Conversion de la voix : Approches et applications

Imen Ben Othmane

Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique

Université de Carthage

Ecole Nationale d'Ingénieurs de Carthage

وزارة التعليم العالي و البحث العلمي
جامعة قرطاج
المدرسة الوطنية للمهندسين بقرطاج

THESE

en vue de l'obtention du

DOCTORAT DE L'ECOLE NATIONALE D'INGENIEURS DE CARTHAGE

Spécialité : Génie Electrique

présentée par

IMEN BEN OTHMANE

intitulée

CONVERSION DE LA VOIX : APPROCHES ET APPLICATIONS

Soutenue le 4 juillet 2019 à l'ENICarthage devant le jury d'examen composé de :

M. Adnen CHERIF	Président	Professeur, Faculté des Sciences de Tunis
Mme Sofia BEN JEBARA	Rapporteur	Professeur, Sup'Com
M. Mondher FRIKHA	Rapporteur	Maître de Conférences, ENET'Com
Mme Afef ELLOUMI	Examinateur	Maître de Conférences, ENICarthage
M. Kais OUNI	Directeur de thèse	Professeur, ENICarthage
M. Joseph DI MARTINO	Co-Directeur de thèse	Maître de Conférences, Université de Lorraine

Thèse préparée au Laboratoire de Recherche Electricité Intelligente et TIC, EI&TIC, LR18ES44, ENICarthage

Dédicaces

Ce travail est dédié tout d'abord à mes parents. À ma très chère mère, symbole d'amour, de tendresse, et d'affection. Grâce au sens du devoir et aux sacrifices immenses qu'elle a consentis, je suis parvenu à terminer ce travail.

À mon très cher père, qui est pour moi un symbole de responsabilité, d'optimisme et de confiance en soi face aux difficultés de la vie. Grâce au sens du devoir et aux sacrifices immenses qu'il a consentis il a su m'offrir un avenir.

C'est à vous deux que je dois cette réussite et je suis fière de vous la faire partager.

À mon mari Houssem, qui a su partager ma joie lorsqu'un résultat s'est avéré correct et qui a su me réconforter dans le cas contraire. J'aimerais bien que tu trouves dans ce travail l'expression de mes sentiments de reconnaissance les plus sincères car grâce à ta patience et à ton aide la finalisation de ce travail a pu voir le jour.

À mon fils Jed : je te remercie d'avoir été gentil et patient durant mes journées d'études. Je te souhaite tout le bonheur du monde.

À mes beaux-parents : Jouda et Mohammed Habib pour les encouragements et leur soutien. Puisse Dieu, vous garder, vous préserver du mal, vous combler de bonheurs et vous procurer une longue vie.

À ma meilleure amie Asma, pour son aide, son encouragement et son soutien durant ma thèse.

Enfin je dédie cette thèse à toute ma famille que j'aime infiniment.

Remerciements

Ce travail a été réalisé à l'Ecole Nationale d'Ingénieurs de Carthage, ENICarthage, au sein du Laboratoire de Recherche Electricité Intelligente et Technologies de l'Information et des Communications, EI&TIC, sous la direction de Professeur **Kaïs OUNI**.

Je voudrais remercier Monsieur **Adnen CHERIF**, Professeur à la Faculté des Sciences de Tunis, FST, pour l'honneur qu'il m'a fait en acceptant de présider le jury d'examen de cette thèse.

J'adresse mes remerciements les plus sincères à Madame **Sofia BEN JEBARA**, Professeur à l'Ecole Supérieure des Communications, Sup'Com, et Monsieur **Mondher FRIKHA**, Maître de Conférences, à l'Ecole Nationale d'Electronique et des Télécommunications, ENET'Com, pour l'intérêt qu'ils ont porté à ce travail en acceptant d'être rapporteurs des travaux de ma thèse et pour leur participation à mon jury.

Que Madame **Afef ELLOUMI**, Maître de Conférences à l'Ecole Nationale d'Ingénieurs de Carthage, ENICarthage, trouve ici l'expression de ma reconnaissance et de mes remerciements pour avoir accepté d'examiner ce travail et de participer à mon jury.

Je tiens à exprimer ma profonde reconnaissance à Monsieur **Kaïs OUNI**, Professeur à l'Ecole Nationale d'Ingénieurs de Carthage, ENICarthage, et Directeur du Laboratoire de Recherche EI&TIC. Je voudrais le remercier sincèrement pour avoir cru en mes capacités et pour m'avoir fournie d'excellentes conditions de travail au sein du Laboratoire de Recherche EI&TIC et pour sa disponibilité, ses remarques pertinentes et les conseils qu'il m'a prodigués pour faire aboutir mes travaux de thèse.

