'un système de synthèse par corpus.

Pour conclure ce document, nous dressons une synthèse des résultats du travail effectué, et proposons quelques perspectives de recherches.

Chapitre 1

Caractérisation de l'identité vocale

Pour développer un système de conversion de voix, il est nécessaire de connaître les paramètres acoustiques caractérisant le locuteur. Pour cela, une bonne compréhension du processus de production de la parole est nécessaire. Dans la première partie du présent chapitre, nous décrivons le processus de production de la parole ainsi que les mécanismes mis en jeu lors de la phonation. Ensuite, nous décrivons les variabilités inter-locuteurs, ainsi que les paramètres acoustiques servant à la discrimination des locuteurs par des humains.

1.1 Mécanismes de production de la parole

Le processus de production de la parole est un mécanisme très complexe qui repose sur une interaction entre les systèmes neurologique et physiologique. La parole commence par une activité neurologique. Après que soient survenues l'idée et la volonté de parler, le cerveau dirige les opérations relatives à la mise en action des organes phonatoires. Le fonctionnement de ces organes est bien, quant à lui, de nature physiologique.

Une grande quantité d'organes et de muscles entrent en jeu dans la production des sons des langues naturelles. Le fonctionnement de l'appareil phonatoire humain repose

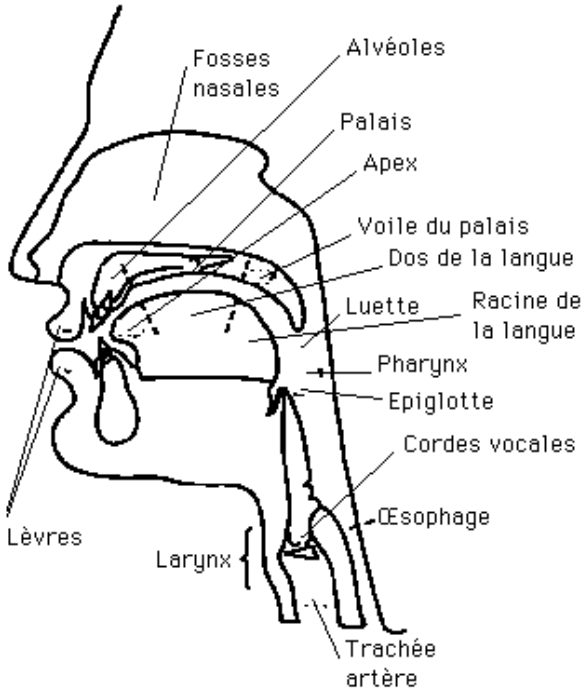


Figure 1.1 – Organes de production de la parole [PX03].

sur l'interaction entre trois entités : les poumons, le larynx, et le conduit vocal.

La figure 1.1 représente une vue globale de l'appareil de production de la parole. Le larynx est une structure cartilagineuse qui a notamment comme fonction de réguler le débit d'air via le mouvement des cordes vocales. Le conduit vocal s'étend des cordes vocales jusqu'aux lèvres dans sa partie buccale et jusqu'aux narines dans sa partie nasale.

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. L'air des poumons est comprimé par l'action du diaphragme. Cet air sous pression arrive ensuite au niveau des cordes vocales. Si les cordes sont écartées, l'air passe librement et permet la production de bruit. Si elles sont fermées, la pression peut les mettre en vibration et l'on obtient un son quasi-périodique dont la fréquence fondamentale correspond généralement à la hauteur de la voix perçue. L'air mis ou non en vibration poursuit son chemin à travers le conduit vocal et se propage ensuite dans l'atmosphère. La forme de ce conduit, déterminée par la position des articulateurs tels que la langue, la mâchoire, les lèvres ou le voile du

palais, détermine le timbre des différents sons de la parole. Le conduit vocal est ainsi considéré comme un filtre pour les différentes sources de production de parole telles que les vibrations des cordes vocales ou les turbulences engendrées par le passage de l'air à travers les constrictions du conduit vocal.

Le son résultant peut être classé comme voisé ou non voisé selon que l'air émis a fait vibrer les cordes vocales ou non. Dans le cas des sons voisés, la fréquence de vibration des cordes vocales, dite fréquence fondamentale ou pitch, noté F_0, s'étend généralement de 70 à 400 hertz. L'évolution de la fréquence fondamentale détermine la mélodie de la parole. Son étendue dépend des locuteurs, de leurs habitudes mais aussi de leurs états physique et mental.

Un exemple de signal de parole correspondant à la prononciation du mot /sa/ est donné à la figure 1.2. Le son /sa/ est représenté dans le domaine temporel, la première partie (de 0 à 80 ms) est non voisée, c'est un signal non périodique de faible énergie. La dernière partie représente un signal quasi-périodique avec une énergie plus grande, et est donc voisée.

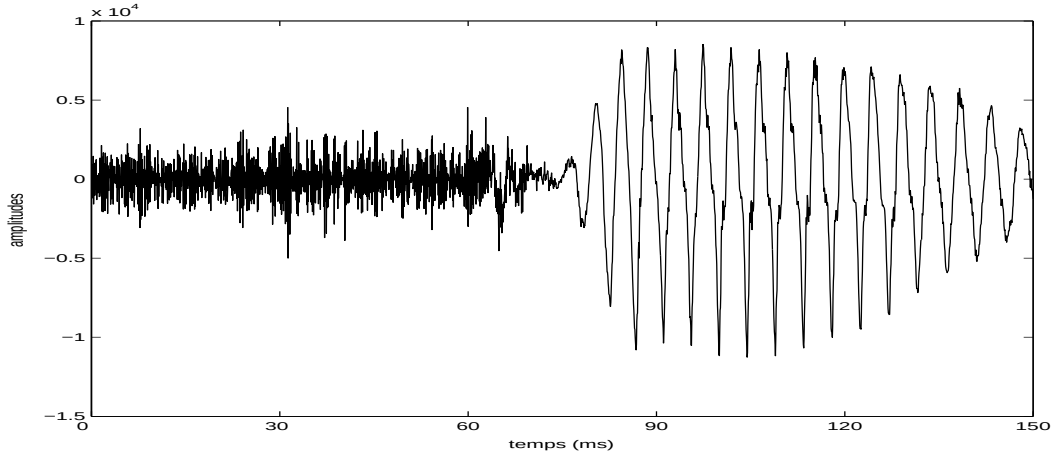


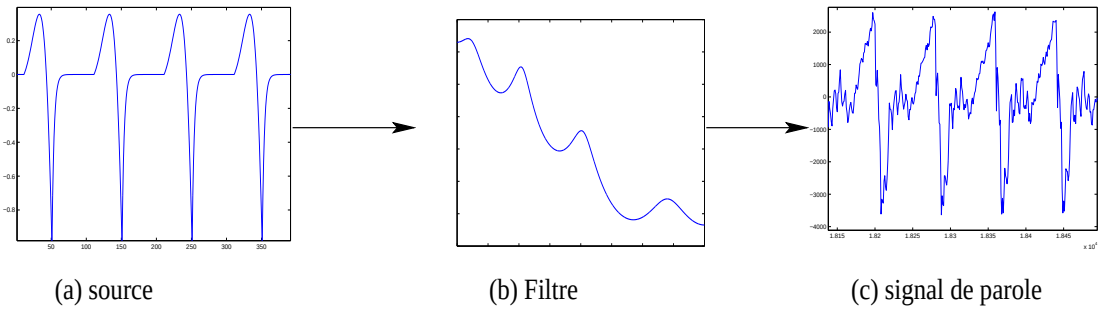
Figure 1.2 – Exemple de signal de parole (son /sa/).

Le processus de production de la parole peut être représenté par le modèle source-filtre (Figure 1.3(b)). Le signal de parole est modélisé comme la sortie d'un filtre linéaire variant dans le temps, qui simule les caractéristiques spectrales de la fonction de trans-

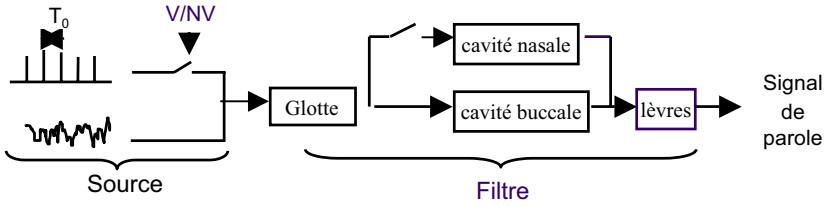
fert du conduit vocal, excité par un signal source qui reflète l'activité des cordes vocales dans les zones voisées et le bruit de friction dans les zones non voisées. Quoique simpliste, cette réprésentation est capable de décrire la majorité de phénomènes de la parole et a été à la base de nombreux codeurs et synthétiseurs de parole.

La décomposition source/filtre est une théorie particulièrement bien adaptée au problème de la conversion de voix. Transformer les paramètres de filtre revient à simuler la modification des caractéristiques du conduit vocal alors que la modification des paramètres du signal source simule les changements de la prosodie et des caractéristiques du signal d'excitation glottique. Des travaux de recherche [KK90, FLL85] ont permis d'apporter des informations a priori sur la forme du signal d'excitation glottique dans le cas des sons voisés. Ces études ont abouti à une modélisation théorique du signal glottique par un ensemble de paramètres pertinents : fréquence fondamentale, quotient d'ouverture, bruit de friction, etc... Cependant, l'extraction des paramètres pertinents du signal glottique reste un problème épineux. C'est d'ailleurs le manque de robustesse de ces techniques de déconvolution source-filtre qui fait que le signal glottique est encore peu utilisé tel quel en conversion de voix.

Une approximation classiquement employée consiste à considérer que le signal de source est constitué d'impulsions générées aux instants de fermeture de la glotte auxquelles s'ajoute un bruit blanc. Dans un tel modèle présenté en figure 1.3(a), le spectre de la partie "filtre" appelée aussi enveloppe spectrale est composée du spectre du filtre décrivant le conduit vocal auquel s'ajoute la partie lisse du spectre glottique. Suivant le modèle du signal glottique utilisé [KK90, FLL85], cette partie lisse du spectre du signal glottique peut être modélisée par un modèle AR d'ordre 2 ou 4. Certaines caractéristiques de ce modèle AR telles que la position du formant glottique et la pente spectrale sont d'ailleurs utilisées pour caractériser la qualité vocale du signal de parole [Hen01]. La partie "filtre" ainsi modélisée est porteuse des informations relatives à "l'empreinte" vocale d'un locuteur, c'est pourquoi elle est également dénommée timbre.



(a) Modèle-source filtre équivalent (Sons voisés)



(b) Modèle source-filtre

Figure 1.3 – Production de la parole : le modèle source-filtre.

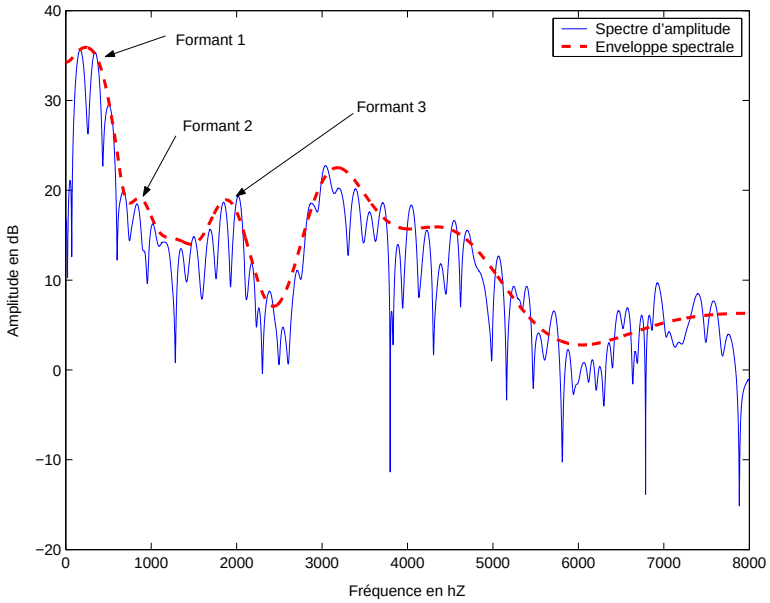


Figure 1.4 – Enveloppe spectrale.

1.2 Variabilité inter-locuteurs

Les signaux de parole véhiculent plusieurs types d'informations. Parmi eux, la signification du message prononcé est d'importance primordiale. Cependant, d'autres informations tels que le style d'élocution ou l'identité du locuteur jouent un rôle important dans la communication orale. Écouter un interlocuteur permet d'avoir des indications le concernant (homme ou femme, calme ou angoissé, etc) et bien souvent de l'identifier si on l'a déjà entendu. Dans notre vie quotidienne, ces informations, sont très utiles. Elles nous permettent, par exemple, de différencier les divers messages que nous entendons selon le locuteur et leur degré d'importance. Si toutes les voix étaient perçues de la même façon, il serait par exemple impossible de suivre une émission de radio faisant participer des personnes différentes. Le but de la conversion de voix est donc de changer les paramètres caractérisant le locuteur tout en gardant les autres informations inchangées.

La grande variabilité entre les locuteurs est due, d'une part, à l'héritage linguistique et au milieu socioculturel de l'individu, et d'autre part aux différences physiologiques des organes responsables de la production vocale. L'expression acoustique de ces différences peut être traduite par une variation de la hauteur mélodique (variation de la fréquence fondamentale), dans l'échelle des formants (plus haute chez les femmes et les enfants que chez les hommes) et dans le timbre de la voix (richesse en harmoniques due à la morphologie du locuteur et au mode de fermeture des cordes vocales).

Kuwabara et Sagisaka [KS95] ont décrit l'interaction entre ces paramètres en terme de logiciel (Software) et de matériel (Hardware) : les facteurs psychologiques et socioculturels viennent du système de commande des organes de la parole, qui peut être assimilé à un logiciel programmable. Les facteurs physiologiques sont essentiellement liés à la nature des organes phonatoires et sont décrits comme un matériel inchangeable. Quand une personne essaye d'imiter une autre, elle essaie plutôt de copier le "logiciel" de la cible. D'après cette étude, il apparaîtrait que les paramètres caractéristiques d'un locuteur viennent davantage de mécanismes de commande des organes de la parole que des organes eux-mêmes.

1.2.1 Caractéristiques non anatomiques

Dans cette catégorie, on peut citer tous les éléments non liés à l'anatomie de l'individu et qui jouent un rôle influant sur la perception de la voix. On trouve par exemple le contexte de l'interaction, la syntaxe, l'intonation, l'accent régional, la façon d'articuler, le lieu géographique, etc.

Le style d'élocution, qu'un individu acquiert en grandissant, est particulièrement sensible au milieu socio-culturel et dialectal de cet individu (l'enfant tente d'imiter les sons prononcés par les adultes qui l'entourent : la famille, l'école, les voisins, etc). Ceci se traduit par la variation d'un individu à l'autre de certains paramètres comme le débit d'élocution, les durées syllabiques, les intonations. Certaines personnes ont tendance à nasaliser tous les sons, d'autres produisent des explosions particulièrement fortes dans les occlusives sourdes. Ces variations dépendent habituellement de la communauté à laquelle le locuteur appartient mais aussi des facteurs anatomiques tels que l'âge et le sexe.

Le style d'élocution est généralement décrit par le biais des paramètres prosodiques tels que les contours de la fréquence fondamentale, la durée des mots et des pauses, les contours d'énergie. Dans beaucoup de cas, il s'avère également que la signification du message comme l'intention du locuteur ont une influence forte sur l'identité de la voix perçue.

1.2.2 Caractéristiques anatomiques

Comme nous l'avons évoqué précédemment, le système de production de la parole se décompose en trois entités : les poumons qui génèrent l'air et fournissent l'énergie nécessaire à la génération de l'excitation glottique, le larynx qui module ce flux d'air et génère la source quasi-périodique, le conduit vocal qui joue le rôle d'un résonateur et filtre le signal d'excitation.

Le volume d'air généré par les poumons est directement lié à l'énergie estimée sur de courts intervalles à partir du signal acoustique. Le volume d'air dans les poumons dépend essentiellement des dimensions du thorax, il varie donc avec le sexe et l'âge de

l'individu. La source acoustique des sons voisés est très sensible aux caractéristiques des cordes vocales telles que leur élasticité, leur masse et leur forme. Toute variation de l'un de ces paramètres engendre des modifications de la périodicité et de la forme de l'onde glottique. Ainsi, les fluctuations de l'ouverture de la glotte peuvent entraîner des variations des fréquences des formants des sons voisés. Le conduit vocal est aussi un facteur important de variabilité inter-locuteurs du signal acoustique, dans la mesure où, par exemple, les différences de taille du conduit vocal induisent des modifications des fréquences formantiques des voyelles.

1.2.3 Caractéristiques acoustiques de l'identité vocale

L'étude des caractéristiques acoustiques de la voix d'un individu reste une tâche ambitieuse. Nombreux sont les chercheurs qui se sont penchés sur ce problème sans parvenir jusqu'à présent à déterminer les paramètres qui définissent l'identité vocale d'un locuteur. Pourtant l'enjeu est important, tant en synthèse de parole - conversion de voix - qu'en identification de locuteur, voire en reconnaissance de la parole. Le paragraphe suivant présente succinctement les observations et les résultats de recherches dans ce domaine.

Le problème de la conversion de voix est très proche de celui de l'identification de locuteur. Dans ce dernier domaine, la recherche de paramètres caractérisant le locuteur a une longue histoire. Des études en psychologie et en phonétique ont mis en évidence les relations entre les paramètres acoustiques d'un locuteur et son âge, son sexe, sa taille, son poids et d'autres propriétés physiques [SR68, HD76, LB78]. Matsumoto et al. [MHSN73] ont conclu que la fréquence fondamentale est le paramètre le plus important en terme d'identité vocale. Par contre, Itaho et Saito [IS88] ont trouvé que l'enveloppe spectrale a plus d'influence sur l'identité vocale du locuteur que la fréquence fondamentale.

La plupart des techniques courantes de reconnaissance de locuteur sont basées sur la caractérisation de la distribution statistique des enveloppes spectrales [Dod85, RSb, GS94]. Il apparaît qu'on peut efficacement distinguer des locuteurs par la comparaison de leurs enveloppes spectrales respectives [IS88, Fur86]. On admet généralement que

la forme globale de l'enveloppe aussi bien que les caractéristiques des formants sont les principales caractéristiques de l'enveloppe spectrale servant à l'identification du locuteur [KS95, Hol90, Gol75].

Discrimination des locuteurs par des humains

Si en reconnaissance du locuteur les chercheurs s'intéressent aux paramètres acoustiques permettant à un ordinateur de distinguer entre différents locuteurs, l'enjeu en conversion de voix est différent. Les liens avec la perception sont beaucoup plus ténus.

Matsumoto et al. [MHSN73] ont étudié la corrélation entre les scores de reconnaissance du locuteur et certains paramètres acoustiques tels que les fréquences des trois premiers formants, la pente du spectre de la source glottique, le pitch moyen, et les fluctuations rapides de la période de pitch. Leur expérience a été réalisé sur des séries de 24 voyelles japonaises prononcées par 8 locuteurs masculins et dont les signaux de parole ont été modifiés à l'aide d'un modèle d'analyse-synthèse. Le test consistait à faire écouter à des auditeurs deux échantillons de parole à la fois et les inviter à indiquer s'ils pensent que les deux échantillons ont été prononcés par le même locuteur ou par deux locuteurs différents. Les auteurs ont conclu que la fréquence fondamentale moyenne seule conduit à 55% de taux de reconnaissance du locuteur. En ajoutant la pente moyenne du spectre de la source glottique et la fluctuation de la période de pitch ce taux de reconnaissance passe à 63.8%. Si d'autre part, les fréquences des trois premiers formants sont ajoutées à la fréquence fondamentale moyenne seule, ce taux atteint 69,3%. Ainsi, ils ont conclu que la fréquence fondamentale moyenne joue un rôle important dans la perception de l'identité vocale, et que la contribution relative des caractéristiques du conduit vocal à l'identité du locuteur est plus grande que celle des caractéristiques de la source glottique. Tous ces paramètres ensemble contribuent à 86 % de taux identification du locuteur par l'audition.

Dans une étude similaire, Itoh et Saito [IS88] ont abouti à une conclusion différente. Les auteurs ont utilisé des phrases courtes et des voyelles en japonais prononcées par des locuteurs connus pour les auditeurs qui ont participé au test. Puis, en utilisant un système LPC d'analyse-synthèse, ils ont créé des nouveaux sons en manipulant certains

paramètres (enveloppe spectrale, fréquence fondamentale, énergie). À l'aide d'un test ABX (un test dans lequel des auditeurs jugent si le son prononcé par un locuteur "X" est plus proche du locuteur "A" ou de "B") les auteurs ont conclu que l'enveloppe spectrale du signal de parole est le facteur le plus déterminant en terme d'identification du locuteur par l'audition.

Dans une expérience menée par Van Lancker et al. [Lan85] sur l'identification des voix familières, des enregistrements de voix célèbres (connues des auditeurs) ont été présentés dans un test où le signal de parole était joué normalement, ou en inversant le signal. En inversant le signal certaines voix sont devenues presque méconnaissables, alors que d'autres étaient presque aussi bien identifiées que les mêmes voix jouées normalement. Les auteurs ont conclu que différents paramètres acoustiques interviennent lors de l'identification de différentes voix, et que l'ensemble de paramètres utiles n'est pas le même pour toutes les voix. D'autres expériences avec des voix dont le débit de la parole a été modifié mènent à la même conclusion.

1.3 Conclusion

A partir de l'analyse précédente, il n'est pas évident d'identifier les paramètres acoustiques qui jouent un rôle décisif sur la caractérisation de l'identité vocale. Comme l'a mentionné Kawabara dans [KS95], l'identité vocale est un amalgame entre divers paramètres acoustiques dont le degré et l'ordre d'importance diffèrent d'un individu à l'autre. Ces paramètres peuvent être classés en trois catégories selon les niveaux d'observation adoptés.

Niveau segmental

Au niveau segmental, les caractéristiques du signal de parole sont analysée à l'échelle centi-seconde. Les descripteurs acoustiques des paramètres segmentaux incluent les formants et leurs largeurs de bande, la pente spectrale, la fréquence fondamentale et l'énergie. Les paramètres segmentaux dépendent principalement des propriétés physiologiques et physiques des organes de la parole, mais également de l'état émotionnel du locuteur [KK90].

Niveau supra-segmental

Les paramètres supra-segmentaux décrivent la variation des paramètres segmentaux sur des échelles d'observation plus grande que le segment acoustique (phonémique, groupe de souffle, ...). Ils incluent les contours de pitch, les trajectoires spectrales, les durée des phonèmes, et l'évolution de l'énergie sur une expression. Ces paramètres dépendent généralement du style d'élocution du locuteur, de son état psychologique ainsi que du sens des mots utilisées [KS95]. Pour traiter les paramètres supra-segmentaux, des modèles complexes doivent être utilisés pour caractériser les paramètres prosodiques et acoustiques à l'échelle de l'énoncé pour chacun des locuteurs et d'autre part pour formaliser le lien entre ces locuteurs.

Niveau linguistiques

Au niveau linguistique, d'autres types d'informations peuvent également être considérées comme porteurs de l'identité vocale : l'univers lexical du locuteur, la phonétisation propre au locuteur ou encore d'autres caractéristiques relevant de dialectes ou d'accents régionaux. Ces phénomènes linguistiques sont au-delà de la portée de cette thèse et ne seront pas étudiés dans le présent document.

Les chercheurs ont prouvé que les paramètres segmentaux et supra-segmentaux sont perceptuellement significatifs pour l'identification du locuteur. Parmi les paramètres supra-segmentaux, les contours de la fréquence fondamentale sont d'une importance particulière. Parmi les paramètres segmentaux, les chercheurs ont considéré que l'enveloppe spectrale et notamment les paramètres de formant sont d'importance majeure. Une approche de conversion de voix tenant compte d'un ensemble plus complet de paramètres acoustiques est susceptible de surpasser toutes les approches basées sur un ensemble plus réduit de paramètres acoustiques.

Dans ce travail, nous allons nous concenter sur la transformation des paramètres segmentaux. En effet, les études effectuées dans [MHSN73, IS88] sur la discrimination des locuteurs par des humains montrent que les paramètres segmentaux jouent un rôle plus important que les paramètres supra-segmentaux dans la discrimination entre locuteurs par des humains. Les paramètres supra-segmentaux ne serons pas abordés de manière explicite.

Chapitre 2

Introduction à la conversion de voix

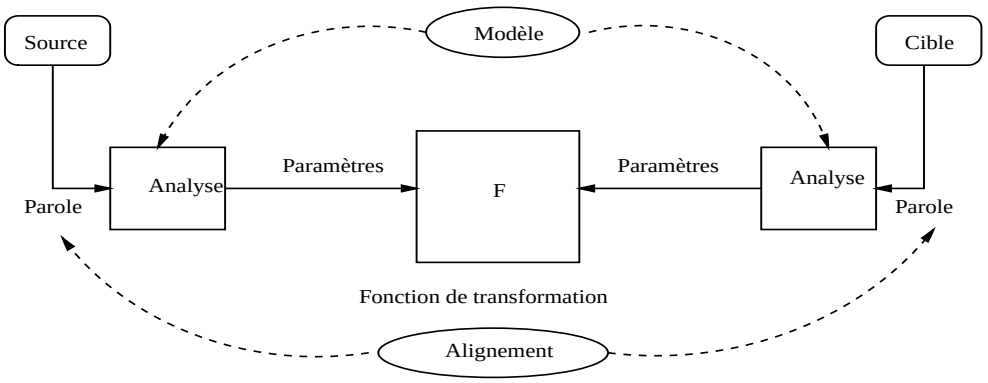
Dans la première partie de ce chapitre, nous exposons les principes de base d'un système de conversion de voix. Puis, dans la deuxième partie, nous dressons un état de l'art des techniques existant dans la littérature.

2.1 Principes d'un système de conversion de voix

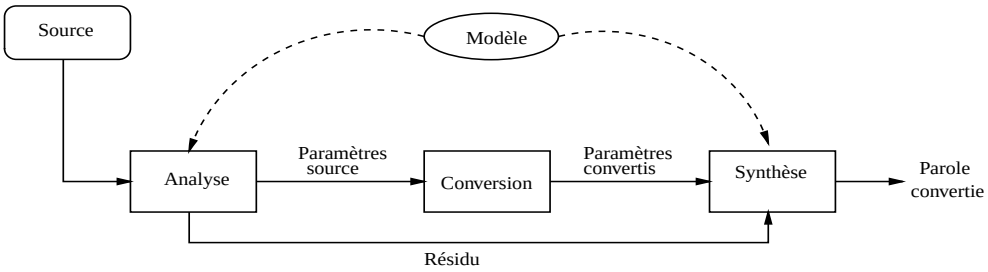
Dans la mise en oeuvre d'un système de conversion de voix, on peut donc distinguer deux phases principales. La première est une phase d'apprentissage durant laquelle des signaux de parole source et cible sont utilisés pour estimer une fonction de transformation. La deuxième est une phase de transformation durant laquelle le système utilise la fonction de transformation précédemment apprise pour transformer des nouveaux signaux de parole source d'une façon qu'ils semblent, à l'écoute, avoir été prononcés par le locuteur cible.

Phase d'apprentissage

La figure 2.1(a) présente le principe général de la phase d'apprentissage d'un système de conversion de voix. La phase d'apprentissage nécessite trois composantes principales :



(a) Phase d'apprentissage



(b) Phase de transformation

Figure 2.1 – Phases d'apprentissage et de transformation d'un système de conversion de voix.

- Deux corpus de parole source et cible comprenant le même contenu phonétique ;
- Un modèle mathématique du signal de parole pour déterminer quels paramètres seront modifiés par le système ;
- Une fonction de transformation décrivant la manière dont les paramètres sources seront modifiés.

La phase d'apprentissage commence par une étape d'analyse des corpus de parole source et cible suivant le modèle mathématique, afin d'extraire les paramètres acoustiques utiles à l'estimation de la fonction de transformation. La nature des paramètres acoustiques utilisés dépend du système de conversion. La plupart des travaux menés dans ce domaine traitent essentiellement de la transformation de l'enveloppe spectrale, modélisée généralement par une variante des coefficients de prédiction linéaire (LPC, LSF, LAR) [Abe92, Kai01, Val92], par cepstre discret [Sty96b] ou par des paramètres relatifs aux formants [MA95]. A ces modifications d'enveloppe spectrale sont

généralement associées des modifications de pitch allant d'une simple mise à l'échelle à une véritable prédiction de contours de pitch [EnRC03a].

Le but de la fonction de transformation est d'établir le lien entre les paramètres de la source et ceux de la cible. Naturellement, le style d'élocution des locuteurs source et cible ne sont pas rigoureusement identiques, ce qui se traduit par une différence entre les unités linguistiques observées (durées de phonèmes par exemple). Or, pour l'apprentissage de la fonction de transformation, les paramètres de la source et de la cible doivent être temporellement alignés de manière à décrire le même contenu phonétique. L'alignement temporel est réalisé à l'aide de l'algorithme DTW [RSa] dans la plupart des systèmes de conversion de voix présentés [EnRC04a, ANK88, VMT92, MA95, SCM95, LYC96]. Cependant, il est possible de réaliser cet alignement avec d'autres méthodes comme les chaînes de Markov cachées [Ars99]. Toutes ces techniques d'alignement opèrent essentiellement sur des paramètres d'enveloppe spectrale.

Cette base de données alignées sera, ensuite, utilisée pour l'estimation de la fonction de transformation. Dans la littérature, la fonction de transformation a été implémentée par diverses méthodes, comme la quantification vectorielle, les réseaux de neurones, l'alignement fréquentiel dynamique, et les modèle de mélange de gaussiennes. Certaines de ces méthodes seront exposées dans la section suivante.

Phase de transformation

Après avoir défini la forme de la fonction de transformation et le modèle d'analyse du signal de parole, il reste à élaborer la stratégie de conversion de voix proprement dite. Cette stratégie est commune à tous les systèmes de conversion de voix : la conversion est simulée par l'application trame par trame de la fonction de transformation à des signaux de parole source.

La figure 2.1(b) présente l'architecture de la phase de transformation. D'un point de vue général, trois étapes sont nécessaires à la conversion de voix. Tout d'abord, une analyse est menée sur chaque trame de façon à extraire les paramètres de la source. Une partie des paramètres, par exemple ceux relatifs au timbre voire au pitch, sont modifiés par le module de conversion. Ensuite, les paramètres modifiées ainsi qu'un résidu (les

paramètres non modifiés) sont transmis à un module de synthèse qui réalise ainsi la génération du signal de parole converti.

2.2 État de l'art

Dans cette section, nous décrivons les modélisations spectrales les plus utilisées avant de décrire les principales techniques de conversion présentées dans la littérature.

2.2.1 Modélisation de l'enveloppe spectrale

L'enveloppe spectrale correspond au spectre d'amplitude du filtre modélisant le conduit vocal et le spectre de la source glottique. Ce filtre contient les informations de phases et d'amplitudes des composantes spectrales du signal produit. L'enveloppe spectrale sert alors, à coder les amplitudes. Dans le cas de transformations spectrales, on choisit une enveloppe paramétrique d'ordre faible (de 16 à 20) pour manipuler les amplitudes des harmoniques.

Analyse LPC

Une modélisation de l'enveloppe spectrale largement employée est la modélisation AR. Celle-ci stipule que le conduit vocal peut être considéré comme une cavité résonnante et approché par un filtre tout pôles. Une technique bien connue pour estimer un tel filtre consiste à appliquer des algorithmes de prédiction linéaire. Ces méthodes détaillées dans [Kay88] supposent que le signal de parole résulte du passage d'un bruit blanc gaussien par un filtre AR. Mais si cette hypothèse peut être vérifiée pour les sons sourds, elle ne l'est plus dans les parties voisées. En effet, le spectre d'une trame de parole voisée s'apparente davantage à un spectre de raies qu'à un bruit blanc et dès lors la technique de prédiction linéaire classique n'est plus valide. Il est à noter que cette erreur de modélisation croît avec la valeur du pitch, dans la mesure où plus l'espacement des harmoniques est important, plus le spectre a une allure "discrète". Ceci explique pourquoi l'estimation de l'enveloppe par prédiction linéaire devient de plus en plus erratique au fur et à mesure que le pitch augmente [Val92]. Dans les cas les plus extrêmes, la prédiction linéaire se limite à associer à des harmoniques bien marquées un formant

relativement étroit, plutôt que de faire ressortir la structure formantique réelle du signal. Pour palier ce problème, El-Jaroudi et Makhoul [EJM91] proposent une version discrète de la prédiction linéaire, le but étant de contraindre le filtre AR à respecter les amplitudes des harmoniques du signal.

Étant donné un filtre AR obtenu par analyse LPC, il convient maintenant de rechercher une paramétrisation utilisable en conversion de voix. La principale qualité requise est le fait de pouvoir supporter des transformations. A ce titre, les coefficients du filtre LPC ne sont pas bien adaptés, car il est a priori difficile de contrôler l'effet d'une modification qui leur serait appliquée sur leur stabilité et sur la perception du signal produit. Parmi les nombreuses paramétrisations possibles d'un filtre AR, les coefficients LSF (Line Spectral Frequency) semblent être ceux qui offrent les meilleures propriétés d'interpolation [Ita75, PA93]. Ils ont d'ailleurs été utilisés récemment par Kain dans un but de conversion de voix [Kai01]. L'estimation de ces paramètres sera détaillée dans la section (3.3.4).

Analyse cepstrale

L'analyse cepstrale est issue de la modélisation source-filtre classiquement utilisée en traitement de la parole. L'hypothèse sous-jacente est que les variations rapides du spectre d'un signal de parole sont dues à l'excitation alors que le conduit vocal contient la partie lisse du spectre. L'utilisation du modèle source-filtre conduit à l'expression du spectre d'un signal de parole $S(\omega)$ suivante :

$$S(w) = E(\omega)H(\omega), \tag{2.1}$$

où $E(\omega)$ et $H(\omega)$ désignent respectivement les contributions de la source et du conduit vocal. En prenant le logarithme de cette expression, puis en faisant une transformée de Fourier inverse, on obtient le cepstre :

$$c = F^{-1}\left(\log|E(\omega)|\right) + F^{-1}\left(\log|H(\omega)|\right), \tag{2.2}$$

où F^{-1} désigne le transformé de fourier inverse.

D'après cette équation, en considérant uniquement les premiers coefficients de la transformée de Fourier, on peut isoler la contribution due au conduit vocal.

Notons d'emblée que contrairement à la modélisation AR précédemment décrite, l'analyse cepstrale ne fait aucune hypothèse quant à la présence de formants dans le spectre. Ceci peut être vu comme un désavantage par rapport à une analyse LPC dans la mesure où si un signal suit effectivement un modèle AR, mieux vaut a priori utiliser cette information pour l'estimation de l'enveloppe spectrale. Cependant, cette modélisation AR peut être mise en défaut dans le cas des sons nasaux. En effet, la mise en parallèle de la cavité nasale avec le conduit vocal se traduit par l'apparition de zéros dans le spectre qu'un modèle AR ne saurait mettre en évidence. Pour avoir une modélisation correcte il faudrait alors utiliser des filtres de type ARMA, mais l'estimation de tels filtres est beaucoup plus délicate. L'analyse cepstrale parvient quant à elle à faire ressortir aussi bien des formants que des anti-formants et peut de ce fait être considérée comme plus robuste.

Une des propriétés du cepstre est qu'il effectue un filtrage passe-bas du spectre du signal et tend donc à lisser les irrégularités du spectre. De ce fait, les amplitudes des harmoniques ne sont pas conservées. Pour palier ce problème et obtenir une enveloppe spectrale passant par les amplitudes des harmoniques du signal, Galas et Rodet ont proposé une méthode dite du cepstre discret [GR91]. L'idée est simplement d'introduire des contraintes sur le cepstre afin que l'enveloppe estimée respecte les amplitudes des harmoniques. Cappé et al. [CLM95] ont amélioré cette technique en la rendant plus robuste par l'introduction d'un terme de régularisation privilégiant l'obtention d'une enveloppe spectrale suffisamment lisse.

2.2.2 Fonctions de transformation

Transformation par quantification vectorielle

Proposée par Abe [Abe92], cette technique est la première à avoir été utilisée en conversion de voix. Elle tire son nom du fait que les espaces acoustiques de chacun des locuteurs sont partitionnés à l'aide d'algorithmes de quantification vectorielle. Pendant la phase d'apprentissage, il s'agit de formaliser les relations entre les différentes classes sources et cibles. Pour cela, après un alignement par DTW des trames des deux enregistrements d'apprentissage, on calcule pour chaque centroïde source C_i^s, et cible C_j^c les

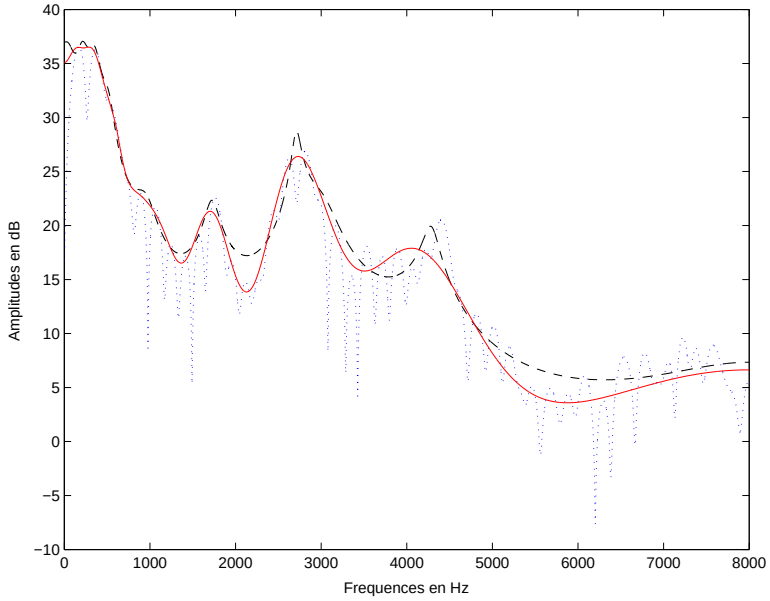


Figure 2.2 – Comparaison entre la modélisation cepstrale (ligne continue) et la modélisation AR (ligne discontinue).

probabilités $P_{ij} = P(C_j^c|C_i^s)$. Ces poids permettent ensuite de définir un nouveau dictionnaire quantifié appelé dictionnaire de "mapping" de même taille que le dictionnaire source et dont les éléments C_i^t sont :

$$C_i^t = \sum_{i=1}^{Q_c} P_{ij} C_j^c \ . \tag{2.3}$$

Lors de la transformation, il suffit alors d'associer à chaque trame d'entrée codée par C_i^s son homologue dans le dictionnaire transformée C_i^t.

Cette méthode a l'avantage d'être simple et peu coûteuse en temps de calcul. Mais son principal inconvénient est qu'elle n'offre qu'une représentation discrète de l'espace transformé. De ce fait, lorsqu'il se produit une transition entre vecteurs d'entrée, cette transition se répercute automatiquement sur sa transformée, ce qui provoque l'apparition, lors de la conversion, de discontinuités spectrales dues à la quantification.

Une manière de résoudre ce problème est d'utiliser la quantification vectorielle floue (fuzzy VQ). L'idée est de lisser les discontinuités spectrales en associant à chaque trame

d'entrée X non pas un seul centroïde source, mais une combinaison linéaire de ces derniers, dépendant de X. Testée dans le cadre de la conversion de voix par Nakamura et Shikano [NS89], cette technique adoucit effectivement les transitions entre enveloppes spectrales converties.

Régression linéaire multiple (LMR)

Dans les techniques de conversion par quantification vectorielle présentée ci-dessus, les transformations sont appliquées sur les centroïdes de l'espace source soit par simple mapping (VQ) soit par combinaison linéaire de centroïdes (Fuzzy VQ). Il en résulte que les paramètres spectraux convertis ne dépendent pas directement des valeurs des paramètres d'entrée, mais plutôt du ou des centroïdes les plus proches.

Dans [Val92], Valbret et al. prennent explicitement en compte la valeur des paramètres d'entrée pour calculer la fonction de transformation. La première étape de l'algorithme consiste à partitionner l'espace acoustique des deux locuteurs à l'aide d'algorithme de quantification vectorielle, puis à modéliser la fonction de conversion au sein de chaque classe par une simple transformation linéaire. Cela revient à estimer, pour chaque classe q, une matrice P_q par minimisation du critère des moindres carrées.

$$P_q = \arg\min \sum_{n=1}^{M_q} \|(y_{n,q} - C_q^c) - P_q(x_{n,q} - C_q^s)\|^2, \tag{2.4}$$

où C_q^s et C_q^c sont les centroïdes de la $q^{\text{ème}}$ classe source et cible respectivement, M_q le nombre de vecteurs dans la classe q, $x_{n,q}$, et $y_{n,q}$ $n = 1, \cdots, M_q$ étant les vecteurs source et cible appartenant à la classe q. Pour tout vecteur source $x_{n,q}$ appartenant à la classe q, le vecteur transformé est donné par :

$$\hat{y}_{n,q} = C_q^c + P_q(x_{n,q} - C_p^s). \tag{2.5}$$

Cette technique parvient à déplacer les formants de leur position initiale vers leur position dans l'espace cible, mais ne préserve pas bien leur amplitude ni leur bande passante. De ce fait, le timbre perçu reste éloigné du timbre de la cible. De plus, la classification "dure" effectuée introduit des discontinuités gênantes.

Conversion de voix par interpolation de locuteurs

Cette méthode est issue des travaux de recherche en reconnaissance de la parole. Les systèmes de reconnaissance actuels utilisent généralement des modèles multi-locuteurs permettant d'obtenir une couverture acoustique suffisante. Par suffisante, on entend assez importante pour que, à partir d'un nombre restreint de locuteurs, on puisse fournir un modèle adapté à tout nouveau locuteur. En conversion de voix, cette interpolation revient à rechercher pour chaque trame i un vecteur de la forme :

$$\hat{S}_i^t = \sum_{k=1}^{M} \omega_k S_i^k, \tag{2.6}$$

où M est le nombre total de locuteurs disponibles, S_i^k le vecteur acoustique du locuteur k (après alignement par DTW) et où les poids ω_k vérifient :

$$\sum_{k=1}^{M} \omega_k = 1. \tag{2.7}$$

Ainsi, lors de la phase d'apprentissage, on cherche à déterminer les poids ω_k en minimisant le critère suivant :

$$E = \sum_i \|\hat{S}_i^t - S_i^t\|^2. \tag{2.8}$$

Cette technique a été testée en conversion de voix, mais les essais effectués par Iwahashi et Sagisaka [IS94] ne sont pas très concluants. Il semble que pour pouvoir fonctionner correctement, cette méthode nécessite au préalable des modèles multi-locuteurs robustes. Son champ d'application n'est donc pas exactement celui de la conversion de voix pour la synthèse de la parole. En effet, nous disposons actuellement de peu de voix et de relativement peu de matière acoustique pour chaque voix. Cependant, elle peut être un moyen simple à mettre en oeuvre pour enrichir un catalogue de voix relativement fourni. Un point particulièrement intéressant est que cette adaptation nécessite relativement peu de données pour apprendre les caractéristiques d'une nouvelle voix.

Alignement fréquentiel dynamique et modification des formants

L'information spectrale n'est pas uniformément répartie en fréquence. Par exemple, la structure acoustique des voyelles se caractérise principalement par ses formants. Dès

lors, les formants apparaissent comme un résumé acoustique assez facilement accessible, d'où l'intérêt de les utiliser en conversion de voix.

De nombreux travaux ont été menés pour étudier la variabilité des formants, tant d'un point de vue intra-locuteur qu'inter-locuteurs. Dans [Fan66], Fant a mis en évidence que la structure formantique varie avec la longueur du conduit vocal et la classe phonémique de la voyelle analysée. Nordström et Lindblom [NL75] émettent l'hypothèse qu'une simple homothétie de rapport inversement proportionnel à la longueur du conduit vocal permet de convertir le timbre d'une voix de femme en celui d'une voix d'homme. Ces travaux sont néanmoins contredits par ceux de Fant [Fan75] qui stipulent que seule une modification non-linéaire des fréquences formantiques permet d'assurer une conversion convenable.

Cependant, l'estimation d'une telle fonction de transformation non-linéaire reste une tâche délicate. Matsumoto et Wakita [MW79, MW86], proposent une normalisation des voyelles à l'aide de distorsions de l'axe des fréquences. Leur méthode d'alignement fréquentiel dynamique (en anglais DFW pour Dynamic Frequency Warping) est l'homologue dans le domaine fréquentiel de la DTW.

Valbret et al. [VMT92] ont repris cette technique pour effectuer l'apprentissage d'une fonction de transformation sur des vecteurs spectraux préalablement alignés par DTW. Pour chaque couple de vecteurs source/cible de la base d'apprentissage, la DFW détermine un chemin d'alignement optimal. L'intérêt de cet algorithme est qu'au sein d'une même classe acoustique (par exemple obtenue par VQ), ces chemins se superposent assez bien. Par conséquent, en les moyennant on peut en déduire une fonction de transformation valable pour chacune de ces classes acoustiques.

Les résultats obtenus montrent que la DFW parvient effectivement à effectuer des transformations locales du spectre, en modifiant par exemple la position des formants voire leur bande passante. En revanche, comme cette transformation se résume à une simple distorsion de l'axe fréquentiel, elle est incapable de modifier leur amplitude. Au final, Valbret et al. ont jugé cette méthode moins performante que la LMR.

Il est à noter que la modification des formants obtenue par la méthode précédente est implicite, c'est-à-dire qu'elle découle des propriétés de la DFW, sans qu'aucune analyse

ni modélisation des formants ne soient effectuées. Dans [NMRY95b], Narendranath se proposent de modifier explicitement les trois premiers formants. Après extraction de ces derniers, la conversion se fait par un réseau de neurones, méthode particulièrement adaptée pour caractériser des fonctions de transformation non-linéaires. Le signal est ensuite généré à l'aide d'un synthétiseur par formants.

Conversion de voix par GMM

Proposée par Stylianou [Sty96b], cette technique est devenue un standard dans le domaine de conversion de voix. Le principal intérêt de cette méthode est qu'elle utilise une modélisation probabiliste continue de l'espace acoustique. Cette classification "souple" permet de réduire sensiblement les problèmes liés aux discontinuités spectrales présents dans tous les autres algorithmes de conversion de voix comme l'ont montré les tests comparatifs menés dans [BS96].

Le modèle GMM permet une modélisation statistique efficace de l'espace acoustique d'un locuteur. Chaque classe q est alors définie par sa proportion (ou poids) α_q dans le mélange, sa moyenne μ_q et sa matrice de covariance Σ_q.

Après analyse, les paramètres du mélange de gaussiennes du locuteur source $(\alpha_q, \mu_q, \Sigma_q)$ $q = 1, \cdots, Q$ sont estimés à l'aide de l'algorithme EM [DLR77]. La fonction de transformation proposée est de la forme suivante :

$$F(x_i) = \sum_{q=1}^{Q} p(q/x_i) \left[\nu_q + \Gamma_q (\Sigma_q^{xx})^{-1} \left(x_i - \mu_q^x\right)\right]. \tag{2.9}$$

Les paramètres ν_q et Γ_q sont déterminés en minimisant la distance quadratique moyenne entre les vecteurs transformés et les vecteurs cibles donnée par :

$$E = \sum_{i=1}^{N} \|y_i - F(x_i)\|^2, \tag{2.10}$$

où x_i et y_i désignent respectivement les vecteurs source et cible préalablement alignés par un algorithme de DTW. La minimisation de ce critère des moindres carrés mis sous forme matricielle revient à l'inversion d'une matrice carré d'ordre $Q(p+1)$ où Q est le

nombre de classes et p la taille des vecteurs d'entrée, ce qui peut rapidement devenir problématique en terme de temps de calcul voire de place mémoire.

Dans [KM98b], Kain a amélioré la procédure d'apprentissage en estimant la distribution jointe des paramètres des locuteurs source et cible. Cette variante revient à estimer l'ensemble des paramètres (à savoir α_q, μ_q, Σ_q, ν_q, et Γ_q) de la fonction de transformation par un algorithme de type EM. D'un point de vue théorique les deux approches conduisent au même estimateur. Cependant, le fait d'éviter la minimisation du critère des moindres carrés présentée ci-dessus rend l'algorithme plus stable numériquement, notamment lorsque l'ordre des modèles GMM augmente.

De nombreuses techniques dérivées du modèle GMM ont également été proposées. Ainsi, dans [TSS04], Toda a combiné la transformation par GMM avec la méthode DFW. Chen [CCL03] a proposé d'associer les GMM avec une technique d'adaptation au locuteur. Pour cela, il a modélisé les paramètres de la source par un modèle GMM. La fonction de transformation proposée est de la forme :

$$F(x) = x + \sum_{i=1}^{Q} h_i(x) \left(\mu_i^y - \mu_i^x\right), \tag{2.11}$$

où $h_i(x)$ est la probabilité a posteriori que x soit généré par la $i^{\text{ème}}$ composante gaussienne. μ_i^x est la moyenne de la $i^{\text{ème}}$ composante gaussienne et μ_i^y est calculé par adaptation conformément à l'équation :

$$\mu_i^y = \frac{r}{r + \sum_{q=1}^{L} P_l(x_q)} \mu_l^x + \frac{\sum_{q=1}^{L} P_l(x_q) y_q}{r + \sum_{q=1}^{L} P_l(x_q)} \tag{2.12}$$

avec un facteur r fixé [RD00], L le nombre de vecteurs dans la base de données servant à l'adaptation. x_q et y_q sont les couples de vecteurs paramètres source et cible alignés par DTW.

2.3 Conclusion

Parmi les différentes méthodes existantes, la conversion par GMM semble être celle qui offre les meilleurs résultats. En revanche, il reste un important travail à réaliser pour

savoir quels paramètres spectraux utiliser. Par exemple, les techniques par VQ ont été testées sur les coefficients LPC, le modèle GMM sur le cepstre discret [Sty96b] et sur les paramètres LSF [Kai01]. Il est généralement reconnu que le cepstre discret fournit une paramétrisation robuste de l'enveloppe. Cependant, cette modélisation ne permet que des modifications globales de l'enveloppe spectrale. Il s'ensuit que si le partitionnement de l'espace acoustique n'a pas permis de faire ressortir des classes suffisamment homogènes, le risque est d'obtenir un lissage trop important des enveloppes converties. Les paramètres LSF sont davantage liés aux formants et peuvent par conséquent ouvrir la voie à des traitements à la fois plus locaux et sans doute mieux adaptés d'un point de vue perceptuel. Mais il n'existe pas, à ce jour, d'études profondes attestant réellement de la supériorité de l'une ou l'autre des paramétrisations.

Dans le chapitre suivant consacré à la transformation du timbre, nous comparons les performances présentées par deux modélisations spectrales, utilisant respectivement les coefficients cepstraux et les paramètres LSF, dans le cadre de la conversion de voix par GMM.

Chapitre 3

Transformation du timbre

3.1 Introduction

Les performances d'un système de conversion de voix dépendent de deux facteurs : d'une part de la nature des paramètres transformés et d'autre part de la technique de conversion utilisée. La plupart des travaux menés dans ce domaine traitent essentiellement de la transformation de l'enveloppe spectrale. Elles reposent sur une analyse préalable du signal de parole. Le but d'une telle analyse est de fournir une estimation convenable de l'enveloppe spectrale. Par convenable, nous entendons suffisamment précise pour mettre en relief les caractéristiques les plus importantes du spectre reliées au timbre.

La plupart des techniques présentées récemment tournent dans leur grande majorité autour des modèles GMM. Comme nous l'avons évoqué au chapitre 2, Stylianou a utilisé le modèle GMM pour modéliser la densité de probabilité des vecteurs spectraux de la voix source, puis a estimé une fonction de transformation par minimisation d'un critère des moindres carrés entre enveloppes spectrales cible et convertie. Les tests comparatifs menés dans [BS96] ont montré que cette approche présente des performances meilleures que celles présentées par d'autres approches telles que la Quantification Vectorielle, la régression linéaire multiple ou les réseaux de neurones. Cette technique a été améliorée dans [KM98b]. En modélisant la densité conjointe de la source et de la cible, Kain a

montré que sa technique est plus stable numériquement, notamment lorsque l'ordre des GMM augmente. En revanche, la conversion par GMM à été appliquée indépendamment à la modification du cepstre discret et à celle des paramètres LSF sans qu'aucune comparaison véritable n'ait été effectuée.

Le travail présenté dans ce chapitre est une étude de l'état de l'art de la transformation du timbre. Nous comparons, dans le cadre de la conversion par GMM, les modélisations par cepstre discret et paramètres LSF, afin de pouvoir choisir quelle paramétrisation utiliser pour la suite de ce travail. Notons que dans ce chapitre seuls la conversion des paramètres spectraux est considéré, la conversion de la fréquence fondamentale sera traité dans le chapitre suivant.

Ce chapitre est organisé comme suit. Dans les sections 3.2 et et 3.3 nous décrivons le modèle HNM ainsi que la méthode utilisée pour l'extraction des paramètres. La section 3.4 présente une description de la conversion par GMM. La section 3.5 présente la méthode utilisée pour la synthèse des paramètres transformés. Enfin, dans la section 3.6, nous présentons les résultats des expérimentations que nous avons effectuées pour comparer les performances de la conversion obtenues en utilisant les paramètres LSF et le cepstre discret.

3.2 Modélisation HNM du signal de parole

En général, le modèle mathématique utilisé dépend de l'application visée. Dans le domaine de la conversion de voix, le modèle de parole doit satisfaire au besoin de modifications spectrales et prosodiques complexes et être en mesure de produire une grande variété de voix intelligibles, aussi bien que naturelles et précises en ce qui concerne l'identité du locuteur.

Dans ce travail, nous utilisons le modèle Harmonique plus bruit, dit modèle HNM [MQ95, Sty96b]. Ce modèle est une simplification du modèle MultiBand Excitation (MBE) de Griffin et Lim [GL88]. Le modèle HNM suppose qu'un signal de parole $s(t)$ peut être décomposé en une partie harmonique $h(t)$ et une partie bruitée $b(t)$. La partie harmonique modélise la composante quasi-périodique des sons voisés du signal

de parole, la partie bruitée modélise la composante aléatoire du signal, c'est-à-dire le bruit de friction et les variations de l'excitation glottique d'une période à l'autre.

$$s(t) = h(t) + b(t), \tag{3.1}$$

avec

$$h(t) = \sum_{l=0}^{L(t)} A_l(t) \cos(2\pi t l f_0(t) + \varphi_l(t)). \tag{3.2}$$

Les paramètres $A_l(t)$, $\varphi_l(t)$ sont l'amplitudes et la phase de la $l^{\text{ème}}$ harmonique à l'instant t. $f_0(t)$ est la fréquence fondamentale à l'instant t et $L(t)$ est le nombre d'harmoniques incluses dans la partie harmonique à l'instant t. Ces paramètres sont mis à jour à des instants spécifiques t_i appelés instant d'analyse. L'intervalle entre deux instants successifs t_i et t_{i+1} est appelé trame.

Généralement on distingue deux catégories de trames de signaux de parole ; des trames voisées et des trames non voisées. Dans le cas des trames voisées, le spectre du signal est divisé en deux bandes limitées par une fréquence variant dans le temps, $F_c(t)$, dite fréquence maximale de voisement ou fréquence de coupure (figure 3.1). En-deçà de $F_c(t)$, le signal est considéré comme étant purement harmonique et représenté par la partie $h(t)$ (équation 3.2) et au-delà de $F_c(t)$ intervient uniquement une partie aléatoire correspondant au filtrage d'un bruit blanc par le conduit vocal. Pour les trames non voisées apparaît uniquement la partie non déterministe (Figure 3.2).

Le contenu fréquentiel de la partie bruitée est représenté par un modèle AR variant dans le temps. La partie bruitée $b(t)$ peut donc être obtenue en filtrant un bruit blanc gaussien $u(t)$ par un filtre tout pôle $g(t)$ et en multipliant le résultat obtenu par une enveloppe d'énergie $e(t)$.

$$b(t) = e(t) \left[g(t) * u(t) \right]. \tag{3.3}$$

Le signal synthétique $\hat{s}(t)$ est simplement obtenu par l'addition de la partie harmonique $h(t)$ et de la partie bruitée $b(t)$:

$$\hat{s}(t) = h(t) + b(t). \tag{3.4}$$

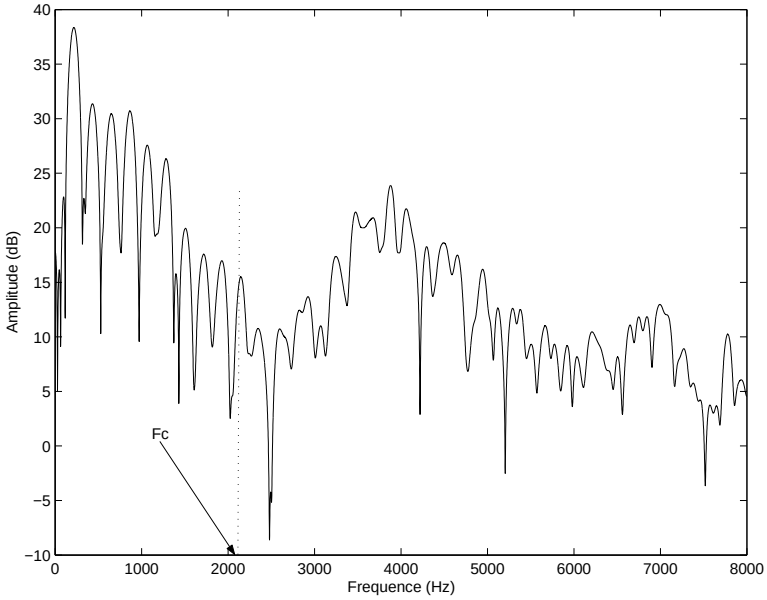


Figure 3.1 – Spectre d'un son voisé.

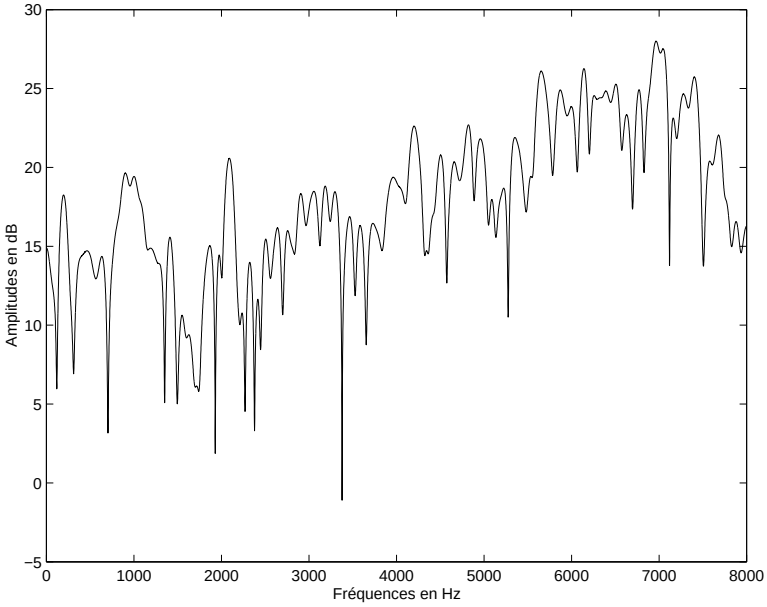


Figure 3.2 – Spectre d'un son non voisé.

L'avantage principal de ce modèle est sa maniabilité dans diverses applications. En effet, cette représentation du signal de parole peut être considérée comme un pré-encodage (le signal est modélisé par un certain nombre de paramètres), et s'adapte facilement à des procédures de codage à débit réduit [MQ95]. L'accès immédiat à des grandeurs telles que la fréquence fondamentale du signal, rend ce modèle pratique lorsqu'on souhaite opérer des transformations prosodiques (de hauteur, de durée, etc) [Sty96b].

3.3 Extraction des paramètres

Le signal de parole est un signal non-stationnaire. Pour avoir des signaux localement stationnaires, l'utilisation d'une fenêtre glissante dans le temps est donc indispensable. Cette stationnarité est assurée pour des fenêtres de durée compatible avec les ordres de grandeurs typiques de variations des paramètres acoustiques. Nous faisons dans la suite, l'hypothèse que la fréquence fondamentale et les phases sont constantes pendant la durée de la fenêtre d'analyse (le centre de la fenêtre est choisi comme origine des temps : $n = 0$). Le lecteur trouvera dans Harris [Har78] une discussion exhaustive sur l'utilisation des fenêtres dans le cadre de l'analyse harmonique. Le découpage temporel consiste alors à multiplier le signal original par la fenêtre choisie, décalée dans le temps. Comme les paramètres acoustiques évoluent rapidement, il est nécessaire d'utiliser un pas d'analyse court (en général de 5 à 20ms).

Théoriquement, les paramètres HNM peuvent être estimés par une technique d'analyse par synthèse c'est-à-dire, par l'optimisation d'une fonction de coût entre le signal original et le signal synthétique. Cependant, cette approche revient à résoudre analytiquement un problème d'optimisation non linéaire de grande dimension. Pour simplifier le problème d'estimation des paramètres HNM, les paramètres de la partie harmonique et de la partie bruitée sont estimés séparément. L'estimation de la fréquence fondamentale et de la fréquence maximale de voisement est isolée de l'estimation des amplitudes et des phases des harmoniques. Ainsi, la première étape d'analyse consiste à estimer la fréquence fondamentale et la fréquence maximale de voisement pour les trames voisées.

3.3.1 Estimation du pitch et décision de voisement

La qualité de la parole synthétisée avec le modèle HNM dépend fortement de la précision de l'estimation de pitch, et de la fréquence maximale de voisement. Cependant, l'estimation du pitch est un problème délicat. Dans la littérature, plusieurs algorithmes ont été proposés pour répondre à ce problème [Hes83, RCRM76, SR79, MQ85, GH87, Her88, DC89, MYC91, Oud98]. Dans cette thèse, nous utilisons un algorithme d'estimation du pitch et de la fréquence maximale de voisement adapté par Stylianou au modèle HNM dans [Sty96a]. L'analyse commence par déterminer un pitch initial $\hat{f}_0$ par une méthode temporelle basée sur une maximisation de la fonction d'autocorrelation. Ce pitch initial est utilisé, par la suite, pour la décision de voisement, pour l'estimation de la fréquence maximale de voisement et finalement pour le raffinement de l'estimation du pitch.

3.3.2 Estimation des amplitudes et des phases

Les amplitudes et les phases des harmoniques sont calculées par minimisation d'un critère des moindres carrés pondérés [Sty96b] :

$$E = \sum_{n=-T_0^i}^{n=T_0^i} W^2(n) \left(s(n) - h(n)\right)^2, \tag{3.5}$$

où $s(n)$ est le signal original, $h(n)$ la partie harmonique définie par (3.2), $W(n)$ la fenêtre d'analyse, et T_0^i la période de pitch de la trame courante. Notons que la trame a une durée égale à deux fois la période de pitch [Sty96b]. La condition de stabilité permet de supposer que les amplitudes, les phases, le nombre d'harmoniques, ainsi que la fréquence maximale de voisement sont constantes durant la trame d'analyse.

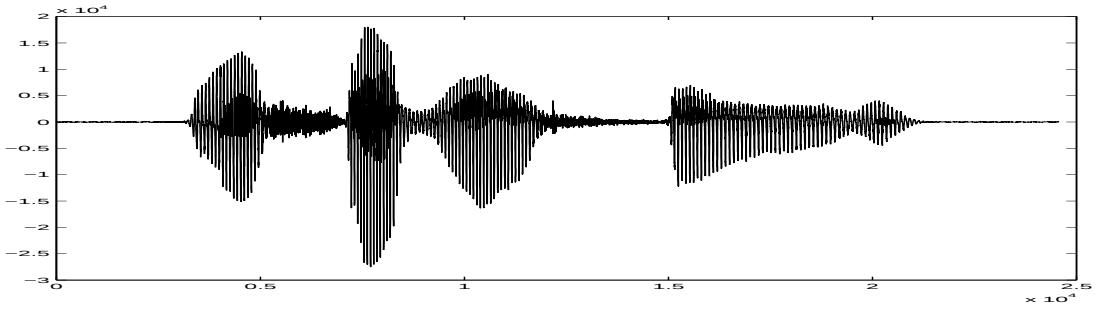
L'équation (3.2) peut s'écrire

$$\hat{h}(n) = c_0 + \sum_{k=1}^{L} c_k \cos(2\pi n k f_0) - s_k \sin(2\pi n k f_0), \tag{3.6}$$

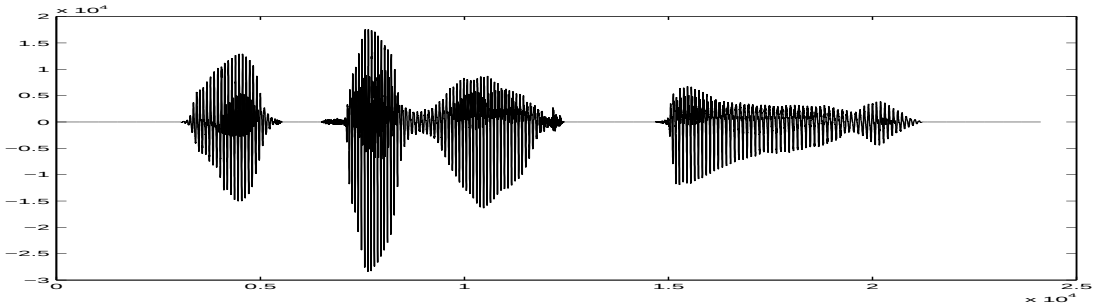
avec

$$c_k = A_k \cos(\varphi_k) \tag{3.7}$$

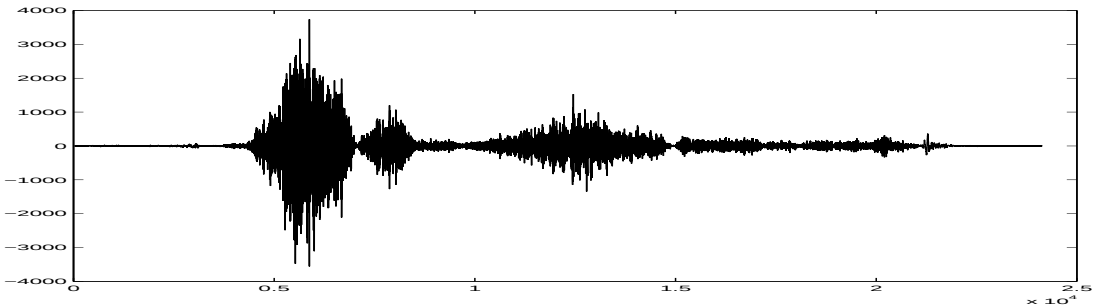
$$s_k = A_k \sin(\varphi_k). \tag{3.8}$$



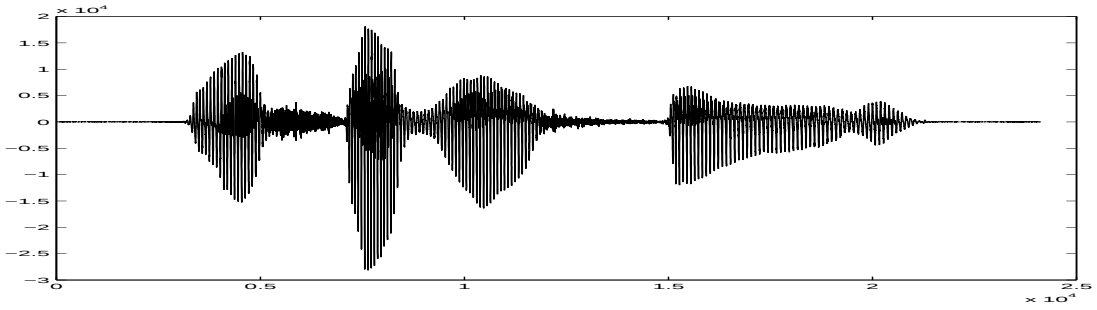
(a) signal original



(b) partie harmonique



(c) partie bruité



(d) signal synthétisé

Figure 3.3 – Analyse et synthèse par HNM : (a) signal original, (b) synthèse de la partie harmonique seule, (c) synthèse de la partie bruitée seule et (d) signal résultant de la somme de (b) et (c).

Les coefficients c_k et s_k sont estimés par minimisation d'un critère des moindres carrées [Sty96b] et les amplitudes et les phases sont données par :

$$A_k = \sqrt{c_k^2 + s_k^2}, \tag{3.9}$$

$$\phi_k = -\arctan \frac{s_k}{c_k}. \tag{3.10}$$

Cette analyse harmonique est importante dans la mesure où elle apporte une information fiable sur la valeur du spectre aux fréquences harmoniques. Une telle information est nécessaire pour avoir une estimation robuste de l'enveloppe spectrale.

3.3.3 Estimation des paramètres du bruit

Pour toutes les trames d'analyse, qu'elles soient voisées ou non, la densité spectrale de puissance est modélisée par un filtre tout-pôle. En utilisant la méthode de l'autocorrélation standard [Kay88], la fonction d'autocorrélation est estimée en utilisant 40 ms du signal centré autour de chaque instant d'analyse. Le gain du filtre est donné par la variance du signal sur la même durée.

Les parties du spectre correspondant à du bruit (qu'il s'agisse de la composante de bruit d'une trame voisée ou d'une trame non voisée) sont modélisées à l'aide d'une simple prédiction linéaire. La réponse fréquentielle du modèle AR ainsi estimé est ensuite échantillonnée à pas constant, ce qui fournit une estimation de l'enveloppe spectrale sur les zones bruitées.

Un des points forts du modèle HNM est que le signal synthétisé $\hat{s}(t)$ et le signal original $s(t)$ sont presque indiscernables perceptuellement. En outre, il permet d'effectuer des traitements du signal de parole de haute qualité, en particulier des modifications du pitch et de la durée des signaux de parole montrant ainsi son utilité dans le cadre de la synthèse de la parole [Sty96b] [Mac96]. Il permet également des modifications directes des phases et des amplitudes du signal en synthétisant la parole avec les paramètres modifiés.

Cependant, pour des modifications spectrales, et donc pour la conversion de voix, l'utilisation directe des paramètres d'amplitudes est peu souhaitable car le nombre de

paramètres est important et variable. Il est, par conséquent, nécessaire d'avoir une représentation paramétrique du spectre d'amplitudes, i.e. l'enveloppe spectrale.

3.3.4 Enveloppes de phase et d'amplitude

L'estimation des enveloppes de phase et d'amplitude peuvent être considérées comme une étape intermédiaire entre l'analyse et la synthèse. Dans la synthèse de la parole, ces enveloppes sont utilisées, par exemple, dans la cas de modifications spectrales et/ou de la fréquence fondamentale. Les amplitudes et les phases calculées lors de l'analyse ne correspondent pas aux amplitudes et aux phases des nouvelles composantes harmoniques. Elles prennent comme valeurs celles que prend l'enveloppe spectrale et l'enveloppe de phase aux nouvelles fréquences.

L'enveloppe de phase est obtenue par une technique de déroulement fréquentiel décrite par Stylianou dans [Sty96b]. Cette technique permet de preserver la continuité de la phase aussi bien dans le domaine fréquentiel que dans le domaine temporel.

Comme nous l'avons vu dans la section 2.2.1, l'enveloppe spectrale correspond au spectre d'amplitude du filtre modélisant le conduit vocal et la partie lisse du spectre de la source glottique. L'enveloppe spectrale est donc une courbe qui passe par les pics des harmoniques sur la partie qui s'étend jusqu'à la fréquence maximale de voisement et qui suit le spectre de la partie bruitée pour les fréquences supérieures à la fréquence maximale de voisement.

Le choix de la représentation de l'enveloppe spectrale dépend, de l'application visée. Par exemple, les LSF sont utilisés en codage de la parole pour leurs bonnes propriétés d'interpolation et de codage, les coefficients cepstraux sont utilisés en reconnaissance de la parole pour leur robustesse au bruit. Ces deux modélisations sont également les plus utilisées en conversion de voix. Dans cette section nous décrivons les techniques d'estimation de ces deux types de modélisation.

3.3.4.1 Estimation de l'enveloppe d'amplitude

Modélisation par les paramètres LSFs

Comme nous l'avons évoqué dans la section (2.2.1), les paramètres LSF sont une variante des coefficients LPC reconnue comme ayant de bonnes propriétés d'interpolation. Pour les calculer, il faut d'abord calculer les coefficients LPC. Pour cela, nous procédons comme dans [MQ95]. Le logarithme de la densité spectrale de puissance, représenté par le log des amplitudes A_l est sur-échantillonné en utilisant une interpolation cubique [UAE93]. Puis, les coefficients du filtre LPC sont estimés par application de l'algorithme de Levinson-Durbin sur les coefficients d'autocorrélation obtenus par une FFT inverse du carré des amplitudes.

Les coefficients a_k du filtre LPC $A_p(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$ sont donc convertis en coefficients LSFs. Dans cette représentation

$$A_p(z) = \frac{1}{2}(P_{p+1}(z) + Q_{p+1}(z)), \tag{3.11}$$

avec

$$P_{p+1}(z) = A(z) + z^{-(p+1)}A(z^{-1}) \tag{3.12}$$

$$Q_{p+1}(z) = A(z) - z^{-(p+1)}A(z^{-1}). \tag{3.13}$$

Les coefficients LSF sont extraits à partir des racines complexes des polynômes $P_{p+1}(z)$ et $Q_{p+1}(z)$. Ces fonctions de transfert possèdent deux propriétés très intéressantes [Sao90] : pour un filtre stable $1/A_p(z)$, toutes les racines de $P_{p+1}(z)$ et $Q_{p+1}(z)$ sont sur le cercle unité et s'alternent deux à deux. D'autre part, ces racines sont conjuguées. En ignorant les racines réelles (1 et -1 selon que p est pair ou impair), le filtre $A_p(z)$ peut être représenté par la séquence w_1, $w_2, \cdots w_p$ des arguments des racines complexe des filtres $P_{p+1}(z)$ et $Q_{p+1}(z)$ se trouvant sur le demi cercle entre 0 et π.

Les paramètres w_i présentent plusieurs propriétés intéressantes. Tout d'abord, ils sont ordonnés : $0 < w_1 < w_2 < \cdots < w_p < \pi$. Cette relation d'ordre est une condition nécessaire et suffisante pour la stabilité du filtre de synthèse $1/A_p(z)$. En outre, les coefficient LSF sont robustes car une erreur sur un seul coefficient LSF aura des repercussions sur une région de spectre située au voisinage de la fréquence correspondant

à ce coefficient. Les coefficients LSF sont des paramètres fréquentiels. La proximité de deux coefficients fait apparaître un pic dans le spectre d'amplitude assimilable à un formant. A partir des coefficients LSF il est donc possible d'identifier grossièrement les zones auditivement importantes dans le spectre du signal de façon très aisée [SI86].

Modélisation par cepstre discret

Cette méthode a été introduite par Galas et Rodet [GR90]. Son objectif est de déterminer les coefficients cepstraux conduisant à une enveloppe spectrale passant le plus proche possible des amplitudes des harmoniques. Étant données les amplitudes spectrales A_l, les coefficients du cepstre discret $c = [c_0 \cdots c_p]$, où p est l'ordre du cepstre, sont obtenus en minimisant un critère des moindres carrées :

$$\varepsilon_r = \sum_{l=1}^{L} |\log A_l - \log |S(f_l, c)||^2, \tag{3.14}$$

où L est le nombre d'harmoniques de la trame. L'amplitude du spectre $|S(f, c)|$ est reliée aux coefficients du cepstre par :

$$\log |S(f, c)| = c_0 + 2 \sum_{i=1}^{p} c_i cos(2\pi f_i). \tag{3.15}$$

La solution de cette équation est donnée par :

$$c = \left[M^T M\right]^{-1} M^T a. \tag{3.16}$$

avec $a = [\log(A_1) \cdots \log(A_L)]$ et

$$M = \begin{pmatrix} 1 & 2cos(2\pi f_1) & 2cos(2\pi f_1 2) & \dots & 2cos(2\pi f_1 p) \\ \vdots & \vdots & & \vdots & \\ 1 & 2cos(2\pi f_L) & 2cos(2\pi f_L 2) & \dots & 2cos(2\pi f_L p) \end{pmatrix}. \tag{3.17}$$

La matrice $M^T M$ dans l'équation (3.16) n'est pas nécessairement inversible. En général, elle est mal-conditionnée lorsque p est proche de L (singulière quand $p > L$).

Pour résoudre ce problème et rendre cette estimation plus robuste, Cappé [CLM95] a introduit un terme de régularisation. L'équation 3.14 s'écrit :

$$\varepsilon_r = \sum_{l=1}^{L} \| \log A_l - \log |S(f_l, c)| \|^2 + \lambda \mathcal{A}[S(f, c)], \tag{3.18}$$

où λ est le paramètre de régularisation et la fonction $\mathcal{A}[S(f,c)]$ est donnée par :

$$\mathcal{A}\left[S(f,c))\right] = c^T R c, \tag{3.19}$$

où R est une matrice (p×p) diagonale donnée par :

$$R = \begin{pmatrix} 0 & & & & & \\ & 8\pi^2 1^2 & & & & \\ & & 8\pi^2 2^2 & & & \\ & & & \ddots & & \\ & & & & 8\pi^2 p^2 \end{pmatrix}. \tag{3.20}$$

La solution de l'équation 3.18 est donnée par :

$$c = \left[M^T M + \lambda R\right]^{-1} M^T a. \tag{3.21}$$

3.3.5 Transformation en échelle de Bark

Des améliorations ont également été proposées afin de mieux prendre en compte les aspects perceptuels. Il est connu que l'oreille humaine est moins sensible aux détails spectraux dans les hautes fréquences que dans les basses fréquences. En exprimant l'enveloppe spectrale à partir d'une échelle de Bark, on parvient à mieux répartir fréquentiellement les erreurs d'estimation. Le but est notamment d'éviter des erreurs grossières aux basses fréquences, car l'oreille humaine y est particulièrement sensible, alors qu'aux fréquences plus élevées ces erreurs sont moins perceptibles. La transformation en échelle de Bark est non linéaire de manière à donner une plus grande importance aux basses fréquences (les basses fréquences sont dilatées tandis que les hautes fréquences sont compressées). La formule de conversion de Hz en Bark est donnée dans [ZT80] par :

$$\begin{cases} f(Bark) = & 13 \arctan\left(0.76 \frac{f(Hz)}{1000}\right) & \text{si } f(Hz) \leq 605 \\ f(Bark) = & 8.7 + 14.2 \log 10\left(\frac{f(Hz)}{1000}\right) & \text{si } f(Hz) > 605. \end{cases} \tag{3.22}$$

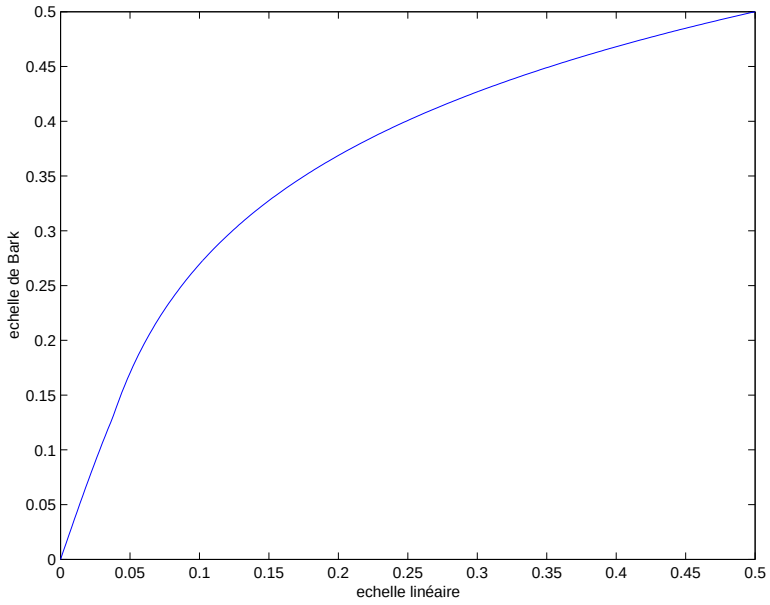


Figure 3.4 – Conversion entre les fréquences en Hertz normalisées (axes des x) et les fréquences en Bark (axes des y) pour une fréquence d'échantillonnage de 16 kHz.

On utilise l'échelle de Bark normalisée dont les fréquences varient entre 0 et $\frac{1}{2}$. L'échelle de bark normalisée est donnée par :

$$f(\text{Bark normalizé}) = \frac{f(Bark)}{2 \times 21.52}, \tag{3.23}$$

où 21.52 correspond à 8000 Hz (la moitié de la fréquence d'échantillonnage) exprimé en échelle de Bark. La figure 3.4 montre la correspondance entre l'échelle linéaire et l'échelle de Bark normalisée.

D'autres transformations de l'échelle des fréquences peuvent être utilisées comme la transformation en échelle de Mel, ou celle donnée par McAulay and Quatieri [MQ91], qui est une fonction linéaire dans les régions basse fréquences et logarithmique dans les régions hautes fréquences. Les figures 3.5 et 3.6 présentent une comparaison entre deux enveloppes spectrale en échelle linéaire puis en échelle de Bark obtenues, repectivement, par une modélisation LSF et une modélisation cepstrale. Ces figures montrent que l'utilisation de l'échelle de Bark permet de réduire sensiblement l'erreur entre l'enveloppe spectrale et les amplitudes des premières harmoniques.

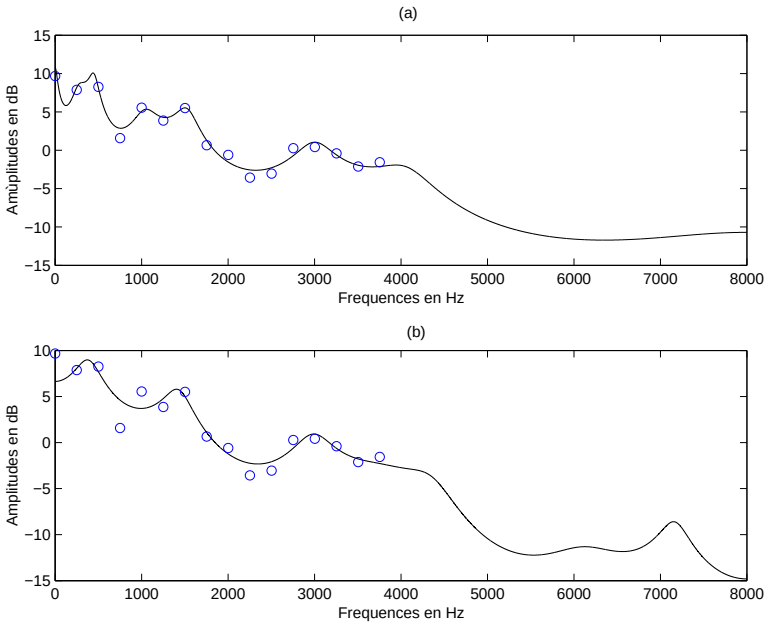


Figure 3.5 – Modélisation AR : enveloppes spectrales en échelle de Bark (a), puis en échelle linéaire (b). Les ronds représentent les amplitudes des harmoniques.

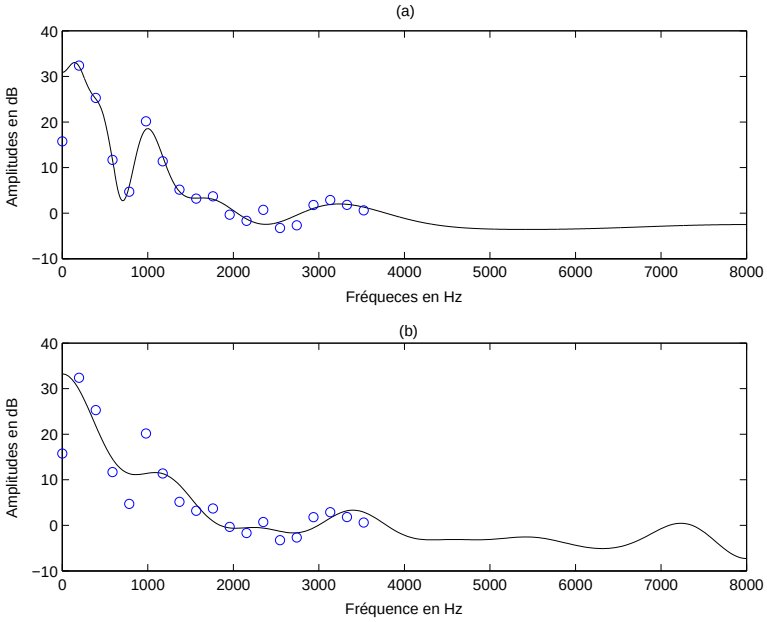


Figure 3.6 – Modélisation cepstrale : enveloppes spectrales en échelle de Bark (a) puis en échelle linéaire (b). Les ronds représentent les amplitudes des harmoniques.

Dans le reste de ce chapitre, nous proposons de comparer les performances de ces deux représentations dans le cadre de la conversion par GMM.

3.4 Conversion de voix par GMM

Dans cette section, nous supposons avoir deux ensembles de vecteurs spectraux $X_{p\times N} = [x_1, x_2, \cdots x_N]$ et $Y_{p\times N} = [y_1, y_2, \cdots y_N]$ source et cible temporellement alignés de manière à décrire le même contenu acoustique, N étant le nombre de vecteurs et p leur dimension. Comme dans [Kai01], nous avons utilisé le modèle GMM pour modéliser la densité de probabilité conjointe des vecteurs acoustiques source et cible.

3.4.1 Modèle de mélange de Gaussiennes (GMM)

Formellement, la densité de probabilité d'une variable aléatoire z suivant un modèle GMM d'ordre Q s'écrit :

$$p(z) = \sum_{i=1}^{Q} \alpha_i \mathcal{N}(z; \mu_i, \Sigma_i), \tag{3.24}$$

avec

$$\sum_{i=1}^{Q} \alpha_i = 1, \text{ et } \alpha_i \geq 0,$$

où les α_i sont les coefficients de mélange (α_i est la probabilité *a priori* que z soit généré par la $i^{\text{ème}}$ composante gaussienne) et $\mathcal{N}(z; \mu, \Sigma)$ est la densité de probabilité de la loi normale de moyenne μ et de variance Σ donnée par :

$$N(z; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right). \tag{3.25}$$

Les paramètres (α, μ, Σ) sont estimés par un algorithme EM (Expectation Maximisation) [DLR77], une méthode itérative pour l'estimation des paramètres au sens du maximum de vraisemblance. Dans le cas d'un mélange de densités gaussiennes, chaque itération implique deux étapes successives :

Expectation :

Calcul des probabilités conditionnelles d'appartenance à chaque classe q pour chaque observation z_t :

$$P(q|z_t) = \frac{\alpha_q |\Sigma_q|^{1/2}| \exp\left[-\frac{1}{2}(z_t - \mu_q)^T \Sigma_q^{-1}(z_t - \mu_q)\right]}{\sum_{i=1}^{Q} \alpha_i |\Sigma_i|^{1/2}| \exp\left[-\frac{1}{2}(z_t - \mu_i)^T \Sigma_i^{-1}(z_t - \mu_i)\right]}. \tag{3.26}$$

Maximisation :

Les probabilités *a priori* $\hat{\alpha_q}$ sont ré-estimées à chaque itération par la moyenne des probabilités *a posteriori* $P(q|z_t)$:

$$\hat{\alpha}_q = \frac{1}{N} \sum_{t=1}^{N} P(q|z_t). \tag{3.27}$$

Les moyennes et les matrices de covariance sont réestimées respectivement par les moyennes et les matrices de covariance des observations pondérées par les probabilités *a posteriori*

$$\hat{\mu_q} = \frac{\sum_{t=1}^{N} P(q|z_t) z_t}{\sum_{t=1}^{N} P(q|z_t)}, \tag{3.28}$$

$$\hat{\Sigma_q} = \frac{\sum_{t=1}^{N} P(q|z_t)(z_t - \mu_q)(z_t - \mu_q)_T}{\sum_{t=1}^{N} P(q|z_t)}. \tag{3.29}$$

Dans ces équations, N désigne le nombre d'observations dans la base d'apprentissage. L'algorithme EM est un algorithme déterministe dont seule la convergence vers un optimum local est assurée. En pratique, pour que l'algorithme converge vers une solution satisfaisante, l'initialisation de l'algorithme EM revêt une importance particulière. Dans notre application, l'algorithme EM est initialisé à l'aide d'une technique classique de quantification vectorielle. Les α_i sont initialisés proportionnellement au nombre de vecteurs appartenant à chaque classe, les μ_i et les Σ_i sont initialisés par les moyennes et les variances empiriques des vecteurs appartenant à chaque classe.

3.4.2 Régression par GMM

La régression est un outil qui permet d'étudier et de mesurer la relation existant entre un caractère expliqué et un ou plusieurs caractères explicatifs. En se basant sur les données d'un échantillon, l'analyse de régression cherche à déterminer une estimation d'une relation mathématique entre les deux caractères différents. Le but est d'estimer les valeurs d'une des variables à l'aide des valeurs des autres. L'analyse de régression a des applications multiples dans presque tous les domaines de la science. En effet, lorsqu'on arrive à déterminer la relation entre deux ou plusieurs variables, on peut alors, à l'aide du modèle de régression ainsi construit, prévoir les valeurs futures de ces variables, étant entendu que les conditions demeurent identiques et qu'il existe toujours une marge d'erreur. La modélisation AR présentée plus haut est un exemple de régression linéaire.

Le but est alors d'estimer une fonction F qui permet de faire le lien entre vecteurs source et cible. Nous employons l'approche de la densité jointe [Kam96, GJ94] comme appliquée par Kain [KM98b] au problème de conversion de voix. Cette approche implique de modéliser à l'aide d'un mélange de gaussiennes la densité jointe $P_Z(z) = p(x, y)$ où x et y désignent respectivement les vecteurs spectraux source et cible.

La fonction de transformation est déterminée en minimisant l'erreur quadratique moyenne (3.30)

$$E(F) = \sum_{i=1}^{N} \|y_i - F(x_i)\|^2. \tag{3.30}$$

La solution est donnée par l'espérance conditionnelle de Y sachant X [GJ94] :

$$\begin{aligned} \hat{y} = F(x) &= \mathbb{E}(Y|X = x) \\ &= \sum_{i=1}^{Q} h_i(x) \left[\mu_i^y + \Sigma_i^{xy}(\Sigma_i^{xx})^{-1}(x - \mu_i^x)\right], \end{aligned} \tag{3.31}$$

où

$$h_i(x) = p(q|x) = \frac{\alpha_i \mathcal{N}(x; \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{Q} \alpha_j \mathcal{N}(x; \mu_j^x, \Sigma_j^{xx})} \tag{3.32}$$

est la probabilité *a posteriori* que x soit généré par la $i^{\text{ème}}$ composante gaussienne [Kam96], avec

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} \tag{3.33}$$

et

$$\mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}. \tag{3.34}$$

3.4.3 Implémentation

Après l'analyse HNM des corpus d'apprentissage, nous obtenons deux ensembles de vecteurs spectraux $[x_1, x_2, \cdots x_n]$ et $[y_1, y_2, \cdots y_m]$ correspondant respectivement aux voix source et cible. Il est souhaitable de faire l'analyse de manière asynchrone avec un pas d'analyse fixe. Notons que le nombre de vecteurs de la source n et celui de la cible m ne sont, généralement, pas égaux. Il convient dans un premier temps d'effectuer un appariement entre vecteurs acoustiques source et cible. Ce dernier est obtenu par un algorithme DTW classique [RJ93]. En sortie d'une telle procédure d'alignement, nous obtenons une séquence de vecteurs source et cible appariés. Lors de cette opération, des contraintes sont introduites de manière à respecter les marques de segmentation en phones. Les trames alignées dont l'une est voisée et l'autre non-voisée ne sont pas retenues dans la base d'apprentissage. Après alignement nous obtenons deux suites de vecteurs acoustiques

$$X = [x_1, x_2, \cdots x_N] \quad \text{et} \quad Y = [y_1, y_2, \cdots y_N] \tag{3.35}$$

contenant le même nombre N de vecteurs acoustiques, caractérisant la même séquence de parole, prononcée respectivement par les locuteurs source et cible. N dépend de la taille de la base d'apprentissage.

Après l'alignement des bases de données des locuteurs, la seconde étape d'apprentissage consiste en l'estimation des paramètres GMM de la densité jointe de la source et de la cible. Comme mentionné précédemment dans la section (3.4.2), les paramètres GMM sont estimés par un algorithme EM initialisé à l'aide d'une quantification vectorielle classique. Comme chaque itération augmente la vraisemblance, il est utile de determiner à quelle itération il faut s'arrêter. La figure 3.7 montre l'évolution de la valeur de la

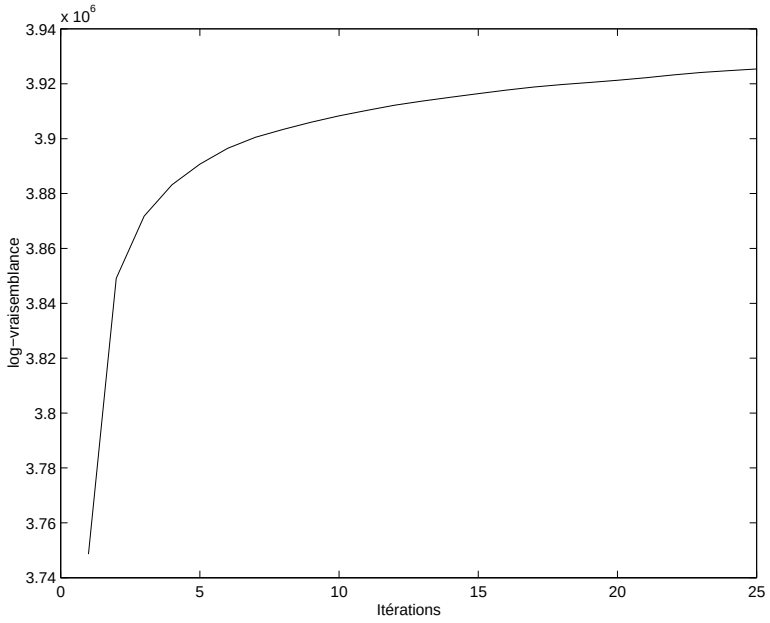


Figure 3.7 – Evolution de la log-vraisemblance.

log-vraisemblance après chaque itération. Il faut noter qu'après une augmentation rapide de la log-vraisemblance, cette augmentation stagne à partir de quelques itérations. La stratégie utilisée est de faire tourner l'algorithme EM jusqu'à ce que la variation relative de la log-vraisemblance soit en deçà d'un certain seuil ou après 25 itérations.

3.5 Transformation et synthèse

Une fois apprise, la fonction de transformation peut être appliquée à la voix source afin de réaliser la conversion souhaitée. Pour cela, les paramètres HNM sont extraits du signal de parole de manière pitch-synchrone. Pour les trames non voisées, un pas d'analyse de 10 ms est utilisé. Puis pour chaque vecteur, les probabilités *a posteriori* sont calculées suivant (3.32), et finalement la fonction de transformation (5.1) est appliquée. Il faut noter que dans le cas où l'enveloppe spectrale est modélisée par le cepstre discret, le coefficient d'énergie c_0 n'est pas pris en compte lors de la transformation.

La synthèse est une opération consistant à calculer les échantillons du signal de parole à partir des paramètres HNM modifiés. Étant données les enveloppes spec-

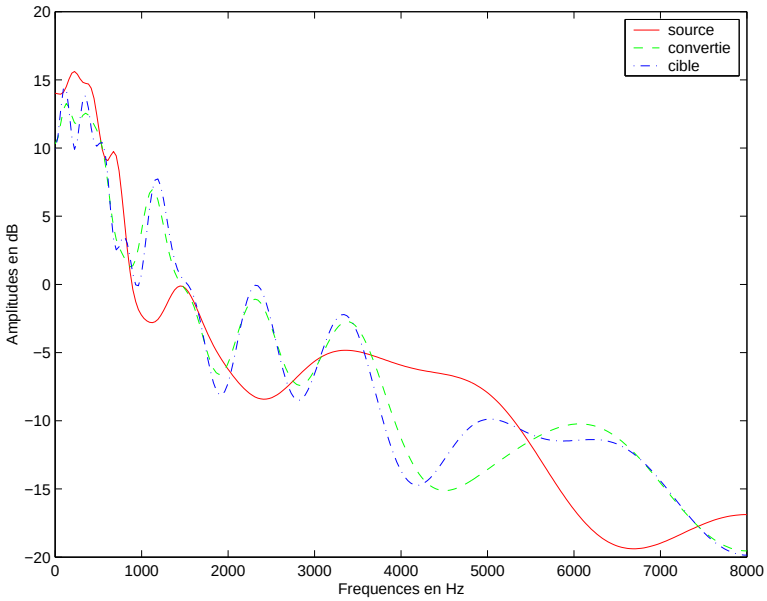


Figure 3.8 – Exemples d'enveloppes spectrales source, cible et convertie.

trales transformées et les nouvelles valeurs de pitch, la première étape est le calcul des phases et des amplitudes aux nouvelles fréquences harmoniques. Ceci est fait en échantillonnant l'enveloppe spectrale transformée et l'enveloppe de phase aux nouvelles fréquences harmoniques. Nous utilisons les enveloppes de phase de la voix source. Une fois les amplitudes et les phases des harmoniques déterminées la synthèse proprement dite est réalisée comme dans [Sty96b].

Synthèse de la partie Harmonique

La synthèse de la partie harmonique se fait échantillon par échantillon. Pour éviter des discontinuités aux frontières des trames, les paramètres de la partie harmonique ne sont pas considérés comme constants sur la durée de la trame de synthèse. En particulier, les amplitudes et les phases sont interpolées linéairement entre les instant de synthèse. Soient $[A_k^i, \phi_k^i, f_0^i]$ et $[A_k^{i+1}, \phi_k^{i+1}, f_0^{i+1}]$ les paramètres de la $k^{\text{ème}}$ harmonique aux instants de synthèse t_s^i et t_s^{i+1}. Les amplitudes instantanées $A_k(t)$ sont données par :

$$A_k(t) = A_k^i + \frac{A_k^{i+1} - A_k^i}{t_s^{i+1} - t_s^i} t, \qquad \text{pour } t_s^i \leq t \leq t_s^{i+1}. \tag{3.36}$$

Pour ce qui concerne la phase, premièrement, la phase à l'instant t_s^{i+1} est prédite à

partir de celle de l'instant t_s^i par :

$$\hat{\phi}_k^{i+1} = \phi_k^i + k2\pi\bar{f}_0^i(t_s^{i+1} - t_s^i), \tag{3.37}$$

avec

$$\bar{f}_0^i = \frac{f_0^i + f_0^{i+1}}{2}. \tag{3.38}$$

Puis la phase ϕ_k^{i+1} est augmentée par un multiple M_k de 2π pour approcher la valeur prédite. M_k est donné par

$$M_k = \langle \frac{1}{2\pi}(\hat{\phi}_k^{i+1} - \phi_k^{i+1}) \rangle, \tag{3.39}$$

où le symbole $\langle\ .\ \rangle$ représente l'entier le plus proche. La phase instantanée est alors donnée par une simple interpolation linéaire :

$$\phi_k(t) = \phi_k^i + \frac{\phi_k^{i+1} + 2\pi M_k - \phi_k^i}{t_s^{i+1} - t_s^i}t, \quad \text{pour } t_s^i \leq t \leq t_s^{i+1}. \tag{3.40}$$

Dans cette analyse, nous avons supposé que la trame i et la trame $i+1$ sont toutes deux voisées. Si la trame i est voisée et la trame $i+1$ est non voisée, les amplitudes de la trame $i+1$ sont mises à zéro, i.e., $A_k^{i+1} = 0$, $\forall k$, tout en gardant la même fréquence fondamentale, i.e., $f_0^{i+1} = f_0^i$. Les phases sont alors données par :

$$\phi_k^{i+1} = \phi_k^i - k2\pi f_0^i(t_s^{i+1} - t_s^i). \tag{3.41}$$

Dans le cas où la trame i est non voisée et la trame $i+1$ est voisée, $A_k^i = 0$, $\forall k$, et $f_0^i = f_0^{i+1}$, alors la phase est donnée par

$$\phi_k^i = \phi_k^{i+1} + k2\pi f_0^i(t_s^{i+1} - t_s^i). \tag{3.42}$$

Dans la procédure d'interpolation décrite ci-dessus, nous avons supposé que les deux trames ont le même nombre d'harmoniques. Or, comme la fréquence maximale de voisement et la fréquence fondamentale sont variables dans le temps, le nombre d'harmoniques n'est pas nécessairement le même. Pour contourner ce problème, on complète par des harmoniques d'amplitudes nulles pour avoir le même nombre d'harmoniques. Les phases sont définies comme précédemment par les équations (3.41) ou (3.42) selon

les cas. Ayant déterminé les amplitudes instantanées $A_l(t)$ et les phases instantanées $\phi_l(t)$ la partie harmonique est donnée par

$$\tilde{h}(t) = \sum_{l=1}^{L} A_l(t) \cos(\phi_l(t)). \tag{3.43}$$

Notons que la partie harmonique $\tilde{h}(t)$ modifiée occupe la même bande de fréquences que la partie harmonique originale $h(t)$, ce qui fait que, dans le cas de modification de la fréquence fondamentale, elle ne contiennent pas le même nombre d'harmoniques. Le nouveau nombre d'harmoniques est obtenu par la division de la fréquence maximale de voisement F_c par la fréquence fondamentale F_0 :

$$L = \langle \frac{F_c}{F_0} \rangle. \tag{3.44}$$

De plus, dans le cas de modification de pitch, les amplitudes sont normalisées de façon à conserver l'énergie de la partie harmonique.

Synthèse de la partie stochastique

Le bruit est localisé sur la plage comprise entre la fréquence de coupure F_c et $F_e/2$ (F_e étant la fréquence d'échantillonnage). Il s'agit de filtrer un bruit blanc gaussien par l'enveloppe spectrale issue de l'analyse et ensuite de le filtrer par un filtre passe haut de fréquence de coupure F_c. Le résultat de ce filtrage est ensuite ajouté à la partie déterministe.

Lorsqu'on dispose de la modélisation AR sous forme de coefficients de réflexion (dits coefficients PARCOR), il n'est pas nécessaire de passer dans le domaine spectral, le simple filtrage d'un bruit blanc par le filtre en treillis correspondant convient. Le signal à court-terme terme résultant est alors agencé aux signaux précédents par la méthode d'addition-recouvrement [Sty96b].

Le signal de bruit ainsi généré trame à trame puis agencé par la méthode d'addition-recouvrement ne se fond pas toujours bien avec le signal harmonique, ce qui est très gênant à l'oreille. Pour éviter cet inconvénient, le bruit doit être modulé temporellement par une fonction dont la périodicité varie comme celle du signal harmonique, et ce uniquement pour les trames mixtes. Des détails sur ce phénomène sont présentés dans [Her91]. Une simple fonction triangulaire (Figure 3.9) donne des résultats satisfaisants.

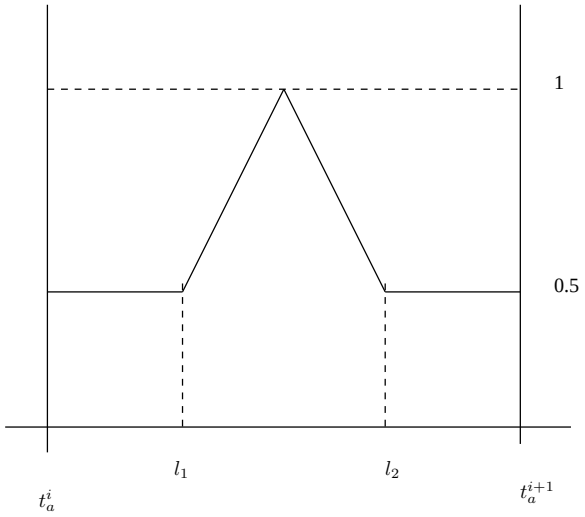


Figure 3.9 – Fonction de modulation temporelle du bruit. t_s^i et t_s^{i+1} sont deux instants de synthèse successifs. $l_1 = 0.15(t_s^{i+1} - t_s^i)$ et $l_2 = 0.85(t_s^{i+1} - t_s^i)$.

3.6 Expérimentations

Dans cette section nous explorons la dépendance de la qualité de la conversion de voix vis à vis du choix de la paramétrisation.

Le choix du nombre de composantes GMM est un problème délicat. Choisir un nombre de composantes petit peut conduire a un modèle incapable de distinguer entre les différentes caractéristiques de la distribution du locuteur. Choisir un nombre très grand peut réduire les performances de la transformation. En effet, il peut alors y a voir beaucoup de paramètres à estimer par rapport à la base d'apprentissage disponible. Notons également, qu'un nombre élevé de composantes conduit à un coût de calcul qui peut s'avérer excessif durant la phase l'apprentissage.

Jusqu'a présent, il n'existe pas de méthode théorique pour le choix du nombre de composantes nécessaires pour bien modéliser l'espace acoustique d'un locuteur. Dans ce travail, nous avons fait varier l'ordre du mélange en considérant des valeurs de Q égales à 8, 16, 32 et 64 pour étudier l'influence du choix de l'ordre du mélange sur la qualité de la conversion. Nous avons, également, fait varier la taille des vecteurs d'apprentissage en prenant p successivement égal à 12, 14, 16, 18 et 20.

La taille des vecteurs acoustiques est également un paramètre important pour l'apprentissage. Pour les différentes combinaisons possibles de ces paramètres, la fonction de transformation est estimée en utilisant les données d'apprentissage pour toutes les combinaisons de locuteurs. Afin d'évaluer objectivement les conversions effectuées, nous nous appuyons, notamment, sur une mesure de distorsion spectrale.

Pour apprendre et tester les fonctions de transformation, nous disposons de quatre corpus de parole échantillonnées à 16kHz et correspondant à deux voix de femme et deux voix d'homme. Chaque corpus contient 15 minutes de parole dont 5 minutes seront utilisées pour l'apprentissage et 10 minutes pour les tests. Comme nous l'avons évoqué précédemment, l'apprentissage d'une fonction de transformation nécessite que les deux corpus d'apprentissage source et cible aient exactement le même contenu acoustique. Par conséquent, sur nos bases, les seules conversion possibles étaient les conversion homme-femme et les conversion femme-homme. Le problème des bases d'apprentissage non parallèles sera traité dans le chapitre 6.

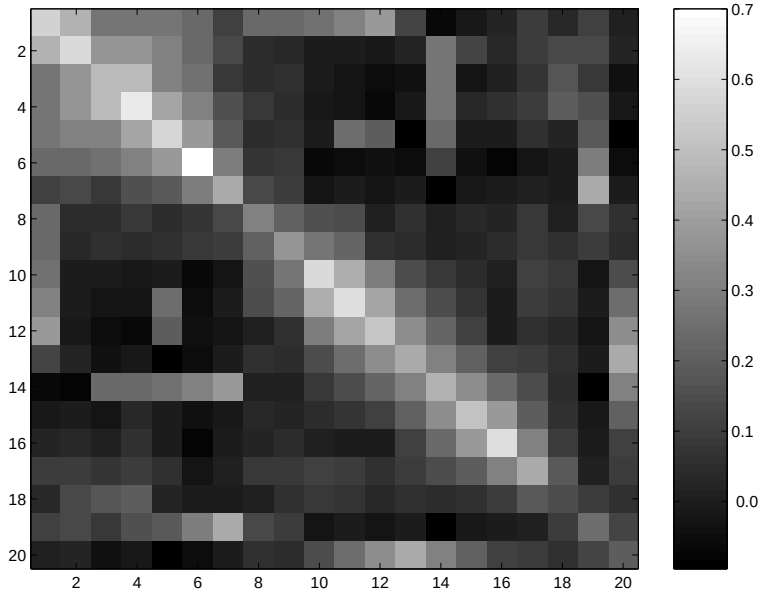


Figure 3.10 – Exemple de matrice de covariance.

Lors de l'apprentissage, nous avons utilisé des matrices de covariances pleines. La figure 3.10 présente un exemple de matrice de covariance de la source en valeur absolue.

La valeur la plus élevée de cette matrice est égale à 0.65, 50% des éléments de cette matrice ont des valeurs supérieures à 0.25 et seulement 5% ont des valeurs inférieures à 0.01. Dans cet exemple, l'énergie de la diagonale ne représente que 17% de l'énergie totale de cette matrice.

3.6.1 Erreurs et indices de performances

Il y a deux sortes d'erreurs particulièrement importantes : l'erreur de transformation $E(C,T)$ entre les vecteurs cible et les vecteurs transformés, et l'erreur inter-locuteurs $E(S,C)$ entre les vecteurs source et cible. Ces erreurs ne peuvent pas être mesurées directement, mais peuvent être approchées empiriquement par une distance moyenne entre les vecteurs spectraux.

Étant donnée une distance d et trois suites de vecteurs spectraux $S = [S_1, S_2, \cdots, S_N]$, $C = [C_1, C_2, \cdots, C_N]$ et $T = [T_1, T_2, \cdots, T_N]$, correspondant respectivement aux vecteurs source, cible et transformés, les erreurs décrites précédemment s'écrivent alors :

$$E(C,T) = \frac{1}{N}\sum_{i=1}^{N} d(C_i, T_i) \tag{3.45}$$

et

$$E(S,C) = \frac{1}{N}\sum_{i=1}^{N} d(S_i, C_i). \tag{3.46}$$

Pour pouvoir mesurer l'apport de la transformation, on utilise une erreur normalisée (EN), qui est le rapport des deux erreurs décrites précédemment, ce rapport (équation 3.47) représentant une mesure normalisée de la proximité entre les paramètres convertis et ceux de la cible :

$$EN = \frac{E(C,T)}{E(S,C)}. \tag{3.47}$$

La distance la plus connue et la plus utilisée est la distance euclidienne :

$$d_{h,2}(A,B) = \sqrt{\sum_{i=1}^{p} |a_i - b_i|^2}. \tag{3.48}$$

Dans l'expression de la distance euclidienne nous supposons implicitement que perceptuellement, les différentes composantes des vecteurs contribuent, de manière identique à la distance globale. En pratique, certaines composantes ont une plus grande

importance. Par exemple, dans le cas des coefficients LSFs, l'oreille humaine est beaucoup plus sensible à une variation des premiers coefficients qu'à la même variation des derniers. Pour le cas des coefficients de réflexion (coefficients PARCOR), l'importance dépend directement de l'ordre de prédiction : le spectre du signal synthétisé est beaucoup plus sensible à une faible variation du premier coefficient qu'à la même variation du dernier. Pour répondre à ce problème, il est nécessaire de faire intervenir une pondération pour bien refléter ce phénomène. L'expression générale d'une distance quadratique pondérée est la suivante :

$$d(A, B) = (A - B)^T W (A - B), \tag{3.49}$$

où W est une matrice carrée définie positive. Lorsque W est l'inverse de la matrice de covariance des vecteurs spectraux, on parle alors de la distance de Mahalanobis.

Cependant, une distance quadratique, pondérée ou non, entre deux vecteurs de cepstre discret, ou deux vecteurs LSF n'a pas la même signification dans le domaine spectral. En d'autres termes, le fait que la distance entre deux vecteurs de coefficients cepstraux soit plus petite que la distance entre deux vecteurs de coefficients LSF ne signifie pas que les deux spectres synthétisés par les deux vecteurs cepstraux seront plus proches perceptuellement que ceux synthétisés par les deux vecteurs LSFs. Pour respecter la sensibilité de l'oreille aux distorsions spectrales, des distances spectrales particulières ont été utilisée [GBGM80, GM76]. Ces distances sont fonction des densités spectrales des signaux originaux et transformés, ce qui fait d'eux une jauge mieux adaptée pour une comparaison de la qualité de conversion du cepstre et des LSFs. Dans ce travail, nous utilisons une distance dans le domaine spectral décrite dans [Pal95] donnée par :

$$d_{DS} = \frac{1}{N} \sum_{i=1}^{N} \|P_{dB}(A) - P_{dB}(B)\|^2, \tag{3.50}$$

où $P_{dB}(A)$ désigne la densité spectrale en échelle de Bark issue de A exprimée en dB échantillonnée régulièrement sur 512 points.

3.6.2 Résultats

Les figures 3.11, 3.12, 3.13 et 3.14 présentent les résultats obtenus avec le cepstre discret et les paramètres LSF en faisant varier l'ordre du mélange et la taille des vecteurs spectraux. Les histogrammes ont des allures similaires, et font apparaître une diminution de la distorsion spectrale moyenne lorsque le nombre de classe augmente quelle que soit la taille des vecteurs spectraux. En effet, l'augmentation de l'ordre du mélange permet une bonne modélisation de l'espace acoustique. Cependant, la taille de la base d'apprentissage étant limitée, il faut donc trouver un compromis entre le volume de données et l'ordre du mélange. Sur les expériences que nous avons effectuées, un GMM d'ordre $Q = 128$ entraîné sur 5 minutes de parole conduit, dans certains cas, à la divergence de l'algorithme d'estimation.

En outre, sur les mêmes figures, nous pouvons comparer l'effet de la taille p des vecteurs d'apprentissage sur les performances de la transformation. Nous constatons que, pour un ordre de mélange fixe, la diminution de la distorsion spectrale est marginale pour $p > 14$, quelle que soit la représentation spectrale utilisée. Cela implique que le choix de la taille p des vecteurs utilisés dépend de leur bonne modélisation de l'enveloppe spectrale.

Les figures 3.15 et 3.16 présentent une comparaison directe des performances présentées par les deux paramétrisations pour un ordre $p = 20$. Les courbes ont des allures similaires. Pour un mélange de 64 composantes, la distorsion spectrale est identique pour une conversion homme-femme, et légèrement différente pour une conversion femme-homme.

Dans une étude antérieure [EnRC03b], basée sur la même distorsion spectrale, nous avions constaté que les paramètres LSF présentent des performances légèrement supérieures. Cette situation est un rappel des limites de l'utilité des mesures objectives pour juger de la qualité de la parole. Cependant, ces mesures sont utiles pour comparer différentes méthodes en utilisant le même système de paramétrisation.

Nous ne pouvons pas attester de la supériorité d'une paramétrisation sur autre à partir de ces mesures. Il est pertinent de mesurer sur le plan de la perception la différence entre les deux modélisation, par le biais de tests d'écoute. Pour qu'une telle

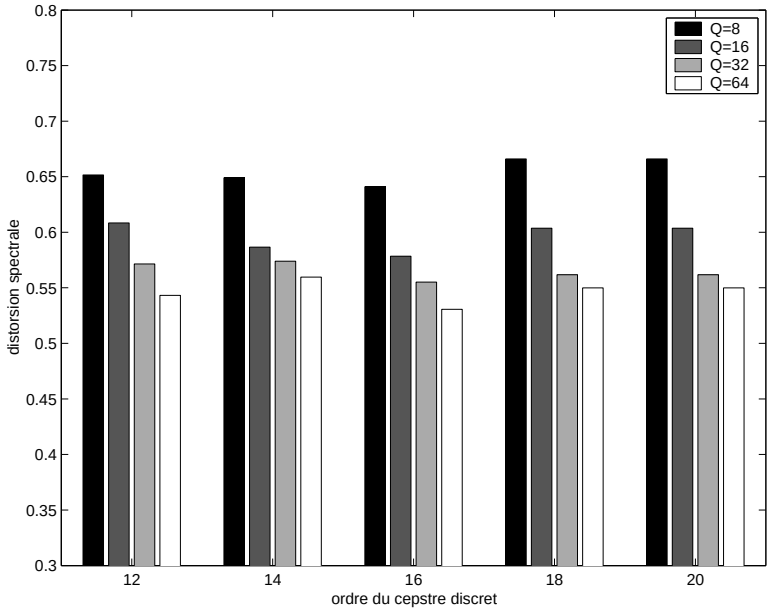


Figure 3.11 – Erreurs de conversion en utilisant des paramètres cepstraux (conversion femme-homme).

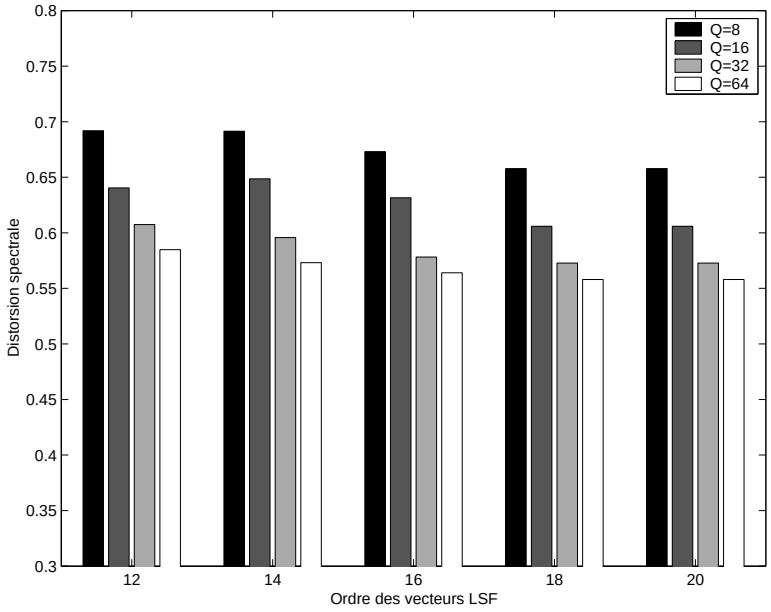


Figure 3.12 – Erreurs de conversion en utilisant des paramètres LSF (conversion femme-homme).

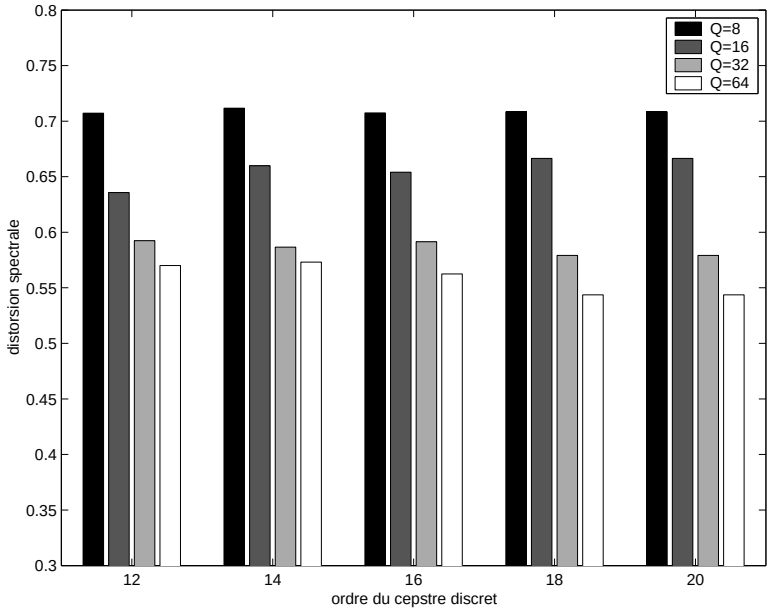


Figure 3.13 – Distorsion spectrale moyenne entre enveloppes cible et convertie en utilisant des paramètres cepstraux (conversion homme-femme).

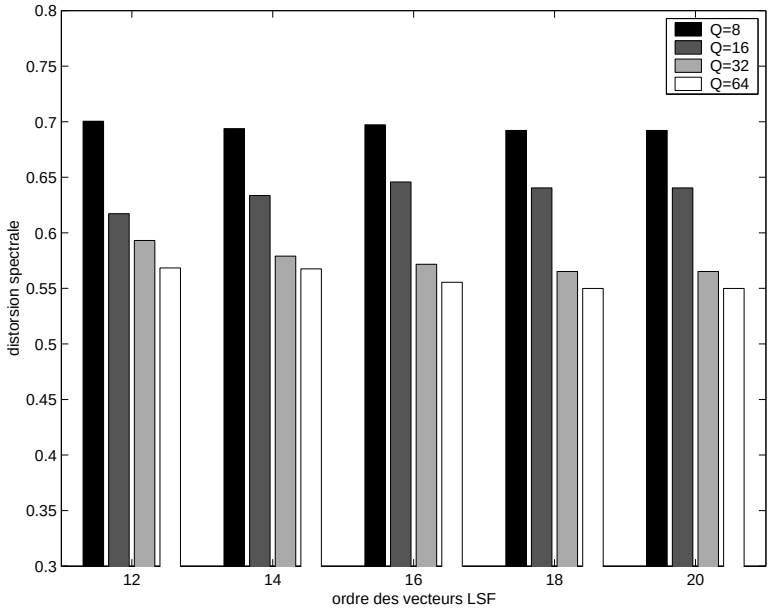


Figure 3.14 – Distorsion spectrale moyenne entre enveloppes cible et convertie en utilisant des paramètres LSF (conversion homme-femme).

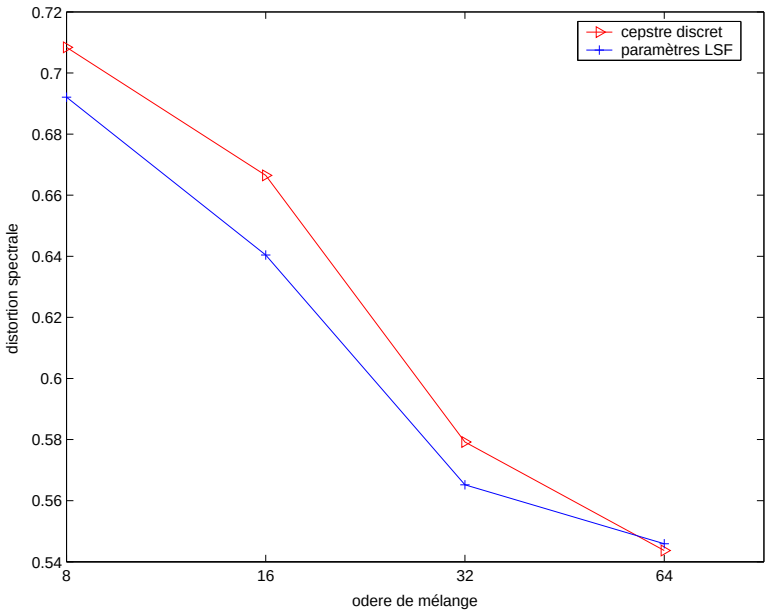


Figure 3.15 – Distorsion spectrale moyenne entre enveloppes cible et convertie pour une transformation homme-femme et $p = 20$.

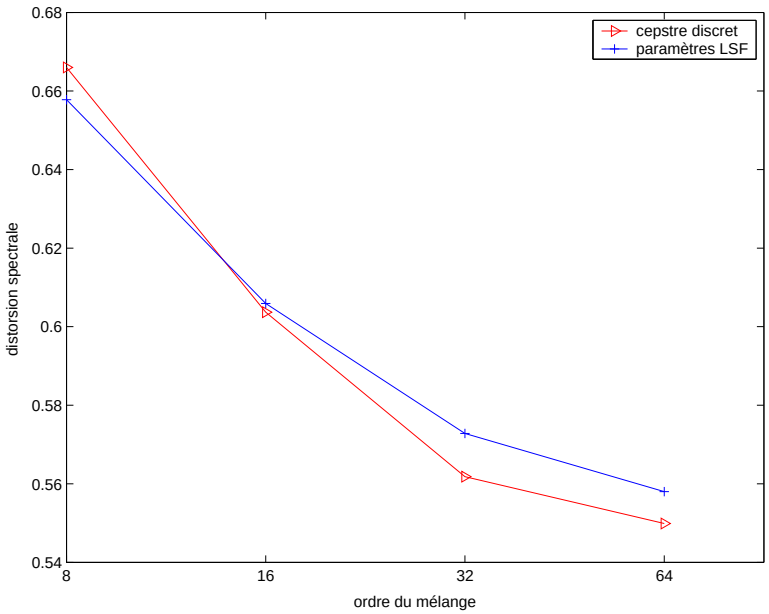


Figure 3.16 – Distorsion spectrale moyenne entre enveloppes cible et convertie pour une transformation femme-homme et $p = 20$.

évaluation puisse être faite, il est cependant nécessaire de pouvoir modifier d'autres paramètres tels que le pitch et le rythme d'élocution.

Les deux paramétrisations conduisant à des mesures objectives similaires dans une application de conversion de voix, nous avons testé de comparer leurs performances en terme de codage. Pour cela un test subjectif de type MOS [1] a été réalisé. Nous avons regroupé une dizaine d'auditeurs et leur avons présenté plusieurs stimuli synthétisés en utilisant le cepstre discret et les paramètres LSF de façon aléatoire. Les auditeurs ont jugé la qualité globale en donnant à chaque fois une note de 1 à 5. Pour tester la cohérence des notes données par les auditeurs, nous avons répété certains stimuli au cours du test. Les auditeurs qui donnent à la même phrase de test deux notes différentes de deux point sont éliminés.

Les MOS moyens, ne dégagent pas de préférence nette entre les deux modélisations. En effet, la modélisation par cepstre discret et par paramètres LSF ont obtenu un MOS égal à 4.3 et 4.2 respectivement. En considérant les paires de notes attribuées par sujets et par phrases, 93% des phrases ont été jugé équivalentes, et 7% des phrases donnent la préférence à la modélisation par cepstre discret. Sur ces dernières phrases, la différence entre les notes attribuées à chaque modélisation est supérieure à 2. En effet, la modélisation par les paramètres LSF conduit dans ces cas à des artéfacts qui sont perçues comme gênants.

3.7 Conclusion

Dans ce chapitre nous avons étudié l'influence de la modélisation spectrale sur les performances de la conversion par GMM, en s'appuyant sur des tests objectifs et subjectifs. Nous avons comparé les performances accessibles par la modélisation par cepstre discret et par les paramètres LSF dans le cadre d'une procédure d'analyse et de synthèse par HNM.

Les tests effectués ont montré que les deux paramétrisations conduisent à des résultats similaires. Cependant, l'utilisation des LSF en conversion de voix nécessite

[1] Mean Opinion Score

un contrôle des vecteurs convertis. En effet, lors de la modification de ces paramètres, des problèmes numériques tels que le non respect de l'ordonnancement des vecteurs LSF peuvent apparaître. Il est donc nécessaire d'introduire, dans un système de conversion de voix par LSF, un mécanisme de correction visant à garantir l'ordonnancement des coefficients et donc la stabilité du filtre AR résultant. Un tel mécanisme peut être difficile à contrôler, la conversion des LSF pouvant elle même engendrer des artéfacts. De plus, le calcul des LSF demande beaucoup plus de ressource lors de l'analyse. Pour ces raisons, nous avons décidé d'utiliser la modélisation par cepstre discret dans la suite de ce travail.

Chapitre 4

Transformation de la fréquence fondamentale

4.1 Introduction

Dans le chapitre précédent, nous avons traité le problème de transformation du timbre. Dans le présent chapitre nous nous intéressons à la transformation de la fréquence fondamentale. Contrairement à la conversion du timbre, ce problème a suscité peu de travaux.

Traditionnellement, la stratégie de conversion de pitch se résume à une simple mise à l'échelle du pitch entre locuteurs source et cible, de manière à respecter globalement la moyenne et la variance du pitch du locuteur cible. Ainsi, le pitch converti s'écrit sous la forme

$$\hat{F_c} = ((F_s - \mu_s)/\sigma_s)\sigma_c + \mu_c, \tag{4.1}$$

où F_s désigne la fréquence fondamentale source, μ_s et σ_s sont la moyenne et l'écart type de la fréquence fondamentale du locuteur source, et μ_c et σ_c ceux du locuteur cible.

Des études plus récentes ont essayé d'introduire des informations liées à la dynamique du pitch au niveau de la phrase. Ainsi, dans [CVW02] Cyessens a enrichi cette

transformation en prenant en compte la pente moyenne du contour du pitch à l'échelle de la phrase.

Gillet et King ont proposé une fonction de transformation linéaire par morceaux [GK03] en se basant sur un travail de Patterson [Pat00]. Pour le contour de pitch de chaque phrase, Patterson a calculé quatre fréquences ; "sentence-initial high (S), non-intial accent peaks (H), Post accent valleys (L), and sentence-final low(F)". Ces fréquences sont estimées sur une base d'apprentissage constituée d'une minute de parole, pour avoir quatre fréquences caractéristiques de chaque locuteur. A partir de ces points les auteurs ont proposé une fonction de transformation de pitch. Les quatres fréquences (S,H,L,F), sont calculées pour les locuteurs source et cible, et donc pour une valeur de pitch source f_s la valeur convertie est donnée par :

$$T(f_s) = \begin{cases} F_c + \frac{(f_s - F_s)(L_c - F_c)}{L_s - F_s} & \text{si } f_s < L_s \\ L_c + \frac{(f_s - L_s)(H_c - L_c)}{H_s - L_s} & \text{si } L_s \leq f_s \leq H_s \\ H_c + \frac{(f_s - H_s)(S_c - H_c)}{S_s - H_s} & \text{si } f_s > H_s \end{cases} \tag{4.2}$$

Ces modifications de pitch restent globales, en ce sens qu'elles ne s'appliquent qu'à des caractéristiques du pitch définies sur l'ensemble de la base de données analysée. Par conséquent, elles ne permettent pas de refléter des différences de style prosodique entre les locuteurs source et cible.

L'estimation de caractéristiques plus locales relatives au pitch est beaucoup plus délicate. Une façon d'aborder le problème est de le reformuler en terme d'apprentissage prosodique. L'objectif est alors de relier des informations de type linguistique à des contours de pitch. Ce domaine de recherche a été et demeure toujours à l'origine de nombreux travaux. En synthèse de la parole notamment, l'apprentissage prosodique a pour but de prédire, en fonction du texte à synthétiser, les consignes qui permettront d'associer au message vocalisé une certaine prosodie. Certes, ces techniques automatiques de prédiction permettent de faire ressortir des caractéristiques prosodiques importantes, mais elle ne parviennent tout de même pas à restituer fidèlement la prosodie d'un locuteur.

Dans ce chapitre, nous nous proposons d'étudier la transformation du pitch de manière plus locale. Trois contributions sont proposées dans ce chapitre. Dans un premier temps, nous commençons par une étude préliminaire qui consiste en une conversion du pitch par GMM. Puis, nous proposons deux méthodes de conversion tenant compte de l'interaction entre le pitch et l'enveloppe spectrale. La première est une technique de transformation en deux étapes : tout d'abord, une fonction de transformation du timbre est appliquée aux enveloppes spectrales source, puis, une fonction de prédiction de pitch est appliquée aux enveloppes converties afin d'estimer les valeurs de pitch converties [EnRC03a]. La deuxième méthode consiste à modifier conjointement le pitch et l'enveloppe spectrale [EnRC04a].

4.2 Etude préliminaire : conversion de pitch par GMM

4.2.1 Motivations et principe

Dans cette étude préliminaire, nous essayons de répondre à la question suivante : est-il possible de relier simplement les valeurs de F_0 entre deux locuteurs ?

Les fonction de transformations (4.1) et (4.2) peuvent être considérées comme des transformations par des modèles GMM à une et trois composantes respectivement avec une classification dure, où chaque valeur de fréquence fondamentale issue du locuteur source est affecté à une classe unique par quantification scalaire. Une extension évidente de ces techniques est la modélisation du pitch des locuteurs source et cible par un modèle GMM.

Soient $F^x = [f_1^x, f_2^x, \cdots, f_N^x]$ et $F^y = [f_1^y, f_2^y, \cdots, f_N^y]$ deux séquences de fréquences fondamentales caractérisant le même contenu acoustique prononcé par les locuteurs sources et cibles respectivement. Ces séquences sont obtenues après appariement des trames source et cible par DTW comme mentionné à la section 2.1. Les paires de valeurs de pitch dont l'une est voisée et l'autre non voisée ne sont pas prises en compte.

Pour l'apprentissage d'une telle fonction de conversion de pitch, nous procédons de la même façon que pour la conversion de timbre dans le chapitre précédent, i.e.

nous utilisons un modèle GMM pour modéliser la densité jointe $p(z) = p(f_0^x, f_0^y)$ avec $z = (f_0^x, f_0^y)$. Puis la fonction de conversion de pitch est estimée par régression :

$$F(f_0^x) = \sum_{i=1}^{Q} h_i(f_0^x) \left(\mu_i^y + \rho_i^{xy}(f_0^x - \mu_i^x)/\sigma_i^x\right), \tag{4.3}$$

où $\mu_i = [\mu_i^x, \mu_i^y]$ et $\Sigma_i = \begin{bmatrix} \sigma_i^x & \rho_i^{xy} \\ \rho_i^{xy} & \sigma_i^y \end{bmatrix}$ sont la moyenne et la matrice de covariance de la $i^{\text{ème}}$ gaussienne, et $h_i(f_0^x) = \alpha_i \mathcal{N}(f_0^x; \mu_i^x, \sigma_i^x) / \sum_{j=1}^{Q} \alpha_j \mathcal{N}(f_0^x; \mu_j^x, \sigma_j^x)$ est la probabilité *a posteriori* que f_0^x soit généré par la $i^{\text{ème}}$ classe.

4.2.2 Expérimentation et résultats

Nous avons testé cette méthode dans le cadre d'une conversion homme-femme. Les valeurs de F_0 sont extraites d'une base d'apprentissage de 10 minutes de parole échantillonnée à 16kHz. Le modèle GMM a été implémenté avec un ordre de 1, 2, 4, 8, 16 et 32. La transformation utilisant une seule composante gaussienne correspond à la conversion linéaire simple décrite par l'équation (4.1).

Trouver une distance permettant de quantifier les performances de la conversion du pitch n'est pas une question facile. Dans la littérature, le pitch a toujours été traité comme paramètre suprasegmental, et les chercheurs s'intéressaient plus à son évolution au niveau de la phrase. Or, dans ce travail, le pitch est considéré comme paramètre segmental et la transformation est faite trame par trame.

Dans ce travail, nous avons évalué les performances de la transformation pour les différents ordres de GMM en utilisant la mesure définie par :

$$EN = \sqrt{\frac{\sum_{n=1}^{N}(\hat{F}_{c,n} - F_{c,n})^2}{\sum_{n=1}^{N}(F_{s,n} - F_{c,n})^2}} \tag{4.4}$$

où F_s, F_c et $\hat{F}_c$ désignent le pitch source, cible et convertie respectivement.

Comme nous pouvons le constater sur le tableau 4.1, la distance EN est identique quel que soit l'ordre du mélange. Ces résultats ont été confirmé par des tests d'écoute informels. Cela prouve qu'un "mapping" basé uniquement sur le pitch est illusoire, d'autres informations doivent être prises en compte.

	1	2	4	8	16	32
EN	0.19	0.19	0.19	0.19	0.19	0.19

Tableau 4.1 – Erreur de conversion de pitch en fonction du nombre de composantes GMM.

4.3 Prédiction du pitch à partir de l'enveloppe spectrale

Les études menées par Syrdal et al. dans [SS95] ont montré que le premier formant et le pitch sont liés, ces résultats impliquant que pour conserver une bonne qualité du signal de parole, chaque changement de l'un de ces paramètres doit être accompagné d'une modification appropriée de l'autre. Ainsi, dans le cadre de la synthèse de la parole par concaténation, cette dépendance a été exploitée pour améliorer la qualité de la parole synthétique dans le cas de facteurs de modifications de pitch importants.

En outre, Tanaka [TA97] a proposé d'adjoindre des modifications spectrales en fonction des modifications de pitch. Pour cela, il considère trois classes de pitch (bas, moyen et haut). A chacune de ces classes, il associe un dictionnaire d'enveloppes spectrales par quantification vectorielle. La correspondance entre les vecteurs spectraux et les trois dictionnaires est établie par une technique de "codebook mapping" détaillée dans [ANK88]. Kain et Stylianou ont proposé dans [KS00] une généralisation de cette méthode. A l'aide du modèle GMM les auteurs ont défini une fonction de prédiction de l'enveloppe spectrale à partir du pitch. L'intégration de cet ajustement spectral dans la chaîne de traitements de synthèse conduit à une amélioration notable de la qualité de la parole synthétique.

Ces recherches montrent que la modification conjointe des informations de pitch et de l'enveloppe spectrale est très souhaitable. Dans cette section nous présentons une nouvelle technique de conversion de voix prenant en compte la dépendance entre le pitch et l'enveloppe spectrale. Cette technique de conversion se compose de deux étapes. Tout d'abord une conversion de l'enveloppe spectrale est effectuée. Puis, une fonction de prédiction de pitch à partir de l'enveloppe spectrale est appliquée aux paramètres spectraux convertis afin d'estimer la valeur de pitch convertie.

Nous décrivons tout d'abord la fonction de prédiction avant de détailler son application dans le cadre de la conversion de voix.

4.3.1 Apprentissage de la fonction de prédiction

Pour l'apprentissage de la fonction de prédiction, un enregistrement du locuteur cible est nécessaire. Les vecteurs cepstraux des trames voisées et les valeurs de pitch correspondantes sont extraits du signal par une analyse HNM.

Soient $F = [f_1, f_2, \cdots, f_N]$ une séquence de valeurs de F_0 pour N trames voisées, et $C = [c_1, c_2, \cdots, c_N]$ la séquence de vecteurs cepstraux correspondants. L'approche utilisée dans ce travail est de combiner chaque valeur de pitch avec le vecteur cepstral correspondant, et de modéliser leur densité de probabilité par un modèle GMM.

Avant de décrire plus avant la procédure d'apprentissage, il convient de s'intéresser à la manière de combiner ces deux types d'informations hétérogènes. Trouver les poids à donner à chaque paramètre est un problème difficile. Dans ce travail, nous avons adopté la même stratégie que dans [MHT+01], où les valeurs de pitch sont normalisées selon l'équation :

$$F_{log} = \log\left(\frac{F_0}{\bar{F}_0}\right), \tag{4.5}$$

où F_0 est la fréquence fondamentale en Hz, et $\bar{F}_0$ est la moyenne des valeurs de pitch sur toute la base d'apprentissage.

Les paramètres cepstraux c sont ensuite combinés au pitch normalisé noté g, afin de modéliser la densité conjointe $p(z) = p(c, g)$ à l'aide d'un modèle GMM. Les paramètres (α, μ, Σ) de ce modèle sont estimés à l'aide de l'algorithme EM décrit précédemment. Une fois le modèle appris, la fonction de prédiction de pitch est donnée par :

$$F(c) = \mathbb{E}(F|C = c) = \sum_{i=1}^{Q} h_i(c)\left[\mu_i^f + \Sigma_i^{fc}(\Sigma_i^{cc})^{-1}(c - \mu_i^c)\right], \tag{4.6}$$

où

$$h_i(c) = \frac{\alpha_i \mathcal{N}(c, \mu_i^c, \Sigma_i^{cc})}{\sum_{j=1}^{Q} \alpha_j \mathcal{N}(c, \mu_j^c, \Sigma_j^{cc})} \tag{4.7}$$

est la probabilité *a posteriori* que c soit généré suivant la $i^{\text{ème}}$ gaussienne.

La procédure d'apprentissage de la fonction de prédiction est résumée dans la figure 4.1.

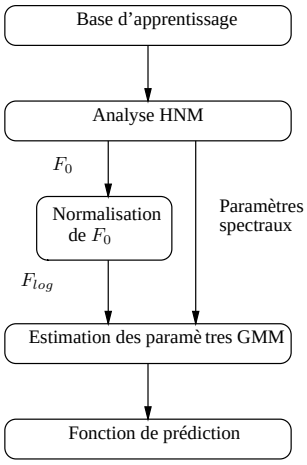


Figure 4.1 – Apprentissage de la fonction de prédiction du pitch à partir de l'enveloppe spectrale.

4.3.2 Résultats et discussion

Nous avons testé la prédiction de pitch sur plusieurs locuteurs. A titre d'exemple, nous présentons dans ce paragraphe quelques résultats de l'application de cet algorithme pour une voix de femme. Les résultats obtenus sur d'autres locuteurs sont similaires.

Nous avons étudié les performances de la technique de prédiction de pitch en fonction de la taille de la base d'apprentissage et du nombre de composantes GMM. Pour la mise en oeuvre de cette méthode nous disposons de 15 minutes de parole échantillonnée à 16 kHz. Cette base a été divisée en deux bases de 10 et de 5 minutes une pour l'apprentissage et l'autre pour le test. De la base d'apprentissage, nous avons constitué 6 bases de 1, 2, 3, 4, 6 et 10 minutes de parole.

Seules les trames voisées sont prises en compte pour l'apprentissage ce qui représente environ 70% de la base initiale. Les vecteurs cepstraux utilisés étant de taille 20, le premier coefficient cepstral qui est relatif à l'énergie de la trame acoustique n'est pas pris en compte lors de l'apprentissage. Nous avons fait varier le nombre de composantes

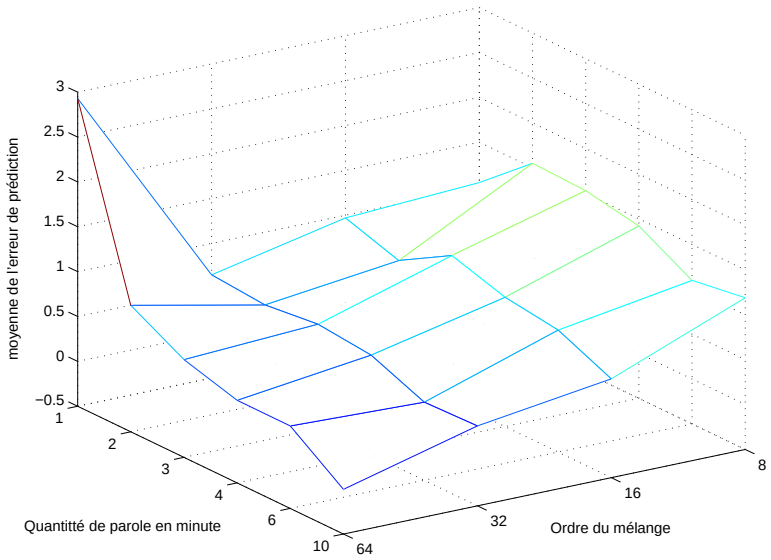


Figure 4.2 – Moyenne de l'erreur de prédiction en fonction de la taille de la base d'apprentissage et du nombre de composantes gaussiennes.

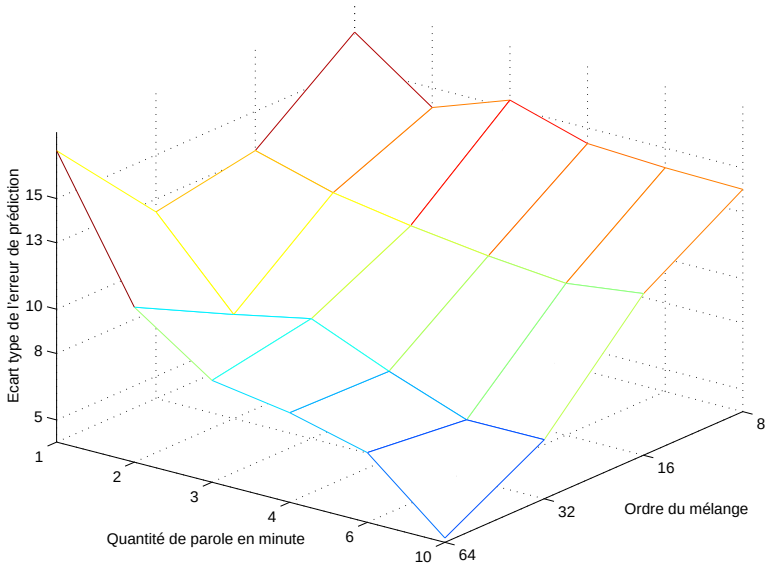


Figure 4.3 – Écart type de l'erreur de prédiction en fonction de la taille de la base d'apprentissage et du nombre de composantes gaussiennes.

GMM entre 8 et 64 par puissance de 2.

Afin d'évaluer les performances de la fonction de prédiction, nous avons utilisé une distance quadratique moyenne entre les valeurs de pitch prédites et celles calculées à partir du signal. Les figures 4.2 et 4.3 montrent les résultats de la prédiction de pitch en fonction de la taille de la base d'apprentissage et du nombre de composantes GMM. Ces résultats montrent clairement que les performances de la fonction de prédiction augmentent avec la taille de la base d'apprentissage. Notons qu'en utilisant une base d'apprentissage de 1, 2, 3, 4 et 6 minutes de parole, la moyenne et l'écart type de l'erreur sont assez élevés. Avec une base d'apprentissage suffisante (10 minutes), L'erreur de prédiction décroît rapidement avec le nombre de de composante GMM. Ces résultats indiquent qu'une base d'apprentissage d'au minimum 10 minutes de parole est nécessaire pour avoir des résultats de prédiction satisfaisants.

Nous avons étudié de plus près la fonction de prédiction en utilisant une base d'apprentissage de 10 minutes de parole avec 64 composantes GMM. Dans un premier temps, nous omettons la normalisation du pitch présentée précédemment. La figure 4.4 montre la distribution des valeurs de pitch ainsi que les moyennes des gaussiennes pour chaque composante du mélange. Nous associons chaque valeur de pitch à la classe maximisant la probabilité *a posteriori* que le vecteur cepstral correspondant soit généré par cette classe conformément à l'équation (4.7). Cette figure fait certes apparaître une certaine dépendance entre pitch et enveloppe spectrale, mais montre aussi une variance élevée du pitch au sein de chaque classe, ainsi qu'un fort recouvrement des valeurs de pitch entre classes distinctes. Après normalisation du pitch conformément à l'équation (4.5), l'estimation des paramètres de la fonction de prédiction du pitch est nettement améliorée, comme le souligne la figure 4.5.

Le tableau 4.2 montre les résultats obtenus sur trois intervalles de pitch. L'erreur de prédiction est globalement faible, la moyenne de l'erreur de prédiction étant de 0.02 Hz et l'écart type de 4.2 Hz. Cependant en ne considérant que les valeurs de pitch dans l'intervalle [150-250 Hz] qui contient 87.4% de données, cet écart type est réduit à 2.5 Hz. En revanche, pour les valeur extrêmes de pitch, cet écart type augmente à 28.5 Hz. Cette dégradation de performance est liée, d'une part, au fait que pour ces valeurs extrêmes, peu d'exemples étaient présents dans la base d'apprentissage.

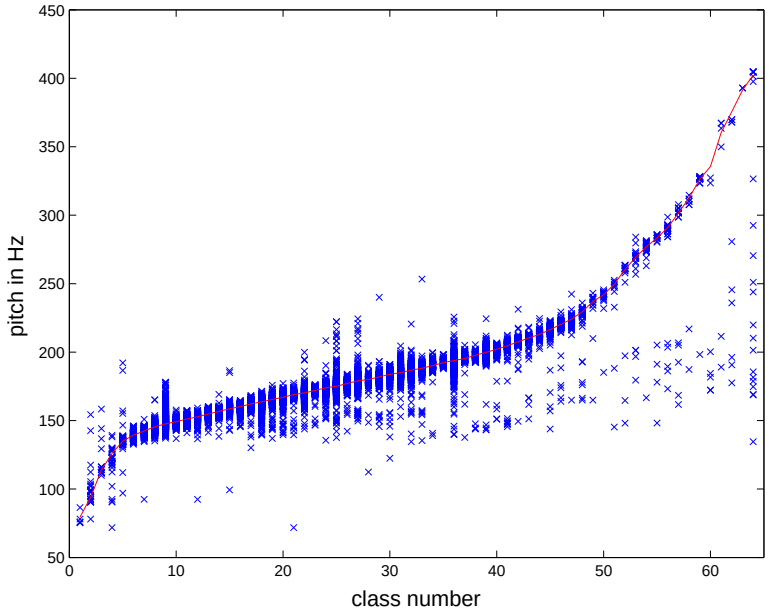


Figure 4.4 – Distribution des valeurs de pitch observées autour des moyennes des gaussiennes : apprentissage sans normalisation.

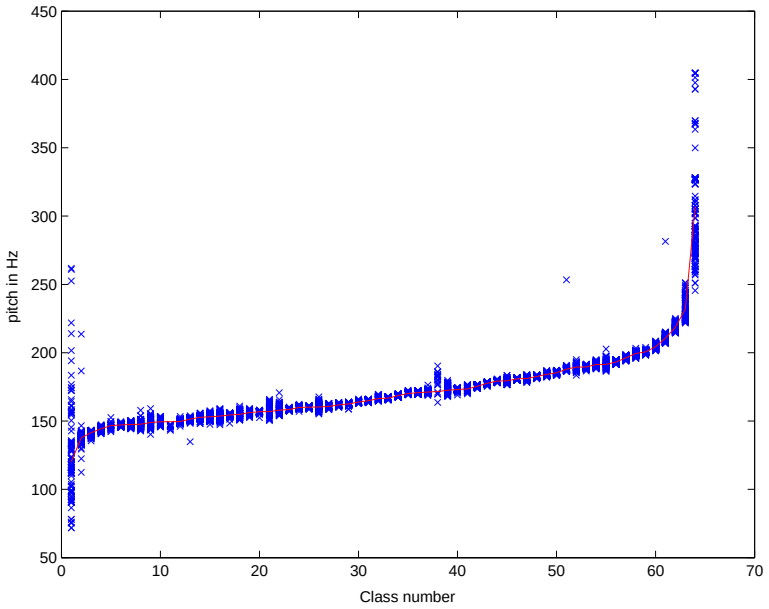


Figure 4.5 – Distribution des valeurs de pitch observées autour des moyennes des gaussiennes : apprentissage avec normalisation.

Ainsi l'algorithme d'apprentissage n'a pas réussi dans ce cas à mettre en évidence une corrélation vraiment marquée entre le pitch et l'enveloppe spectrale. D'autre part, certaines erreurs sont imputables à des mauvaises estimation du pitch. En effet, la méthode d'estimation de pitch utilisée (voir section 3.3) peut conduire dans certains cas, à des valeurs qui sont des multiples ou des divisions des valeurs de pitch réelles. La prédiction à partir des enveloppes spectrales estimées en utilisant ces mauvaises valeurs de pitch conduit à des valeurs prédites erronées.

F_0	$<$ 150 Hz	150-250 Hz	$>$ 250 Hz	Total
Moyenne (Hz)	0.6	-0.1	0.6	-0.02
Écart Type (Hz)	4.7	2.5	28.5	4.2
Fréquence relative (%)	11.4	87.4	1.2	100

Tableau 4.2 – Moyenne et écart type de l'erreur de prédiction de pitch pour une voix de femme.

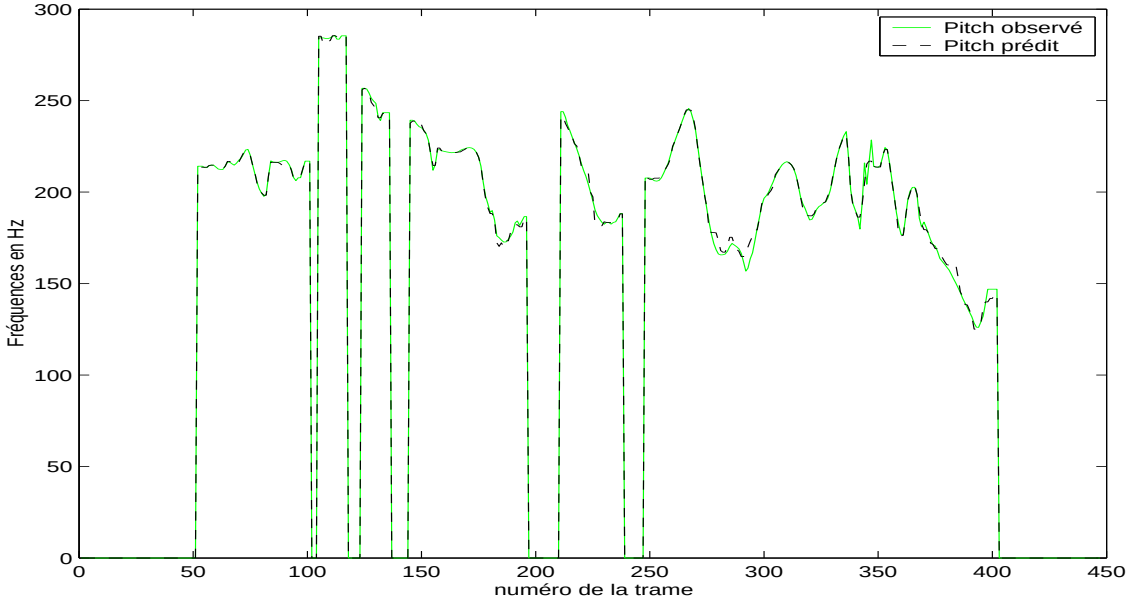


Figure 4.6 – Contours de pitch observé (ligne continue) et prédit (ligne discontinue) pour la phrase : " Ils ont tous obtenu leur C.A.P. en juillet dernier".

La figure 4.6 montre un exemple de résultat de prédiction sur une phrase de la base de test. En traits continus, nous avons représenté le contour de pitch de la phrase origi-

nale et en traits discontinus le contour de pitch prédit à partir des vecteurs cepstraux de chaque trame. La méthode que nous avons proposée permet de reproduire de manière assez fidèle le contour du pitch d'un locuteur à partir uniquement de l'information d'enveloppe spectrale.

4.3.3 Application à la conversion de voix

Après avoir décrit l'algorithme de prédiction de pitch, nous nous intéressons à son application dans le domaine de la conversion de voix. La figure 4.7 présente l'architecture d'un système de conversion de voix permettant la conversion du timbre ainsi que la prédiction du pitch à partir du timbre transformé.

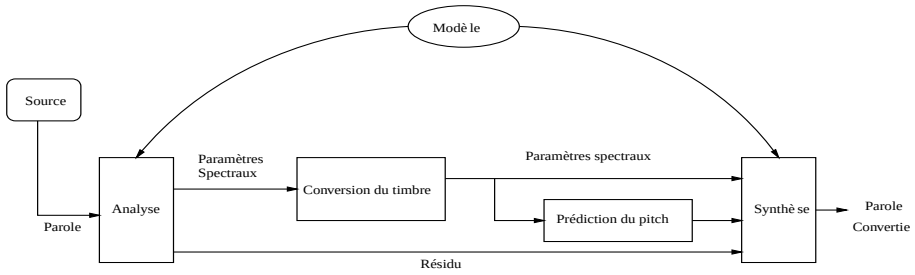


Figure 4.7 – Système de conversion de voix associant modifications du timbre et prédiction de pitch.

Après l'analyse du signal source, une fonction de transformation du timbre est appliquée aux paramètres spectraux source de manière à approcher le timbre de la cible. Puis, pour les trames voisées, la fonction de prédiction du pitch, apprise sur une base d'apprentissage du locuteur cible, est appliquée afin d'estimer le pitch du locuteur cible. L'enveloppe spectrale et le pitch modifiés sont récupérés par un module de synthèse pour produire la parole transformée. Notons que l'étape d'analyse et de synthèse sont les mêmes que dans le chapitre précédent. En pratique, dans le cadre d'une application en synthèse de la parole, l'analyse se fait de manière off-line et ne fait donc pas partie du système de conversion.

4.3.3.1 Evaluation et discussion

Nous avons testé cette méthode dans le cadre de la transformation homme-femme, en utilisant une base d'apprentissage de 10 minutes de parole pour chaque locuteur. Les paramètres cepstraux et le pitch sont extraits à l'aide d'une analyse HNM. Pour la transformation du timbre, nous avons utilisé la transformation par GMM décrite dans le chapitre précédent avec 64 composantes. Nous avons appris la fonction de prédiction de pitch sur la voix cible en utilisant 64 composantes GMM.

Lorsque l'enveloppe transformée est proche de la cible, les résultats de prédiction du pitch sont satisfaisants. En revanche, toute erreur de transformation du timbre se répercute automatiquement sur le pitch prédit, ce qui rend cet algorithme peu robuste, et limite son intérêt pour la conversion de voix.

La figure 4.8 montre un exemple de pitch prédit à partir des enveloppes spectrales transformées. La courbe de pitch prédit a une allure très fluctuante et est peu conforme à celle que l'on pourrait observé sur un signal de parole naturelle. Pour lisser cette courbe un filtrage médian a été effectué, mais ce dernier n'a pas permis d'améliorer significativement les résultats.

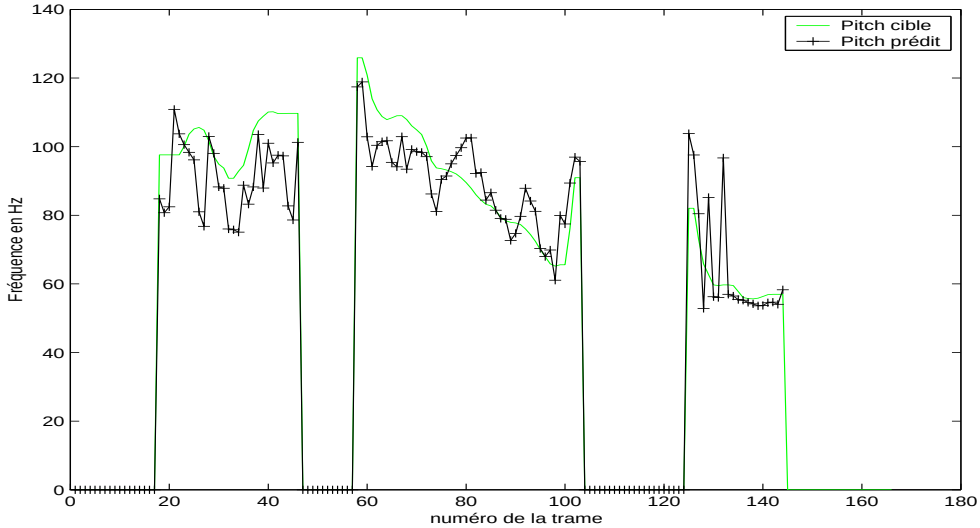


Figure 4.8 – Exemple de mauvais résultats de prédiction : Contours de pitch cible (ligne simple) et prédit (ligne avec des +) pour la phrase : " édition de luxe".

Ces expériences ont montré qu'il existe une forte dépendance entre le pitch et l'enveloppe spectrale, ce qui permet d'estimer de façon satisfaisante le pitch à partir uniquement d'information d'enveloppe spectrale. En revanche, l'intégration de l'algorithme de prédiction proposé dans un système de conversion de voix s'avère peu concluante. Cependant, cet algorithme de prédiction de pitch peut être envisagé pour d'autre contextes applicatifs. Parmi eux, notons la reconnaissance de la parole distribuée (DSR : Distributed Speech Recognition) [SR03]. Dans cette application, la reconstruction de la parole reconnue par le système de reconnaissance à partir des coefficients MFCC est, dans certains cas, nécessaire. A titre d'exemple, nous pouvons citer :

- Les services de réponse interactifs basés sur la reconnaissance des informations "sensibles" comme pour les transactions bancaires ou le courtage. Les paramètres issus de la DSR peuvent être stockés pour une future vérification par des humains, ou pour répondre à des exigences de légalité ;
- La vérification par des humains des expressions de la base de parole collectée par un système DSR. Une telle base de données peut être employée pour améliorer les performances du système ;
- La reconnaissance automatique assistée par des humains.

Dans ce cadre, Shao et al. [SM04] ont utilisé cette méthode de prédiction de pitch pour reconstruire de la parole à partir des coefficients MFCC. Elle a été utilisée pour la décision de voisement et pour la prédiction de pitch et semble donner de bons résultats.

Cette méthode peut également être utilisée pour la correction automatique du pitch. En effet, en apprenant la fonction de prédiction du pitch à partir des paramètres MFCC, par exemple, et des valeurs de pitch vérifiées manuellement. Les nouvelles valeurs de pitch calculées par une méthode d'estimation peuvent être comparées à celles obtenues par prédiction. En outre, la fonction de prédiction semble être multi-locuteurs, ce qui peut ouvrir la voix pour son application en codage de la parole.

4.4 Transformation conjointe du pitch et de l'enveloppe spectrale

Ayant mis en évidence la corrélation entre enveloppe spectrale et fréquence fondamentale, nous nous proposons de traiter ces deux informations de façon conjointe dans un contexte de conversion de voix. Dans cette section, nous proposons d'utiliser deux fonctions de transformation basées sur le modèle GMM : une fonction pour les trames voisées prenant en compte la conversion du pitch et de l'enveloppe spectrale, et une fonction pour les trames non voisées prenant en compte la transformation de l'enveloppe spectrale uniquement.

Apprentissage des fonctions de transformation

Avant de faire l'apprentissage proprement dit des fonctions de transformation, il faut tout d'abord créer les bases d'apprentissage des fonctions de transformation pour les trames voisées et pour les trames non voisées. Ce pré-traitement se fait en plusieurs étapes :

- Analyse des corpus d'apprentissage source et cible ;
- Alignement temporel des vecteurs cepstraux source et cible par un algorithme de DTW [RJ93] ;
- Séparation des couples de vecteurs spectraux alignés en deux bases d'apprentissage, une pour les trames voisées et une autre pour les non voisées.

Lors de l'étape d'alignement, seules les enveloppes spectrales sont prises en compte. Les trames alignées dont l'une est voisée et l'autre est non voisée ne sont pas prises en compte lors de l'apprentissage des paramètres GMM.

Pour les trames voisées, le modèle GMM sera appliqué à des vecteurs de paramètres comprenant à la fois les paramètres cepstraux et la fréquence fondamentale. Donc, comme dans la section 4.3.1, nous effectuons une normalisation des valeurs de pitch conformément à l'équation (4.5). Notons que la normalisation des valeurs de pitch de la source et de la cible est effectuée en utilisant la moyenne des valeurs de pitch de la source et de la cible respectivement.

Pour l'apprentissage de la fonction de transformation des trames voisées, nous supposons avoir deux suites de vecteurs de paramètres $X = [x_1, x_2, \cdots, x_N]$ et $Y = [y_1, y_2, \cdots, y_N]$ source et cible temporellement alignées. Chaque vecteur x (ou y) est obtenu en incorporant le vecteur des coefficients cepstraux c_x et le pitch normalisé g_x du locuteur source ; $x = [c_x^T, g_x]^T$ ($.^T$ étant l'opérateur de transposition).

Après apprentissage des modèles GMM, la fonction de transformation est estimée de la même manière qu'à la section 3.4.2 du chapitre 3. i.e.

$$F(x) = \mathbb{E}(Y|X = x) = \sum_{i=1}^{Q} h_i(x) \left[\mu_i^y + \Sigma_i^{xy}(\Sigma_i^{xx})^{-1}(x - \mu_i^x)\right]. \tag{4.8}$$

La figure 4.9 décrit le processus d'apprentissage de la fonction de transformation conjointe pour les trames voisées, ce processus se compose de trois étapes principales :

- Normalisation des valeurs de pitch source et cible et construction de vecteurs d'apprentissage par combinations des valeurs des pitch normalisées et des vecteurs cepstraux correspondant ;
- Estimation des paramètres GMM par un algorithme EM initialisé à l'aide d'un algorithme de VQ ;
- Estimation de la fonction de conversion par régression.

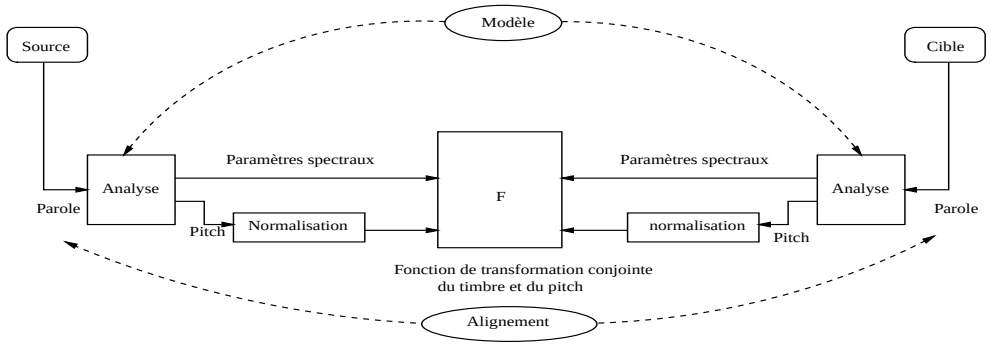


Figure 4.9 – Apprentissage de la transformation conjointe du pitch et de l'enveloppe spectrale pour les trames voisées.

Pour les trames non voisées, seuls les paramètres de l'enveloppe spectrale sont transformés. L'étape d'apprentissage reste la même dans la mesure où le modèle GMM est

utilisé pour l'estimation de la densité jointe des paramètres spectraux source et cible ce qui conduit à une fonction de conversion similaire à la fonction (4.8).

4.4.1 Transformation

Après l'analyse, les trames de la parole source sont transformées par l'une ou l'autre des fonctions de transformation préalablement apprises selon qu'elles sont voisées ou non. Pour les trames non voisées, seule l'enveloppe spectrale est modifiée en utilisant la fonction de transformation pour les trames non voisées.

La transformation des trames voisées est expliquée sur la figure (4.10). Le pitch est normalisé conformément à l'équation (4.5), puis combiné avec l'enveloppe spectrale comme décrit dans la sous section 4.4, avant d'appliquer la fonction de transformation des trames voisées. Le pitch normalisé ainsi converti $\hat{g}_y$ est alors modifié selon :

$$\hat{f}_y = \bar{f}_y \exp(\hat{g}_y). \tag{4.9}$$

où $\bar{f}_y$ désigne la valeur moyenne du pitch du locuteur cible. Puis le pitch $\hat{f}_y$ et les paramètres cepstraux $\hat{c}_y$ sont utilisés pour produire la parole convertie.

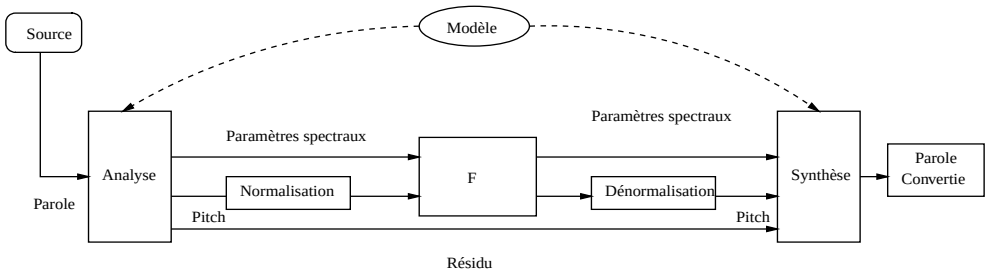


Figure 4.10 – Système de conversion de voix permettant la transformation conjointe du pitch et de l'enveloppe spectrale.

4.4.2 Expérimentation

Dans cette section, nous présentons les expériences effectuées en vue de l'évaluation de la méthode de conversion conjointe. Nous utilisons les mêmes bases de données que dans le chapitre précédent pour construire les bases d'apprentissage et de test.

Les bases d'apprentissage sont analysées de manière asynchrone avec un pas d'analyse de 10 ms et des trames de 20 ms. Notons que seules les données d'apprentissage sont analysées d'une manière asynchrone. Pour l'application de la fonction de conversion de voix l'analyse est effectuée d'une manière pitch-synchrone afin d'obtenir des modifications prosodiques de haute qualité. Nous avons utilisé des vecteurs cepstraux d'ordre 20. Le premier coefficient cepstral c_0 n'est pas pris en compte pour l'apprentissage de la fonction de transformation.

Après alignement, nous obtenons 30000 couples de vecteurs, ce qui correspond à environ 5 minutes de parole. Les couples de vecteurs alignés sont divisés en deux bases d'apprentissage suivant qu'ils sont voisés ou non. Les couples de trames telles que l'une est voisée et l'autre non voisée ne sont pas prises en compte. Pour chaque trame voisée, le pitch est normalisé conformément à l'équation (4.5), puis combiné avec les coefficients cepstraux.

L'estimation des paramètres GMM est effectué à l'aide de l'algorithme EM initialisé par une technique de quantification vectorielle classique. L'apprentissage est mené en utilisant des matrices de covariances pleines. Pour éviter des singularités, une petite valeur à été ajoutée aux élément diagonaux des matrices de covariances après chaque itération. Pour chaque base d'apprentissage, nous faisons varier le nombre de composantes du mélange en considérant les puissances de 2 comprises entre 8 et 64. Lors de ces expériences, 20 itérations sont jugées suffisantes pour atteindre la convergence de l'algorithme EM.

Comme dans le chapitre précédent, nous avons utilisé, lors de l'apprentissage, des matrices de covariances pleines. La figure 4.11 présente un exemple de matrice de covariance en valeur absolue de la source dans le cas de la conversion conjointe. La valeur la plus élevée de cette matrice est égale à 0.63, 63% des éléments de cette matrice ont des valeurs supérieures à 0.25 et seulement 4.2% ont des valeurs inférieures à 0.01. Dans cet exemple, l'énergie de la diagonale ne représente que 13% de l'énergie totale de cette matrice. Cet exemple souligne également l'importance de la corrélation entre le pitch et l'enveloppe spectrale.

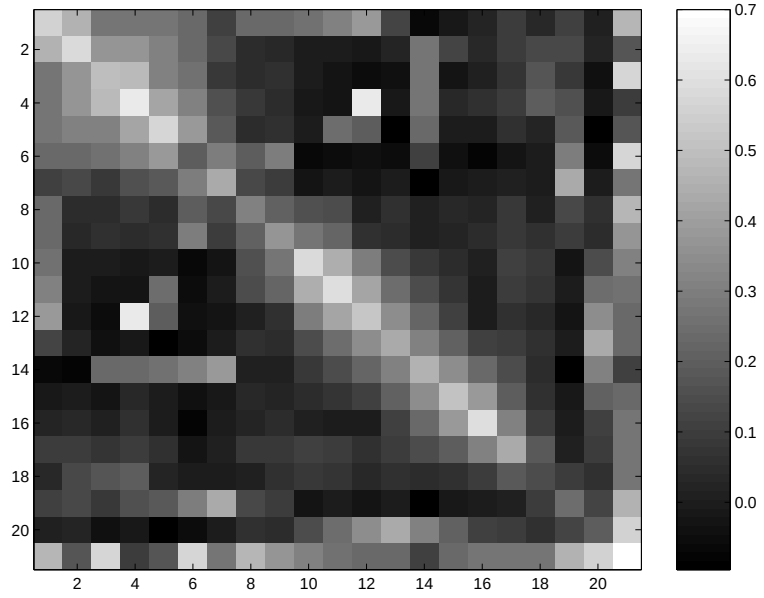


Figure 4.11 – Matrice de covariance de la source pour une des composantes du modèle GMM, obtenue après apprentissage de la densité conjointe.

4.4.2.1 Evaluation objective

Pour évaluer objectivement les performances de la transformation conjointe du timbre et de l'enveloppe spectrale, la détermination de mesures de performances objectives est nécessaire. Le problème de la détermination d'une mesure de distorsion spectrale à été étudié dans la section (3.6.1).

Pour l'évaluation objective de la conversion de pitch, nous avons utilisé une distance quadratique moyenne entre les valeur de pitch normalisées (DPN pour Distorsion de Pitch Normalisé) définie comme suit :

$$DPN = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{g_n^y - \hat{g}_n^y}{g_n^y - g_n^x}\right)^2}, \tag{4.10}$$

où g_n^x, $\hat{g}_n^y$ et g_n^y sont les valeurs de la fréquence fondamentale source, convertie et cible respectivement, normalisées conformément à l'équation (4.5).

La figure 4.12 présente un exemple de conversion de pitch entre une voix d'homme

et une voix de femme par l'application de la technique de conversion conjointe avec 64 composantes gaussiennes, et par conversion linéaire simple. La conversion conjointe du timbre et du pitch offre une transformation de pitch assez satisfaisante. Par comparaison, une simple mise à l'échelle du pitch offre une transformation linéaire destinée à respecter le pitch moyen du locuteur cible (équation 4.1) mais ne permet pas de refléter des différences notables entre les formes des contours source et cible.

La figure 4.13 montre l'évolution de la distorsion de pitch normalisé (DPN) en fonction du nombre de composantes gaussiennes comparée à une transformation linéaire simple. Quel que soit le nombre de composantes, la DPN obtenue par la conversion conjointe est inférieure à celle obtenue par une conversion linéaire simple classique. Contrairement aux résultats obtenus lors de l'étude préliminaire, la DPN décroît avec le nombre de composantes. Ce résultat met en évidence l'importance du rôle de la corrélation entre le pitch et l'enveloppe spectrale.

Pour mesurer la proximité entre les paramètres spectraux convertis et ceux de la cible, nous avons utilisé l'erreur normalisée (EN) décrite dans la section (3.6.1) du chapitre (3) basée sur une distance dans le domaine spectrale, i.e. la distance donnée par :

$$d_{DS} = \frac{1}{N} \sum_{i=1}^{N} \|P_{dB}(A) - P_{dB}(B)\|^2, \tag{4.11}$$

où $P_{dB}(A)$ désigne la densité spectrale en échelle de Bark issue de A exprimée en dB échantillonnée sur 512 points. Les densités spectrales P_{dB} ont été calculées en mettant le premier coefficient cepstral c_0 à zero pour annuler l'influence de l'énergie sur les résultats.

Comme nous pouvons le constater sur la figure 4.14, la prise en compte de la corrélation entre le pitch et l'enveloppe spectrale rend la modification du timbre plus robuste. En effet, pour le même nombre de composantes la distorsion spectrale obtenue par la conversion conjointe est inférieure à celle obtenue par la conversion spectrale seule. Les résultats de la transformation des trames non voisées sont aussi satisfaisants comme le montre la figure 4.15.

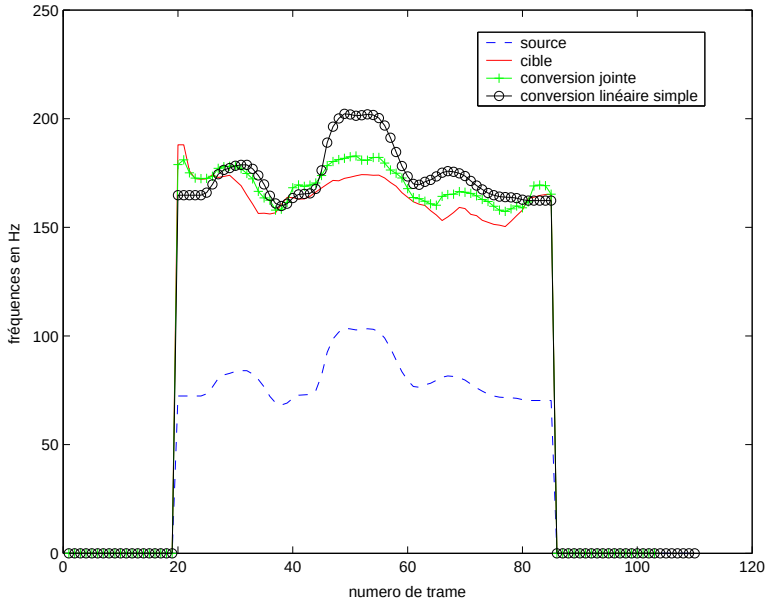


Figure 4.12 – Exemple de contours de pitch transformés pour un modèle GMM à 64 composantes.

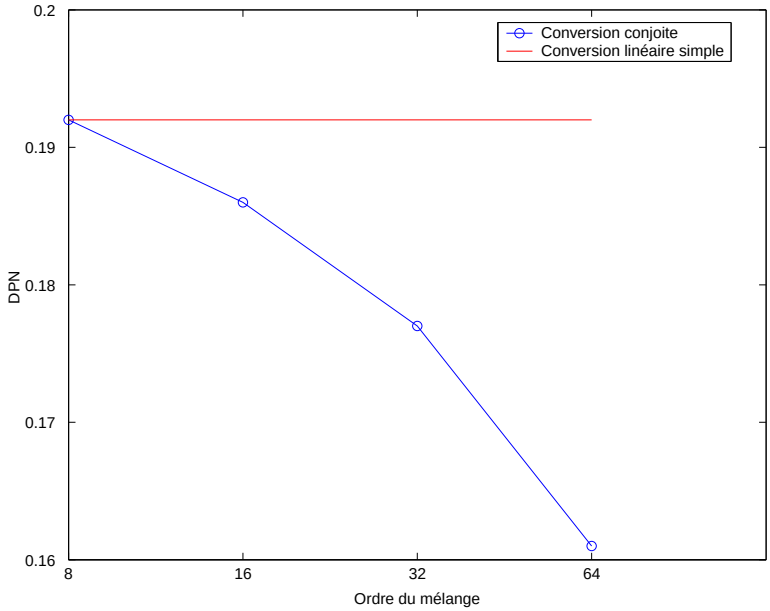


Figure 4.13 – Erreur de conversion de pitch en fonction du nombre de composantes GMM.

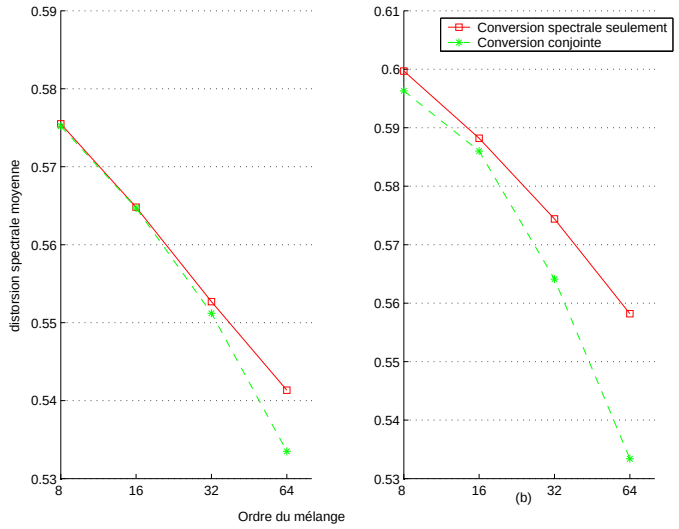


Figure 4.14 – Distorsion spectrale moyenne pour les sons voisés en fonction du nombre de composantes GMM pour une transformation femme-homme (a) et homme-femme (b).

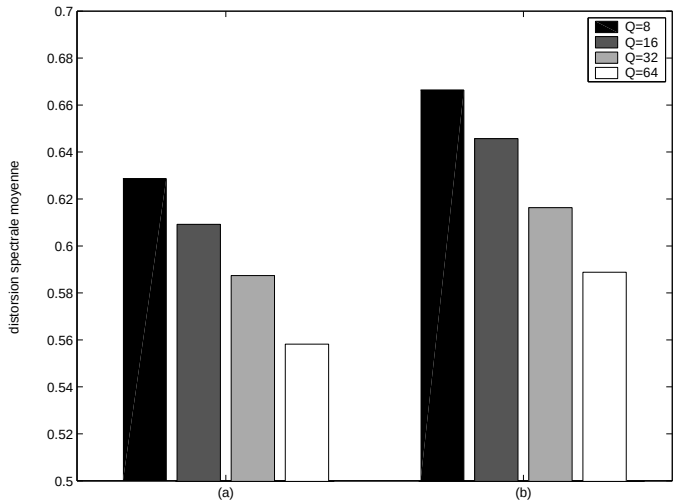


Figure 4.15 – Distortion spectrale moyenne pour les sons non voisés en fonction du nombre de composantes GMM, pour des transformation homme-femme (a) et femme-homme (b).

4.4.2.2 Evaluation subjective

Pour évaluer notre système sur le plan perceptif, nous avons organisé des tests d'écoute. Le test d'écoute le plus utilisé pour valider un système de conversion de voix est le test ABX (c'est un test dans lequel des auditeurs jugent si le son prononcé par un locuteur "X" est plus proche du locuteur "A" ou de "B"). Dans le cas de conversions inter-genre, c'est-à-dire, la conversion homme/femme et femme/homme ce type de test se révèle cependant inadéquat. En effet, dès que le système arrive à simuler le genre cible, la voix convertie est jugée plus proche de la cible que de la source.

Pour évaluer la qualité de la parole convertie par la méthode proposée, nous l'avons comparé aux trois configurations suivantes :

- *Naturelle* : parole naturelle sans modification ;
- *Conversion classique* : parole transformée par la technique de conversion présentée dans le chapitre 3, cette technique, servant de référence ;
- *Plaquage acoustique* : parole source sur laquelle nous avons calqué le timbre et le pitch de la cible. Le plaquage acoustique peut être considéré comme la borne supérieure de la transformation, dans la mesure où il reflète la qualité accessible par le modèle HNM lorsque une modification idéale a été effectuée.

Nous avons regroupé une douzaine d'auditeurs et nous les avons fait passer trois séries de tests. Le premier test a pour objectif la comparaison de notre système et le système de référence, c'est-à-dire, la conversion classique. Les deux derniers tests ont pour objectifs l'étude des limites que peut atteindre une conversion de pitch et de timbre. Dans ces deux derniers, nous comparons la conversion conjointe avec le plaquage acoustique, puis la conversion conjointe et le plaquage acoustique avec la parole naturelle.

Pour l'ensemble de ces trois tests, 20 phrases ont été utilisés. Les trois tests ont été effectués dans les mêmes conditions et sont basées sur le même principe. L'écoute des stimuli est réalisée avec un casque audio professionnel dans un bureau calme, les sujets pouvant ajuster le niveau d'écoute à leur convenance. Les phrases de test sont présentées aux auditeurs dans un ordre aléatoire, chaque phrase n'étant présentée qu'une seule fois. Après avoir écouté une phrase, les sujets doivent noter sa qualité globale sur une échelle

MOS [1] à cinq niveaux. Cette échelle va de un, pour la plus mauvaise qualité, jusqu'à cinq pour une qualité excellente. La présentation des stimuli et l'entrée des notes est informatisée : après l'écoute d'un stimulus, la note est entrée à l'aide du clavier de l'ordinateur. En cas d'entrée incohérente, le sujet est sollicité jusqu'à ce que la nouvelle entrée soit correcte, ce qui permet alors la présentation du stimulus suivant dans le test. Afin que les sujets puissent étaler leur notation sur toute l'échelle de valeur, une session d'apprentissage est dispensée avant le test. Cet apprentissage consiste à présenter au sujets 6 stimuli du test couvrant l'étendue de la gamme de qualité des phrases de test sans avoir à les noter. La session d'apprentissage peut être re-présentée lorsqu'un sujet en fait la demande. De plus, afin d'évaluer la cohérence de chaque sujet, une phrase par configuration est présenté deux fois. Les auditeurs qui donnent à la même phrase de test deux notes différentes d'au moins deux points sont éliminés. Malgré le soin que nous avons apporté à l'élaboration de ce protocole d'écoute, nous sommes conscients des limites de ce test qui demeure très subjectif : les auditeurs ne jugent pas tous selon les mêmes critères.

Résultats du test 1

L'examen préliminaire des notes montre que les auditeurs semble avoir adopté des stratégies d'évaluation similaires. En effet, les répartitions des notes sont homogènes : aucun sujet ne semble avoir un comportement aberrant.

Technique de conversion	conjointe	classique
MOS	3.63	2.44

Tableau 4.3 – MOS obtenus par la conversion conjointe et la conversion classique.

Note	conjointe > classique	conjointe = classique	classique > conjointe
Fréquence relative (%)	71,25	26,25	2,5

Tableau 4.4 – Résultats des tests de comparaison entre conversion conjointe et conversion classique.

En considérant les MOS moyens, il s'avère que les sujets préfèrent la conversion conjointe. En effet, comme le montre le tableau 4.3 la conversion conjointe a obtenu un MOS égal

[1] Mean Opinion Score

à 3.63 contre 2.44 pour la conversion classique. En considérant les paires de notes attribuées par sujet et par phrase, 71,25% donnent la préférence à la conversion conjointe, 26,25% des phrases ont obtenu la même note pour les deux configurations, et seulement 2,5% donnent la préférence à la conversion classique. En effet, la technique de conversion conjointe permet de réduire le nombre des artéfacts, ce qui fait d'elle une technique plus robuste que la conversion classique. Les résultats de ce test confirment ceux obtenus par les tests objectifs.

Résultats du test 2

Dans ce test, nous avons comparé la qualité de la parole convertie par le méthode de la conversion conjointe à sa borne supérieure, c'est-à-dire le plaquage acoustique.

Technique de conversion	conjointe	plaquage
MOS	2.85	3.25

Tableau 4.5 – MOS obtenus par la conversion conjointe et le plaquage acoustique.

Note	conjointe > plaquage	conjointe = plaquage	plaquage > conjointe
Fréquence relative (%)	17.4	38	43.5

Tableau 4.6 – Résultats des tests de comparaison entre conversion conjointe et plaquage acoustique.

Le tableau 4.5 montre les MOS moyens obtenus par la conversion conjointe et le plaquage acoustique. Les auditeurs donnent en moyenne un demi point de plus pour le plaquage acoustique. Cependant, en examinant les paires de notes par phrase et par sujet, les auditeurs attribuent dans 38% des cas la même note aux deux transformations, et dans 17,4% des cas ils préfèrent la conversion conjointe. Sur les 43,5% restant, 95% ont seulement un point de différence avec leur homologues.

De ce test il ressort que, bien que la technique de conversion conjointe apporte des améliorations sensibles par rapport à la conversion classique, la qualité de la parole convertie par la conversion conjointe reste légèrement inférieure à celle obtenue par un plaquage acoustique.

Résultats du test 3

Dans ce test, nous avons introduit des phrases de parole naturelle afin d'évaluer la qualité absolue de la parole convertie.

Technique de conversion	Conjointe	Plaquage	Naturelle
MOS	2.76	3.22	5

Tableau 4.7 – MOS obtenus par la conversion conjointe et le plaquage acoustique.

Note	conjointe > plaquage	conjointe = plaquage	plaquage > conjointe
Fréquence relative (%)	12.8	37,2	48

Tableau 4.8 – Résultats des tests de comparaison entre conversion conjointe et plaquage acoustique.

La décision de laisser la comparaison avec la parole naturelle au dernier test est prise dans l'objectif d'éviter l'appréciation des stimuli de test par rapport à la parole naturelle. Cependant, ce choix a influencé les résultats du dernier test. En effet, les auditeurs n'ont pas tenu compte de l'apprentissage. Ils ont attribué la note 5 automatiquement à la parole naturelle et ont continué à noter les deux autres configurations sur les même critères que dans les tests précédents.

Les résultats du test 3 confirment ceux obtenus par le test 2. Les auditeurs accordent une légère préférence au plaquage acoustique par rapport à la conversion conjointe (tableau 4.8 et 4.7). Ceci montre que malgré l'amélioration de la qualité de parole convertie qu'apporte la conversion conjointe, celle-ci ne permet pas d'atteindre la qualité du plaquage acoustique. Mais le test 3 montre surtout une différence de qualité considérable entre signaux transformés et signaux naturels. En effet, la différence des MOS entre la parole naturelle et plaquage acoustique est beaucoup plus importante que celle entre le plaquage acoustique et la conversion conjointe. Ceci suggère que le problème vient plutôt de la modélisation utilisée que de la technique de conversion elle-même. En effet, la synthèse de la partie harmonique transformée est réalisée en utilisant le résidu de la voix source. Cette non adéquation du résidu avec les enveloppes spectrales transformées se traduit par un bruit gênant à l'écoute. Pour remédier à ce problème il convient de trouver une technique de transformation du résidu en tenant compte de sa dépendance

avec le pitch et l'enveloppe spectrale. Dans [Kai01] l'auteur a proposé d'utiliser une technique de prédiction linéaire entre le résidu et l'enveloppe spectrale. Or, d'une part, cette modélisation n'est pas très fidèle au mécanisme de production de la parole. D'autre part, cette prédiction semble très difficile dans la mesure où les études statistiques que nous avons réalisées n'ont pas permis de dégager une dépendance nette entre ces paramètres. Une autre approche consisterait à utiliser un modèle de déconvolution source/filtre plus performant que le modèle HNM, i.e. permettant une bonne séparation entre le signal glottique et le conduit vocale. On peut par exemple mentionner des méthodes telles que [Alk92, DKA95, Lob01], bien qu'elles puissent être complètement erratique.

4.5 Conclusion

Dans ce chapitre nous nous sommes intéressés à la transformation de la fréquence fondamentale en tenant compte de sa dépendance avec le timbre. Ayant mis en évidence la dépendance entre le pitch et l'enveloppe spectrale, nous l'avons exploitée à des fins de conversion.

Nous avons mis en oeuvre deux techniques de conversion tenant compte de cette dépendance. La première de ces transformations procède en deux étapes : tout d'abord, une fonction de transformation de timbre est appliquée aux enveloppes spectrales source, puis, une fonction de prédiction de pitch est appliquée aux enveloppes converties afin d'estimer les valeurs de pitch convertie. Malgré les bonnes performances de la technique de prédiction du pitch, son association avec une technique de conversion du timbre par GMM conduit à un système de conversion de voix qui manque de robustesse. En effet, lorsque l'enveloppe spectrale transformée est proche de la cible, les résultats de prédiction sont satisfaisants. En revanche, toute erreur de transformation du timbre se répercute automatiquement sur le pitch prédit.

La deuxième technique consiste à transformer conjointement l'enveloppe spectrale et le pitch. Les tests objectifs et subjectifs effectués ont montré que cette conversion conjointe améliore sensiblement la qualité de la parole convertie. En particulier, la transformation conjointe du pitch et de l'enveloppe spectrale rend globalement robuste l'opération de conversion. Cependant des marges de progrès ont été identifiés. En effet,

bien qu'apportant des améliorations significatives par rapport à l'état de l'art, notre technique ne parvient pas à atteindre la qualité de la parole naturelle. La comparaison avec des phrases sur lesquelles nous avions plaqué le timbre et le pitch de la voix cible a montré que cette technique a atteint sa limite. Les défauts présents dans la parole convertie viennent plutôt des paramètres non transformés, à savoir, la fréquence maximale de voisement F_c, les phases et le gain de la partie bruitée qui demeurent inchangées lors de la conversion. Or, ces paramètres contiennent encore de l'information importantes du point de vue de la perception. Parmi elles, les phases semblent particulièrement cruciales, car le manque d'adéquation entre les phases source et cible se traduit par des dégradations audibles et un manque de naturel de la parole convertie. Ces problèmes liés à la phase sont à rapprocher au manque de contrôle explicite que nous avons sur les paramètres liés à la production du signal glottique. L'issue semble résider dans une approche plus globale où l'interaction entre source glottique et conduit vocale serait mieux prise en compte.

Chapitre 5

Mise en oeuvre de la conversion de voix

5.1 Introduction

Dans ce chapitre nous nous intéressons à l'application de la conversion de voix dans le cadre de la synthèse de la parole à partir du texte.

Dans la première partie de ce chapitre, nous traitons le problème de l'intégration de la conversion de voix dans un système de synthèse de la parole à partir du texte. En effet, deux approches peuvent être employées pour appliquer le processus de conversion : modifier la base acoustique source de manière à créer une nouvelle voix qui sera employée ultérieurement dans le système TTS, ou intégrer la fonction de transformation dans le synthétiseur, si toutefois cette dernière n'augmente pas de façon rédhibitoire la complexité du système de synthèse. Cette approche est intéressante dans la mesure où elle permet de réduire drastiquement la taille des bases de données nécessaires à la synthèse. En effet une seule voix de référence doit être stockée (ce qui représente quelques centaines de méga-octets pour les systèmes actuels de synthèse par corpus) ainsi qu'une fonction de transformation pour chaque nouvelle voix (quelques centaines de kilo-octets par voix). Ainsi, l'intégration de la fonction de conversion dans un synthétiseur s'avère très intéressante, notamment dans le cas de systèmes embarqués.

Cependant, les systèmes de conversion de voix actuels sont complexes. Par exemple, pour l'intégration d'une fonction de conversion de voix basée sur les GMM dans un système de synthèse de la parole par HNM, la tâche de conversion de voix à une complexité de 1.5 à 2 fois plus élevée que celle de la synthèse elle-même. Donc, il y a un vrai besoin de diminuer le coût de calcul de la conversion de voix. A cette fin, nous présentons une méthode de conversion de voix simplifiée, permettant de réduire la complexité de la fonction de conversion par GMM par un facteur compris entre 40 et 130 tout en gardant la même qualité qu'une conversion classique par GMM [EnRC04b].

Dans la deuxième partie de ce chapitre, nous traitons le problème de parallélisme du corpus d'apprentissage. En effet, comme nous l'avons évoqué au chapitre 2, l'apprentissage d'une fonction de conversion nécessite deux corpus de parole source et cible parallèles, i.e. contenant le même contenu phonétique. L'acquisition de telles bases est très délicate dans la mesure où il faut que les deux locuteurs prononcent le même texte et réalisent des chaînes phonétiques identiques. Or, il est évident que la conversion de voix ne trouvera de véritable essor (en dehors de la synthèse vocale) que si sa mise en oeuvre peut être simplifiée de manière drastique. A ce titre, le fait de pouvoir apprendre une voix cible à partir d'un enregistrement quelconque du locuteur cible est d'une importance particulière. Actuellement, il n'existe pas de solution convaincante permettant, à partir d'un enregistrement quelconque d'un locuteur cible, d'effectuer de la conversion de voix. Dans ce travail, nous proposons une méthode permettant de contourner la contrainte de parallélisme du corpus d'apprentissage.

5.2 Réduction de la complexité de la conversion

5.2.1 Complexité de la conversion par GMM

Nous nous somme intéressés à la fonction de conversion par GMM décrite dans le chapitre précédent ; la densité jointe de la source x et la cible y est modélisée par un modèle GMM. Puis, la fonction de conversion est donnée par la régression :

$$\begin{aligned}\hat{y} = F(x) &= \mathbb{E}(Y|X=x) \\ &= \sum_{i=1}^{Q} h_i(x)\left[\mu_i^y + \Sigma_i^{xy}(\Sigma_i^{xx})^{-1}(x-\mu_i^x)\right], \end{aligned} \tag{5.1}$$

où

$$h_i(x) = p(q|x) = \frac{\alpha_i \mathcal{N}(x;\mu_i^x,\Sigma_i^{xx})}{\sum_{j=1}^{Q}\alpha_j \mathcal{N}(x;\mu_j^x,\Sigma_j^{xx})}. \tag{5.2}$$

est la probabilité *a posteriori* que x soit généré par la $i^{\text{ème}}$ gaussienne.

Vu les performances de conversion qu'elle offre, cette méthode est devenue un standard dans le domaine de conversion de voix. Cependant, cette méthode de transformation s'avère coûteuse pour des applications en temps réel. La détermination d'un vecteur converti requiert le calcul de toutes les probabilités *a posteriori* (5.2) pour chacune des composantes du mélange, puis, l'évaluation du vecteur transformé conformément à l'équation (5.1). Le volume des calculs est récapitulé dans le tableau 5.1.

équation	Coût de calcul	
	$\times$	$+$
(5.1)	$Q(p^2+p)$	$Q(p^2+2p)$
(5.2)	$Q(p^2+2p)$	$Q(p^2+2p)$
Total	$Q(2p^2+3p)$	$Q(2p^2+4p)$

Tableau 5.1 – Coût des calculs de la méthode de conversion en utilisant la fonction (5.1) : ($\times$) nombre de multiplications, ($+$) nombre d'additions, où p indique la dimension des vecteurs de données à transformer et Q le nombre de composantes du mélange.

5.2.2 Simplification algorithmique

La fonction de conversion donnée dans l'équation (5.1) est une fonction composée d'une somme pondérée de Q fonctions linéaires dont les poids sont les probabilités *a posteriori* de chaque classe. Cependant, si l'apprentissage des paramètres du GMM est fait correctement, les poids de chaque composante du mélange sont très différents. Plus précisément, seulement quelques composantes ont une probabilité *a posteriori*

significatives. Ce phénomène est illustré à la figure 5.1 représentant les histogrammes des probabilités *a posteriori* cumulées pour un nombre de composantes $M = 1$ et $M = 3$. Ainsi, en pratique il n'est pas utile de calculer la somme sur toutes les classes dans l'équation (5.1). Partant de ce constat, nous proposons de limiter les opérations de conversion aux composantes du mélange ayant un poids significatif, c'est-à-dire de retenir un nombre M restreint de composantes ayant les probabilités *a posteriori* les plus élevées. Dans ce cas, la fonction de transformation s'écrit comme :

$$F(x) = \sum_{i \in A} \omega_i(x) \left[\mu_i^y + \Sigma_i^{yx} (\Sigma_i^{xx})^{-1} (x - \mu_i^x) \right], \tag{5.3}$$

avec

$$\omega_i(x) = \frac{h_i(x)}{\sum_{j \in A} h_j(x)}, \tag{5.4}$$

où A est l'ensemble des M indices des composantes choisies et où les $h_i(x)$ sont les probabilités *a posteriori* (5.2). Quand $M = 1$, cette méthode revient à utiliser l'estimateur de maximum de probabilité *a posteriori* (MAP).

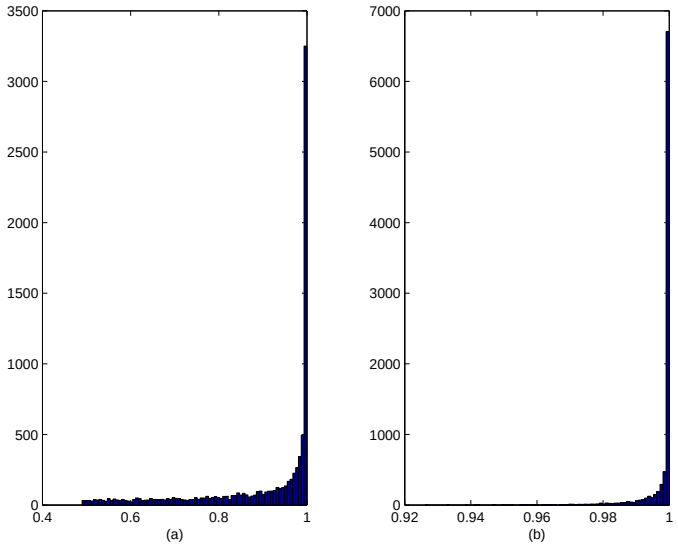


Figure 5.1 – Distribution des probabilités a posteriori cumulées pour (a) $M = 1$ et (b) $M = 3$.

Avantageusement, lorsque les données à convertir sont des données enregistrées, ce qui est le cas par exemple dans une application de synthèse de la parole, la technique

décrite ci-dessus peut être mis en oeuvre de manière "off-line". En effet, dans ce cas, pour des raisons de complexité algorithmique notamment, il est judicieux d'effectuer la classification et la sélection des composantes utiles lors de la préparation de la base de données source. Ceci permet d'éviter l'évaluation de l'équation (5.2)) pour chaque composante GMM lors de la conversion.

Le traitement "off-line" nécessite certes de stocker les indices et les poids des composantes sélectionnées du mélange pour chaque trame d'analyse dans le dictionnaire acoustique, mais a l'avantage de limiter drastiquement la complexité de la conversion. En effet, l'opération de conversion ne requiert que $M(p^2 + p)$ multiplications et $M(p^2+2p)$ additions. Sans prendre en compte le calcul des exponentielles, la complexité de conversion est ainsi réduite d'un facteur de l'ordre de $2Q/M$.

5.2.3 Expérimentation

Ces expérimentations sont effectuées dans les mêmes conditions que dans la section 4.4.2. Notons que la simplification algorithmique est testée uniquement dans le cadre de la transformation conjointe avec 64 composantes GMM.

Pour évaluer cette simplification algorithmique sur le plan objectif, nous avons utilisé la distorsion spectrale normalisée suivante :

$$DSN = \frac{\sum_{n=1}^{N} \|P_{dB}(y_n) - P_{dB}(\hat{y}_n)\|_2}{\sum_{n=1}^{N} \|P_{dB}(y_n) - P_{dB}(x_n)\|_2}, \tag{5.5}$$

où P_{dB} est la distance spectrale décrite dans la section (3.6.1). Les densités spectrales P_{dB} ont été calculées en mettant le premier coefficient cepstral c_0 à zero pour annuler l'influence de l'énergie sur les résultats.

Les résultats sont présentés sur la figure 5.2. Les courbes de la distorsion DSN ont des allures similaires, les résultats obtenus par la méthode de conversion classique, (i.e. la conversion conformément à l'équation (5.1)) et la conversion utilisant les trois composantes les plus probables sont très proches. Pour la méthode MAP la DSN est légèrement supérieure.

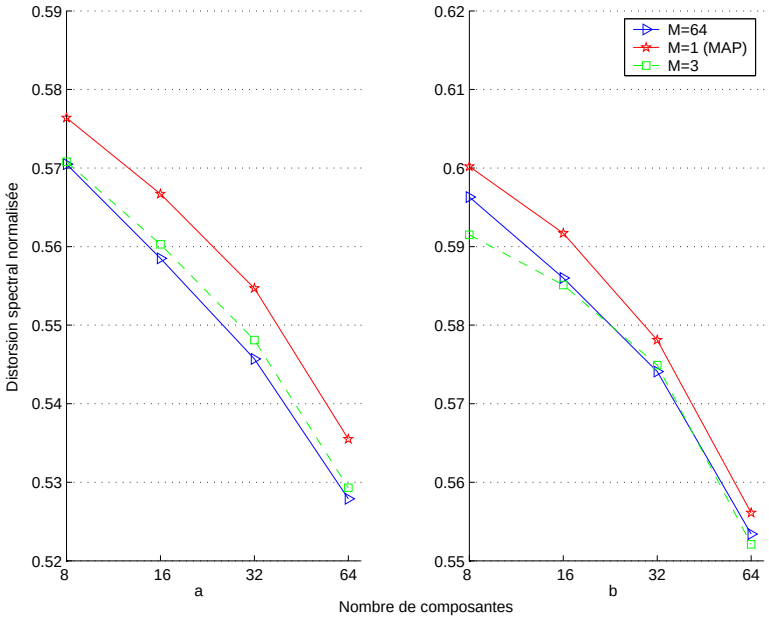


Figure 5.2 – Distorsion spectrale normalisée pour la méthode de conversion classique, ainsi que pour la nouvelle méthode pour $M = 3$ et $M = 1$ (MAP). (a) : conversion femme-homme, (b) : conversion homme-femme.

Avec $Q = 64$ la complexité est divisée par 45 dans le cas $M = 3$ et par environ 130 dans le cas du MAP ($M = 1$).

Notons que la mesure DSN ne donne qu'une idée globale sur les performances de la transformation. Pour avoir des informations plus locales, nous avons calculé la distance spectrale trame par trame. Pour cela, nous avons utilisé la distortion spectrale relative (DSR), définie par :

$$DSR = \frac{\|P_{dB}(y_n) - P_{dB}(\hat{y}_n)\|_2}{\|P_{dB}(y_n) - P_{dB}(x_n)\|_2}. \tag{5.6}$$

La figure 5.3 présente l'histogramme de la DSR sur la base de test. Nous constatons que des erreurs grossières peuvent apparaître localement. Cependant, ces erreurs sont plus probables dans le cas du MAP. Par exemple, la probabilité que la distance entre les paramètres convertis et la cible soit plus grande que celle entre la source et cible (i.e. la probabilité que DSR > 1) est de 3% pour une conversion classique, 3.6% pour une

conversion avec la méthode proposée dans le cas où $M = 3$ et de 5% pour la conversion avec la méthode MAP.

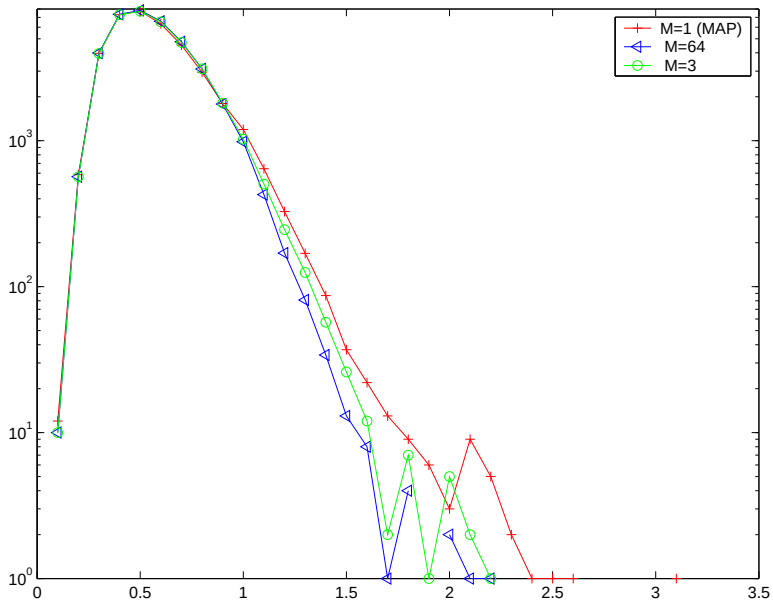


Figure 5.3 – Histogrammes de la DSR trame par trame pour la méthode classique, ainsi que pour la nouvelle méthode pour $M = 3$ et $M = 1$ (MAP) dans le cas d'une conversion homme-femme.

Les tests d'écoute effectués ont montré que la parole convertie obtenue avec la méthode proposée pour $M = 3$ composantes et la méthode classique sont indiscernables. La robustesse de la méthode proposée peut s'expliquer par le fait que la somme des trois plus grandes probabilités *a posteriori* est toujours plus grande que 0.9 comme nous pouvons le voir sur la figure (5.1). Par contre, la méthode MAP conduit à des artéfacts locaux qui ont été perçus comme gênants à l'écoute. Ces artéfacts se produisent généralement quand la probabilité *a posteriori* de la classe sélectionnée ne dépasse pas 0.6, ce qui correspond au cas où une seule composante ne peut pas modéliser correctement l'enveloppe spectrale de la trame.

Notons cependant que ces stratégies de sélection avec un nombre fixe de composantes ne sont pas optimales, aussi le choix d'un nombre variable de composantes sera préférable. Pour optimiser la conversion, nous pouvons songer à utiliser la technique MAP quand la probabilité *a posteriori* est assez élevée (i.e. supérieure à 0.9), au lieu

d'utiliser systématiquement la technique avec $M = 3$ composantes. Donc, un raffinement de la méthode proposée consiste à retenir un nombre minimum de composantes dont les probabilités *a posteriori* cumulées dépassent un certain seuil. Par exemple avec un seuil fixé à 0.9, cette stratégie permet d'utiliser une, deux ou trois composantes pour 59%, 27% et 14% des trames de la base de tests. Comparée à la conversion classique, cette méthode offre une réduction de complexité d'un facteur entre 45 et 130, en fonction du nombre de composantes GMM sélectionnées. En moyenne ce facteur de réduction est de 100. A l'écoute, cette méthode conduit sensiblement aux mêmes résultats que la méthode avec $M = 3$ composantes. Avec un tel facteur de réduction, le volume de calcul lié à la conversion devient négligeable comparé à celui de la synthèse.

5.3 Apprentissage à partir de corpus non parallèles

5.3.1 Limitation des méthodes actuelles

L'utilisation de corpus d'apprentissage parallèles peut se révéler problématique pour la mise en oeuvre de la conversion de voix. En effet, une question se pose d'emblée : comment qualifier le fait que les corpus source et cible puissent être considérés comme parallèles ? Généralement, seules les correspondances phonétiques entre enregistrements des deux locuteurs sont vérifiées. Celles-ci sont bien entendu nécessaires pour pouvoir effectuer un appariement de trames issues des locuteurs source et cible. Mais elles demeurent toutefois insuffisantes, car d'autres différences entre les deux enregistrements peuvent se révéler importantes sur le plan de la perception. Par exemple, des différences d'intonation peuvent apparaître entre les deux locuteurs. Si ces dernières se manifestent de façon systématique, c'est-à-dire si elles permettent de mettre globalement en relief des différences de style prosodique, alors elles n'auront pas de répercussions négatives sur la mise en correspondance des deux locuteurs. En revanche, si ce phénomène apparaît de manière peu reproductible, alors une partie des données d'apprentissage des deux locuteurs risque d'être mal appariée et la fonction de conversion sera alors mal estimée. Pour espérer obtenir une bonne qualité de conversion, il apparaît donc crucial d'introduire un mécanisme de vérification et le cas échéant de correction de l'enregis-

trement du locuteur cible.

En outre, le fait de pouvoir utiliser des enregistrements non parallèles pour effectuer l'apprentissage de fonctions conversion de voix est bien entendu très intéressant sur le plan de la mise en oeuvre. Ainsi, parallèlement au développement des méthodes de conversion basées sur l'utilisation de corpus parallèles, d'autres travaux ont été menés afin de rendre possible la conversion dans le cas où les corpus source et cible ne sont pas parallèles. Ces travaux sont très largement inspirés des techniques d'adaptation au locuteur classiquement utilisées en reconnaissance de la parole par modèles de Markov cachés. Une application intéressante a été proposée dans [YTM+03], où le module d'adaptation au locuteur permet de personnaliser un système de synthèse par HMM. Dans un premier temps, une classification des modèles HMM en contexte par arbre de décision est réalisée pour construire un modèle de voix "moyenne". Ensuite, les paramètres de ces modèles HMM sont adaptés en fonction du locuteur cible. Des tests tant objectifs que subjectifs ont certes montré l'utilité de la méthode dans le cadre de la synthèse par HMM. Mais la qualité de la parole convertie accessible par les systèmes de synthèse par HMM est néanmoins très médiocre.

Dans [MSM04], une technique d'adaptation au locuteur est également proposée pour de la conversion de voix basée sur des corpus non-parallèles. Dans cette application, les auteurs font l'hypothèse que deux corpus parallèles A et B sont disponibles. Pour réaliser la conversion entre les corpus non parallèles source C et cible D, ils supposent en outre que les corpus C et D sont parallèles respectivement à une partie des corpus A et B. Dans ce cas, ils expriment la fonction de conversion entre les locuteurs C et D comme la composée de trois fonctions de conversion, respectivement des locuteurs C vers A, A vers B et B vers D. Le cadre d'application semble assez restrictif, car cette méthode requiert tout de même des portions d'enregistrement parallèles. De plus, aucun mécanisme permettant de contrôler le parallélisme des corpus utilisés n'est proposé. Enfin, la composition des trois fonctions de conversion risque d'entraîner des erreurs de conversion importantes. Au final, la qualité de la parole convertie obtenue par cette méthode est jugée moins bonne que celle obtenue à partir de corpus parallèles.

De ce bref état de l'art, il ressort que d'une part, la mise en oeuvre de la conversion de voix sur des corpus parallèles peut se révéler difficile et que d'autre part, les techniques

de conversion de voix sur corpus non parallèles conduisent actuellement à une qualité de parole jugée médiocre.

5.3.2 Méthode proposée

La solution proposée dans ce travail consiste à utiliser des échantillons de parole générés par un système de synthèse par corpus. En effet, nous supposons avoir une base de données acoustiques d'un locuteur de référence adaptée à la synthèse de la parole à partir du texte. Cette base acoustique est utilisée par un système de synthèse pour créer d'une manière artificielle un corpus parallèle à un corpus donné.

Selon l'application désirée, la voix de référence peut servir comme source ou cible. Par exemple, la transformation de la parole d'un locuteur de référence vers un locuteur cible correspond à la personnalisation d'un système de synthèse par corpus. Une fonction de conversion permettant la transformation de la parole d'un locuteur source vers un locuteur de référence peut, quant à elle, être utilisée pour unifier les messages vocaux issus de différents opérateurs humains, par exemple dans le cadre de services de type centre d'appels.

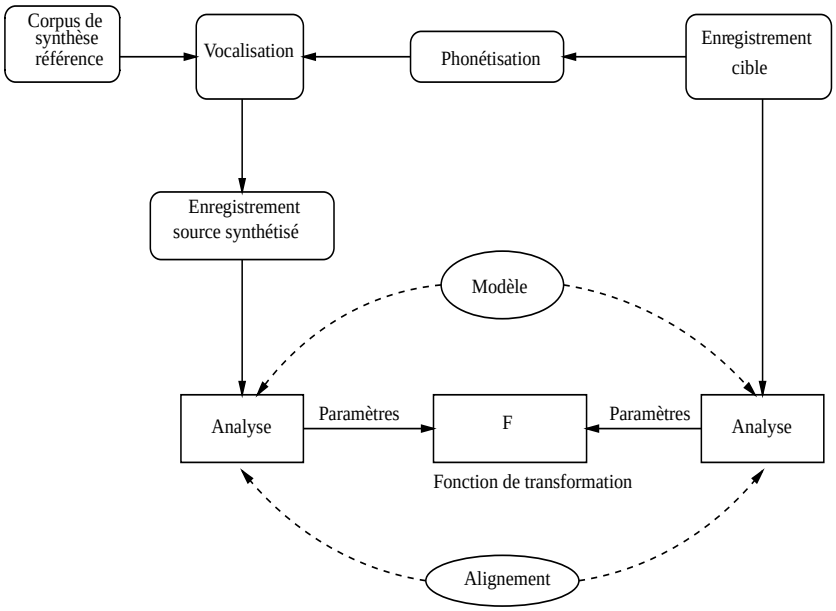


Figure 5.4 – Apprentissage de la fonction de transformation sur corpus non parallèles.

La figure 5.4 montre le processus d'apprentissage de la fonction de transformation en utilisant une voix de synthèse comme enregistrement source. La première étape de ce processus est la phonétisation de l'enregistrement de la voix cible. Les chaînes phonétiques résultantes sont vocalisées par un système de synthèse par corpus. Les enregistrements synthètiques ainsi générés sont utilisés comme voix source pour estimer une fonction de conversion entre la voix de référence et la cible. L'estimation proprement dite de la fonction de conversion est réalisée comme définie dans le chapitre précédent.

Plus généralement, dans des applications où la source et la cible sont bien déterminées, la voix de référence peut être utilisée comme un lien entre les locuteurs source et cible. Deux fonctions de conversion sont alors estimées, d'une part entre les locuteurs source et référence et d'autre part entre les locuteurs référence et cible. Ainsi, une fonction permettant la conversion source-cible désirée peut être obtenue en composant les deux fonctions de transformation source-référence et référence-cible.

$$F_{source-cible}(x) = F_{reference-cible}\ o\ F_{source-reference}(x). \tag{5.7}$$

5.3.3 Expérimentation et résultats

Dans cette section, nous présentons les résultats des expérimentations que nous avons effectuées en vue de l'évaluation de la conversion avec des corpus d'apprentissage synthétique dans le cadre de la conversion conjointe avec 64 composantes GMM.

Dans ces expérimentations, nous avons utilisé des enregistrements synthétiques comme enregistrements source. Pour générer ces enregistrements, nous avons utilisé le système de synthèse par corpus de la division R&D de France Télécom. Lors de la génération de l'enregistrement synthétique, nous avons contraint le synthétiseur à reproduire la phonétisation réalisée par le locuteur cible. Nous avons comparé les performances des fonctions de transformation dans le cadre de la conversion femme-homme et homme-femme dans le cas où les corpus d'apprentissage source et cible sont issus des voix naturelles, dite cas des corpus parallèles, et le cas où le corpus d'apprentissage source est généré par le système de synthèse, dite cas des corpus non parallèles.

5.3.3.1 Evaluation objective

Pour mesurer les performances de cette approche sur le plan objectif, nous avons utilisé la distorsion spectrale normalisée décrite dans le section précédente (équation (5.5)). Les résultats de ces mesures sont présentés sur le tableau 5.2.

	P	NP
femme-homme	0.53	0.54
homme-femme	0.53	0.55

Tableau 5.2 – Comparaison des distorsions spectrales normalisées entre vecteurs spectraux converti et cible en utilisant des corpus parallèles (P) et non parallèles (NP)

Ces résultats montrent que les performances d'une fonction de conversion apprise sur des corpus synthétiques sont comparables sur le plan objectif, avec ceux obtenus par apprentissage sur des corpus naturels.

5.3.3.2 Evaluation subjective

Pour confirmer les résultats obtenus par les tests objectifs, nous avons effectué deux tests d'écoute. Le premier est un test de qualité globale. Il a été effectué dans les mêmes conditions que les tests d'écoute décrits dans le chapitre précédent et suivant le même protocole (section 4.4.2.2). Les résultats de ce test sont résumés dans les tableaux 5.3 et 5.4. Les auditeurs donnent presque le même score aux deux configurations. En considérant les paires de notes par phrase et par sujet, dans 41% des cas la conversion par corpus parallèles obtient un point de plus que la conversion avec corpus non parallèle. Dans 38% des cas, les deux configurations sont jugées équivalentes. Sur les 21% restants, les auditeurs préfèrent la conversion avec corpus non parallèles.

Dans un deuxième test, nous avons présenté aux auditeurs des séries de trois stimuli. Le premier stimulus correspond aux phrases cible, et les deux restants correspondent aux phrases converties par les deux méthodes présentées dans un ordre aléatoire. Les auditeurs étaient invités à dire laquelle des deux dernières phrases est plus proche de la première phrase. Les réponses étaient complètement aléatoires. Les auditeurs n'arrivent pas à faire la différence entre les deux configurations.

Corpus d'apprentissage	P	NP
MOS	2.98	2.75

Tableau 5.3 – Les MOS obtenus par la conversion avec corpus source naturel et avec corpus source synthétique.

Note	P > NP	P = NP	NP > P
Fréquence relative (%)	41	38	21

Tableau 5.4 – Résultats des tests de comparaison entre l'apprentissage sur des corpus parallèles (P) et non parallèles (NP).

Notre méthode permet donc d'obtenir des résultats de conversion comparables à ceux obtenus à partir de corpus parallèles, tant du point de vue de la qualité des signaux transformés que de la proximité de la parole convertie par rapport à la cible. La méthode proposée garantit donc une qualité acceptable tout en simplifiant considérablement l'apprentissage de la fonction de conversion pour un nouveau locuteur.

5.4 Conclusion

Dans ce chapitre nous avons abordé des aspects pratiques relevant de la mise en oeuvre de la conversion de voix dans un système de synthèse de la parole à partir du texte.

Nous avons proposé une méthode de conversion basée sur les GMM, qui consiste à restreindre la fonction de conversion aux composantes GMM les plus représentatives. Cette étude a montré que la méthode MAP, qui consiste à la simplification maximale du nombre de classes prises en compte par la conversion, n'est pas recommandable dans la mesure où cette dernière peut induire des artéfacts locaux perçus comme gênants. Pour rendre cette réduction plus robuste, nous avons présenté une autre variante de cette méthode consistant à utiliser un nombre variable de composantes GMM. Les résultats obtenus ont montré que les performances de cette méthode sont comparables à ceux obtenus par une conversion GMM classique.

Nous avons également étudié l'utilisation des bases d'apprentissage générées par un système de synthèse de la parole à partir du texte. Les résultats des tests effectués ont montré que la qualité de la parole convertie par une fonction de transformation apprise sur des enregistrements synthétiques est comparable à celle de la parole convertie par une fonction de transformation apprise sur des enregistrements naturels. Ainsi, cette technique simplifie la mise en oeuvre de la conversion de voix en permettant l'apprentissage d'une nouvelle voix cible à partir d'un enregistrement quelconque du locuteur cible.

Des expériences supplémentaires mériteraient d'être menées afin d'évaluer cette méthodologie dans le cas beaucoup plus général où l'on dispose de deux corpus source et cible non parallèles. Plus précisément, il s'agirait de comparer les performances d'une telle conversion effectuée en deux étapes (conversion source-référence puis référence-cible) à celle d'une transformation directe source-cible apprise sur des corpus parallèles. Ne disposant pas des bases de données adéquates, nous n'avons pas pu mener ces expériences dans le cadre de cette thèse.

Conclusion

Le travail réalisé au cours de cette thèse a porté sur le développement d'un système de conversion de voix. L'objectif est de modifier le timbre et la prosodie d'un signal émis par un locuteur de sorte que des auditeurs aient l'impression d'entendre parler un autre individu déterminé au préalable.

Synthèse du travail

Pour développer un système de conversion de voix, une bonne connaissance des mécanismes de production de la parole et des paramètres acoustiques caractérisant l'identité du locuteur est nécessaire. Ce fut l'objet du premier chapitre. Le deuxième chapitre a précisé les principes d'un système de conversion de voix et dressé un état de l'art des méthodes de conversion. La plupart des systèmes de conversion existants dans la littérature traitent du problème de la conversion de l'enveloppe spectrale. Les techniques le plus récentes tournent autour de la conversion par GMM. Cette technique a été proposée par Stylianou dans [Sty96b] et améliorée par Kain dans [Kai01]. Les tests comparatifs menés dans [BS96] ont montré que cette technique est celle qui offre les meilleurs résultats.

Dans le chapitre 3 nous avons étudié de plus près cette technique de conversion. En effet, la conversion par GMM a été appliquée indépendamment à la modification du cepstre discret [Sty96b] et au paramètres LSF [Kai01] sans qu'aucune comparaison véritable n'ait été effectuée. Notre contribution dans ce chapitre consistait ainsi à analyser les performances accessibles par ces deux paramétrisations. Les tests effectués ont

montré que ces deux modélisations offrent des performances équivalentes. Cependant, l'utilisation des LSF dans la conversion de voix nécessite des contrôles supplémentaires des vecteurs convertis. En effet, lors de la modification de ces paramètres, des problèmes numériques peuvent apparaître. Par exemple, l'ordonnancement des vecteurs LSF n'est pas respecté pour 7% des vecteurs utilisés pour les tests. Le non ordonnancement d'un vecteur LSF se traduit par l'instabilité du filtre AR résultant, et conduit à un bruit très gênant à l'écoute.

L'indépendance entre la fréquence fondamentale et l'enveloppe spectrale est une idée courante dans le domaine du traitement de parole. Ainsi, tous les codeurs de la parole traitent ces deux informations séparément. De même, dans des applications de synthèse de parole ou de conversion de voix, les modifications spectrale et prosodique sont en général effectuées indépendamment l'une de l'autre. Des études en synthèse de la parole ont mis en évidence une dépendance entre le pitch et l'enveloppe spectrale. Ainsi cette dépendance est un élément important qui ne peut être négligé dans un système de conversion de voix. Dans le chapitre 4, nous avons étudié cette dépendance afin d'en tenir compte lors de la conversion. Les premières études ont abouti à une fonction de prédiction de pitch à partir de l'enveloppe spectrale. Les résultats obtenus par cette fonction de prédiction sont très satisfaisants dans la mesure où, en moyenne, l'erreur de prédiction est de l'ordre de 4 Hz. L'application de cette technique de prédiction dans le domaine de la conversion de voix se traduit par une conversion en deux étapes : tout d'abord une conversion du timbre est effectuée, puis, une fonction de prédiction de pitch est appliquée aux paramètres spectraux transformés afin d'estimer la valeur de pitch convertie. Cependant, la combinaison de la prédiction du pitch avec la transformation de l'enveloppe spectrale par GMM conduit a un système de conversion de voix qui manque de robustesse. En effet, lorsque l'enveloppe spectrale est bien convertie, la prédiction du pitch donne des résultats satisfaisants. En revanche, toute erreur sur la conversion du timbre se répercute sur la prédiction de pitch. Dans le même objectif de tenir compte de la dépendance entre le pitch et l'enveloppe spectrale, nous avons proposé une deuxième technique de conversion permettant de transformer de manière conjointe l'enveloppe spectrale et la fréquence fondamentale. Dans cette technique, nous utilisons deux fonctions de transformation basées sur le modèle GMM : une pour les

trames voisées qui permet de transformer de manière conjointe l'enveloppe spectrale et le pitch, et une deuxième pour les trames nons voisées qui prend en compte uniquement l'enveloppe spectrale. Les mesures objectives que nous avons effectuées ont montré que cette conversion conjointe conduit à une bonne conversion de pitch et rend la transformation du timbre plus robuste.

Cette technique de conversion conjointe à de plus fait l'objet d'une évaluation subjective suivant un protocole rigoureux. Une douzaine d'auditeurs ont été invités à tester la qualité de la conversion. Durant ces tests, nous avons comparé notre méthode avec une technique de conversion de référence : la conversion par GMM classique [Sty96b, Kai01], puis par rapport à une borne supérieure définie par un plaquage acoustique de la cible sur la source. Les résultats ont montré une préférence nette de la conversion conjointe du timbre et de la fréquence fondamentale par rapport à la conversion séparée classique. La conversion conjointe atteint la qualité du plaquage acoustique dans 51% des cas. Dans les 49% restants, la qualité de la conversion conjointe est légèrement inférieure. Cependant, la comparaison du plaquage acoustique avec la parole naturelle montre que même une modification idéale de l'enveloppe spectrale et de la fréquence fondamentale ne peut atteindre la qualité de la parole naturelle. D'autres paramètres tels ques les phases des harmoniques doivent être prises en compte. Or, la transformation des phases est un problème délicat dans la mesure où les études statistiques que nous avons réalisées n'ont pas permis de dégager une dépendance entre les phases et les autres paramètres du modèle HNM.

Dans le chapitre 5, nous nous sommes penchés sur des aspects pratiques de la conversion de voix. Dans un premier temps, nous nous sommes intéressés à l'intégration de la conversion de voix dans un système de synthèse par HNM. Nous avons proposé une simplification algorithmique de la conversion par GMM permettant une réduction par un facteur variant entre 40 et 120 de son coût de calcul lors de la phase de transformation et facilité ainsi son intégration dans un système de synthèse à partir de texte.

Dans ce même chapitre, nous avons abordé le problème d'acquisition des bases d'apprentissage. En effet, traditionnellement, l'apprentissage d'une fonction de conversion nécessite deux corpus parallèles, c'est-à-dire comprenant le même contenu phonétique. Nous avons alors proposé de créer artificiellement deux corpus parallèles en utilisant

pour le locuteur source un enregistrement obtenu par le biais d'un système de synthèse par corpus. Les tests d'écoute effectués pour comparer la conversion en utilisant des bases d'apprentissage naturelles et des bases synthétisées sont satisfaisants : les auditeurs était incapable de faire la différences entre les deux configurations.

Perspectives

Nous avons mis en oeuvre un système de conversion de voix permettant une bonne conversion du timbre et du pitch. Il convient dans un premiers temps de le tester sur d'autres bases, et notamment d'évaluer sa capacité à effectuer des conversions de type homme-homme et femme-femme. Il convient également de tester cette technique de conversion dans le cas de corpus non parallèles "généralisée", c'est-à-dire en utilisant la composition de deux fonctions de transformation avec des enregistrements synthétisés comme voix de référence (voir section 5.3.2).

Pour améliorer le système proposé, plusieurs voies de recherche restent ouvertes. Parmi celles-ci, un premier axe concerne le modèle d'analyse du signal utilisé. En effet, il conviendrait d'utiliser un modèle permettant un contrôle explicite sur les paramètres liés à la production du signal glottique. En particulier, il serait intéressant de modéliser les interactions existant entre le conduit vocal et le signal glottique et de les prendre en compte pour des fins de conversion. Les méthodes d'analyse synthèse proposé dans [Alk92, DKA95] offrent des voies prometteuses.

Un deuxième axe de recherche se rapporte à la technique de transformation. Pour toutes les méthodes existantes à l'heure actuelle, la conversion est faite trame par trame de manière indépendante. Or, la manière dont se succèdent les trames est un élément potentiellement important dont il faudrait tenir compte lors de la conversion. La dépendance entre trames successives pourrait être modélisée à l'aide de chaînes de Markov cachées qui ont largement montré leur intérêt en reconnaissance vocale. Dans cette optique, les problèmes à résoudre sont de deux ordres : d'une part, définir une fonction de conversion intégrant effectivement cette dépendance temporelle et d'autre part, spécifier un modèle suffisamment simple pour qu'il puisse être correctement appris à partir des données disponibles pour la conversion de voix.

Enfin, dans cette étude, seules des modifications à l'échelle segmentale ont été effectuées. Or le style d'élocution, i.e. la prosodie propre à un individu est également un élément clé de son identité vocale. Il est par conséquent nécessaire de mettre en oeuvre des méthodes visant à restituer le style d'élocution du locuteur cible. A cette fin, il conviendrait de définir un formalisme permettant de modéliser et de modifier la prosodie d'un locuteur, à savoir notamment son style d'élocution et son interaction.

Bibliographie

[Abe92] M. Abe – « A Study on Speaker Individuality Control », Thèse, NTT Human Interface Laboratories, Mars 1992.

[Alk92] P. Alku – « Glottal wave analysis with pitch sunchronous iteratavie adaptive inverse filtring », *Speech Communication* **11** (1992), p. 109–108.

[ANK88] M. Abe, S. Nakamura et H. Kawabara – « Voice conversion through vector quatizantion », *Proc. of ICASSP* (1988), p. 655–658.

[Ars99] L. Arslan – « Speaker transformation algorithm using segmental codebook », *Speech Communication* **28** (1999), no. 3, p. 211–226.

[ASK90] M. Abe, K. Shikano et H. Kawabara – « Cross-language voice conversion », *Proc. of ICASSP* (1990), p. 345–348.

[BS96] G. Baudoin et Y. Stylianou – « On the transformation of the speech spectrum for voice conversion », *Proc. of ICSLP* **2** (1996), p. 1405–1408.

[CCL03] Y. Chen, M. Chu, J. Liu et R. Liu – « Voice Conversion with smoothed GMM and MAP adaptation », *Proc. of Eurospeech*(2003).

[CLM95] O. Cappé, J. Laroche et E. Moulines – « Regularized estimation of cepstrum envelope from discrete frequency points », *IEEE ASSP Workshop on application of signal processing to audio and acoustic* (1995), p. 213–216.

[CVW02] T. Ceyssens, W. Verhelst et P. Wambacq – « On the construction of a pitch conversion system », *Proc. of EUSIPCO* (2002).

[DC89] I. Dologlou et G. Carayannis – « Pitch detection based on zero phase filtering », *Speech Communication* **8** (1989), p. 309–318.

[DKA95] W. Ding, H. Kasuya et S. Adach – « Simultatneous estimation of voice tract and voice source parameters based on an arx model », *IEICE Trans. Inf & Syst, vol. E78-D* (1995), p. 738–743.

[DLR77] A. Dempster, N. Laird et D. Rubin – « Maximum likelihood from incomplete data via the EM algorithm », *Journal of the Royal Statistical Society* **39** (1977), no. Serie B, p. 1–38.

[Dod85] G. Doddington – « Speaker recognation-Identifying people by their voices », *Proc. of IEEE* **73** (1985), no. 11, p. 1651–1664.

[Eme77] F. Emerard – « Synthèse par diphones et traitement de la prosodie », Thèse, Université de Grenoble III, 1977.

[EnRC03a] T. En-najjary, O. Rosec et T. Chonavel – « A new method for pitch prediction and its application in voice conversion », *Proc. of Eurospeech* (2003).

[EnRC03b] T. En-najjary, O. Rosec et T. Chonavel – « Influence de la modélisation spectrale sur les performances d'un système de conversion de voix », *Proc. de GRETSI* (2003).

[EnRC04a] T. En-najjary, O. Rosec et T. Chonavel – « A voice conversion method based on joint pitch and spectral envelope transformation », *Proc. of ICSLP* (2004).

[EnRC04b] T. En-najjary, O. Rosec et T. Chonavel – « Fast GMM-based voice conversion for Text-to-Speech synthesis systems », *Proc. of ICSLP* (2004).

[Fan66] G. Fant – « A note on vocal tract size factors and noninuform F-Pattern scalings », *Speech Transmission Laboratory Quaterly Pregress and Status Report 4/66, Royal Institute of Technology,Sweden* (1966), p. 22–30.

[Fan75] G. Fant – « Non-uniform vowel normalization », *Speech Transmission Laboratory Quaterly Pregress and Status Report 2-3/75, Royal Institute of Technology,Sweden* (1975), p. 1–19.

[FLL85] G. Fant, , J. Lijencrant et Q. Lin – « A four parameter model og glottal flow », *STL-QPSR* (1985).

[Fur86] S. Furui – « Research on individuality features in speech waves and automatic speaker recognition », *Speech Communication* **5** (1986), no. 2, p. 183–197.

[GBGM80] R. Gary, A. Buzo, A. Gray et Y. Matsuyama – « Distortion Measures for speech processing », *IEEE Trans. Acoustic. Speech, Signal Processing* **28** (1980), no. 4, p. 367–376.

[GH87] Y. Gong et J. Haton – « Time domain harmonic matching pitch estimation using time-depndent speech modeling », *IEEE Trans. Acoust., Speech, Signal processing* **35(10)** (1987), p. 1386–1400.

[GJ94] Z. Ghahramani et M. Jordan – « Solving inverse problem using EM approach to density estimation », *In Proccedings of the 1993 Connectionist Models Summer School, Lawrence Erlbaum Publishers* (1994), p. 316–323.

[GK03] B. Gilet et S. King – « Transforming F0 Contours », *Proc. of Eurospeech* (2003).

[GL88] D. Griffin et J. Lim – « Multiband excitation vocoder », *In Proc. IEEE Trans. Acoust., Speech, Signal Processing* **vol.36** (1988), p. 1223–1235.

[GM76] A. Gary et J. Markel – « Distortion Measures for speech processing », *IEEE Trans. Acoustic. Speech, Signal Processing* **ASSP-24** (1976), no. 5, p. 380–391.

[Gol75] U. Goldstein – « Speaker-identifying features based on formant tracks », *Journal of the Acoustical Society of America* **59** (1975), no. 1, p. 176–182.

[GR90] T. Galas et X. Rodet – « An improved cepstral methode for deconvolution of source-filter systems with discret spectra : application to musical sound signals », *In ICMC* (1990).

[GR91] T. Galas et X. Rodet – « Generalized functional approximation for source-filter system modeling », *Proc. of Eurospeech* (1991), p. 1085–1088.

[GS94] H. Gish et M. Schmidt – « Text-independent speaker identification », *IEEE Signal Processing Magazine* **11** (1994), no. 4, p. 18–32.

[Har78] F. Harris – « On the use of windows for harmonic analysis with discrete fourier transform », *Proc. of IEEE* **66** (1978), p. 51–83.

[HD76] D. Hartman et J. Danhaur – « Perceptual features for males in four perceived age decades », *J. Acoust. Soc. Amer* **59** (1976), p. 713–715.

[Hen01] N. Henrich – « Etude de la source glottique en voix parlée et chantée », Thèse, Université de Paris6, Novembre 2001.

[Her88] D. Hermes – « Measurement of pitch by subharmonic summation », *J. Acoust. Soc. Amer.* **83(1)** (1988), p. 257–264.

[Her91] D. Hermes – « Synthesis of breathy vowels : some research methods », *Speech Communication* **11** (1991), p. 497–502.

[Hes83] W. Hess – *Pitch determination of speech signal : Algorithmes and devices*, Springer, Berlin, 1983.

[Hol90] H. Hollien – *The acoustic of crime-The new science of forensic phonetics*, Plenum Press, 1990.

[IS88] K. Itoh et S. Saito – « Effects of acoustical features parameters on perceptual speaker identity », *Review of the Electrical Communications Laboratory* **36** (1988), no. 1, p. 135–141.

[IS94] N. Iwahashi et Y. Sagisaka – « Speech spectrum transformation by speaker interpolation », *IEEE ICASSP* (1994).

[Ita75] F. Itakura – « Line spectrum representation of linear predictive coefficients of speech signals », *Journal of the Acoustical Society of America* **57** (1975), p. S35.

[Kai01] A. Kain – « High resolution voice transformation », Thèse, Oregon Health and Science University, Octobre 2001.

[Kam96] N. Kambhatla – « Local models and Gaussian Mixtures Models for Statistical Data Processing », Thèse, Oregon Graduate Institute of Science and Technology, Janvier 1996.

[Kay88] S. Kay – *Modern spectral estimation*, Prentice Hall, Englewood cliffs, New Jersey, 1988.

[KK90] D. Klatt et L. Klatt – « Analysis, synthesis, and perception of voice quality variations among female and male talkers », *J. Acoust. Soc. Amer.* **87** (1990), no. 2, p. 820–857.

[KM98a] A. Kain et M. Macon – « Spectral voice conversion for text to speech synthesis », *Proc. of ICASSP* **1** (1998), p. 285–288.

[KM98b] A. Kain et M. Macon – « Text-to-speech voice adaptation from sparse training data », *Proc. of ICSLP* (1998).

[KS95] H. Kawabara et Y. Sagisaka – « Acoustic characteristics of speaker individuality : Control and conversion », *Speech Communication* **16** (1995), no. 2, p. 165–173.

[KS00] A. Kain et Y. Stylianou – « Stochastic modeling of spectral adjustement for high quality pitch modification », *Proc. of ICASSP* (2000).

[Lan85] D. V. Lancker – « Familiar voice recognition : patterns and parameters. part1 : Recognition of backward voices », *Journal of phonetics* **Vol. 1** (1985), p. 19–38.

[LB78] N. Lass et W. Brown – « Correlation study of speaker's height, body surface areas and speaking fundamental frequencies », *J. Acoust. Soc. Amer* **63** (1978), p. 1218–1220.

[Lob01] A. Lobo – « Glottal flow derivative modeling with wavelet smoothed excitation », *Proc. of ICASSP* (2001).

[LYC96] K. Lee, D. Young et I. Cha – « A new voice transformation method based on both linear and nonlinear prediction analysis », *Proc. of ICSLP* **3** (1996), p. 1401–1404.

[MA95] H. Mizuno et M. Abe – « Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt », *Speech Communication* **16(2)** (1995), p. 153–164.

[Mac96] M. Macon – « Speech synthesis based on sinusoidal modeling », Thèse, Georgia Institute of Technology, Atlanta, Georgia, Octobre 1996.

[MHSN73] H. Matsumuto, S. Hiki, T. Sone et T. Nimura – « Multidimentional representation of personal quality of vowels and its acoustical correlates », *IEE Trans. AU* **AU-21** (1973), p. 428–436.

[MHT+01] C. Miyajima, Y. Hattori, K. Tokuda, T. Masuko, T. Kobayash et T. Kitamura – « Text-independent speaker identification using gaussian

mixture models based on multi-space probability distribution », *IEICE Transactions on Information and Systems* **E** (2001).

[MQ91] R. McAulay et T. Quatieri – « Low-rate speech coding based on the sinusoidal model », Advances in Speech Signal Processing, Marcel Dekker (1991), p. 165–208.

[MQ95] R. McAulay et T. Quatieri – « Speech coding and synthesis », Sinusoidal coding, elsevier science (1995), p. 121–173.

[MQ85] R. MacAulay et T. Quatieri – « Pitch estimation and voicing detection based on a sinusoidal model », *IEEE Trans. Acoust., Speech, Signal processing* (1985), p. 249–252.

[MSM04] A. Mouchtaris, J. V. Spiegal et P. Mueller – « Non parallel training for voice conversion by maximum likelihoud constrained adaptation », *Proc. of ICASSP* **1** (2004), no. 1-4.

[MW79] H. Matsumoto et H. Wakita – « Frequency warping for nonuniform talker normalization », *Proc. of ICASSP* (1979), p. 566–569.

[MW86] H. Matsumoto et H. Wakita – « Vowel normalization by frequency warped spectral matching », *Speech Communication* **5** (1986), p. 239–251.

[MYC91] Y. Medan, E. Yair et D. Chazan – « Super resolution pitch determination of soeech Signals », *IEEE Trans. Acoust., Speech, Signal processing* **39(1)** (1991), p. 40–48.

[Nef04] S. Nefti – « Segmentation automatique de parole en phone :Correction d'étiquitage par l'introduction de mesures de confiances », Thèse, Université de Rennes 1, Décembre 2004.

[NL75] P. Nordstrom et B. Lindblom – « A normalization procedure for vowel formant data », *International Congress of Phonetic Sciences* (1975).

[NMRY95a] M. Narendranath, H. Murthy, S. Rajendran et B. Yegnanarayan – « Transformation of formants for voice conversion using artificiel neural networks », *Speech Communication* **16 , 2** (1995), p. 207–216.

[NMRY95b] M. Narendranath, H. Murthy, S. Rajendran et B. Yegnanarayana – « Transformation of formants for voice conversion using artificial neural networks », *Speech Communication* **16** (1995), p. 207–216.

[NS89] S. Nakamura et K. Shikano – « Spectrogram normalization using fuzzy vector quantization », *J. Acoust. Soc. Japan* **45** (1989), p. 107–114.

[Oud98] M. Oudot – « Etude du modèle Sinusoïdes et Bruit pour le traitement des signaux de parole, Estimation robuste de l'enveloppe spectrale », Thèse, Thèse de doctorat de l'ENST paris, France, Novembre 1998.

[PA93] K. Paliwal et B. Atal – « Efficient vector quantization of LPC parameters at 24 bits/frames », *IEEE Trans. on acoustics Speech and Audio Proc.* **1(1)** (1993), p. 3–14.

[Pal95] K. Paliwal – « Interpolation properties of lineair prediction parametric representation », *Proc. of Eurospeech* (1995), p. 1029–1032.

[Pat00] D. Patterson – « A Linguistic approach to pitch range modeling », Thèse, University of Edinburgh, 2000.

[PX03] Pythoud et Xanthos – « Cours de phonétique », Disponible à : http ://www.unil.ch/ling/phon/index.html, 2003.

[RCRM76] L. Rabiner, M. Cheng, A. Rosenberg et C. McGonegal – « A comparative performance study of several pitch detection algorithms », *IEEE Trans. Acoust., Speech, Signal processing* **24** (1976), p. 399–418.

[RD00] D. Reynolds et T. Q. R. Dunn – « Speaker verification using adapted gaussian mixture models », *Digital Signal Processing* **10** (2000), p. 19–41.

[RJ93] L. Rabiner et B.-H. Juang – *Fundamentals of speech recognation*, PTR Prentice Hall, 1993.

[RSa] L. Rabiner et R. Scahfer – *Digital Processing of speech signal*, Signal Processing Series, Prentice Hall.

[RSb] A. E. Rosenberg et F. K. Soong – « Recent research in automatic speaker recognation. ».

[Sao90] S. Saoudi – « Codage de la parole par les paires de raies spectrales », Thèse, Université de Rennes I, Décembre 1990.

[SCM95] Y. Stylianou, O. Cappé et E. Mouline – « Statistical methods for voice quality transformation », *Proc. of Eurospeech* **6** (1995), no. 2, p. 447–450.

[SCM98] Y. Stylianou, O. Cappé et E. Mouline – « Continous probabilistic transform for voice conversion », *IEEE Transaction on Speech and Audio Processing* **6** (1998), no. 2, p. 131–142.

[SI86] N. Sugamura et F. Itakura – « Speech analysis and synthesis methods developed at ECL in NTT from LPC to LSP », *Speech Communication* **5(2)** (1986), p. 199–215.

[SM04] X. Shao et B. Milner – « Pitch prediction from MFCC vectors for speech reconstruction », *Proc. of ICASSP* (2004), no. 841-844.

[SNB96] A. Schmidt-Neilsen et D. Brock – « Speaker recognizability testing for voice coders », *Proc. of ICASSP* **2** (1996), no. 1249-1152.

[SR68] M. Schwartz et H. Rine – « Identification of speaker sex from isolated,whispered vowels », *J. Acoust. Soc. Amer* **44** (1968), p. 1736–1737.

[SR79] T. Sreenivas et P. Rao – « Pitch extraction from corrupted harmonics of the power spectrum », *J. Acoust. Soc. Amer.* **65(1)** (1979), p. 223–228.

[SR03] A. Sorin et T. Ramabdran – « Extended advanced front-end (XAFE) algorithm aescription », Tech. report, ETSI STQ-Aurora DSR Working Group, 2003.

[SS95] A. Syrdal et S. Steele – « Vowel F1 as a function of speaker fundamental frequency », 110^{th} *Meeting of JASA* **78** (1995).

[Sty96a] Y. Stylianou – « Decomposition of speech signal into a deterministic and a stochastic part », *Proc. of ICSLP* (1996).

[Sty96b] Y. Stylianou – « Harmonic plus noise models for speech, combined with statistical methods for speech and speaker modifications », Thèse, Telcom Paris, Janvier 1996.

[TA97] K. Tanaka et M. Abe – « A new fundamental frequency modification algorithm with transformation of spectrum envelope according to F0 », *Proc. of ICASSP* **2** (1997).

[TSS04] T. Toda, H. Saruwatari et K. Shikano – « Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum », *Proc. of ICASSP* (2004), no. I-97-I-100.

[UAE93] M. Unser, A. Adroubi et Eden – « B-spline signal processing », *IEEE Transactions on Speech and Audio Processing* **41(2)** (1993), p. 821–833.

[Val92] H. Valbret – « Système de conversion de voix pour la synthèse de la parole », Thèse, ENST Paris, septembre 1992.

[VMT92] H. Valbret, E. Mouline et J. Tubach – « Voice transformation using PSOLA technique », *Speech Communication* **11** (1992), p. 175–187.

[YS98] P. Yang et Y. Stylianou – « Real time voice alteration based on linear prediction », *Proc. of ICSLP* (1998), p. 1667–1670.

[YTM+03] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda et T. Kobayashi – « A context clustering technique for average voice models », *IEICE Trans. Inf & Syst, vol. E86-D* (2003), p. 534–542.

[ZT80] E. Zwicker et E. Terhardt – « Analytical expression for critical-band rate and critical bandwidth as a function of frequency », *J. Acoust. Soc. Amer.* **68** (1980), no. 5, p. 1523–1525.