Je tiens à exprimer aussi toute ma gratitude à Monsieur **Joseph DI MARTINO**, Maître de Conférences à l'Université de Lorraine et chercheur au sein de l'Equipe SMarT du Laboratoire Lorrain de Recherche en Informatique et ses Applications, LORIA, de m'avoir accueillie durant mes stages au LORIA et de m'avoir accordée de son temps et de son énergie. Les conseils et idées qu'il m'a prodigués ont toujours été pour moi très enrichissants et clairs, me facilitant largement la tâche et me permettant d'aboutir à la finalisation de cette thèse.

Enfin, je réserve une pensée particulière à tous les membres du Laboratoire de Recherche EI&TIC pour l'ambiance conviviale et l'esprit d'équipe ainsi qu'à ceux qui m'ont motivée, soutenue et encouragée.

Résumé

La conversion vocale est un problème important dans le domaine du traitement du signal audio. Le but de la conversion de voix est de transformer le signal de parole d'un locuteur source de telle sorte qu'il soit perçu comme s'il avait été prononcé par un locuteur cible tout en conservant le contenu linguistique du signal converti identique à celui du signal d'origine. La conversion basée sur un modèle de mélange gaussien (GMM) est la technique la plus couramment utilisée dans le domaine de la conversion vocale, mais elle est souvent sensible aux problèmes de sur-apprentissage et de lissage excessif. Pour résoudre ces problèmes, nous proposons une classification secondaire en appliquant une classification, par la technique des K-moyennes, dans chaque classe obtenue par une classification primaire afin d'obtenir des fonctions de conversion locales plus précises. Cette proposition évite le recours à des algorithmes d'apprentissage complexes car les fonctions de transformation locales sont déterminées en même temps.

La deuxième contribution de cette thèse inclut une nouvelle méthodologie pour concevoir la relation entre deux ensembles d'enveloppes spectrales. Nos systèmes fonctionnent : 1) en cascadant des réseaux de neurones profonds avec un modèle de mélange gaussien pour construire des modèles DNN-GMM et GMM-DNN-GMM, ceci afin de trouver une fonction de transformation performante entre les vecteurs cepstraux des deux locuteurs ; 2) en utilisant un nouveau processus de synthèse spectrale mettant en œuvre des prédicteurs de cepstres en cascade avec une excitation et une phase extraites de l'espace d'apprentissage cible codé sous la forme d'un arbre binaire KD-tree.

Les résultats expérimentaux des méthodes proposées exhibent une nette amélioration de l'intelligibilité, de la qualité et du naturel des signaux de parole convertis par rapport aux résultats obtenus avec une méthode de conversion de base. L'extraction de l'excitation et de la phase de l'espace d'apprentissage cible permet de préserver l'identité du locuteur cible.

Notre dernière contribution de cette thèse concerne l'implémentation d'un nouveau système d'aide à la parole pour améliorer la parole œsophagienne (ES). La méthode adoptée dans cette thèse vise à améliorer la qualité de la voix œsophagienne en combinant une technique de conversion vocale et un algorithme de dilatation temporelle. Dans le système proposé, un réseau de neurones profonds (DNN) est utilisé pour transformer de manière non linéaire les vecteurs cepstraux relatifs au conduit vocal. Ensuite, les trames converties obtenues sont utilisées pour déterminer les vecteurs d'excitation

et de phase réalistes à partir de l'espace d'apprentissage cible préalablement codé sous la forme d'un arbre binaire. Nous montrons que la méthode proposée améliore considérablement l'intelligibilité et le naturel de la voix œsophagienne convertie.

Mots clés Conversion vocale, modèles de mélange gaussien, classification, fonctions de transformation locale, réseaux de neurones profonds, cepstre, arbre KD-tree, prédicteurs de cepstres en cascade, espace d'apprentissage, excitation, phase, algorithme de dilatation temporelle.

Abstract

Voice conversion (VC) is an important problem in the field of audio signal processing. The goal of voice conversion is to transform the speech signal of a source speaker such that it sounds as if it had been uttered by a target speaker while preserving the same linguistic content of the original signal. Gaussian mixture model (GMM) based conversion is the most commonly used technique in VC, but is often sensitive to overfitting and oversmoothing. To address these issues, we propose a secondary classification by applying a K-means classification in each class obtained by a primary classification in order to obtain more precise local conversion functions. This proposal avoids the need for complex training algorithms because the estimated local mapping functions are determined at the same time.

The second contribution of this thesis, includes a new methodology for designing the relationship between two sets of spectral envelopes. Our systems perform by : 1) cascading Deep Neural Networks with Gaussian Mixture Models for constructing DNN-GMM and GMM-DNN-GMM models in order to find an efficient global mapping relationship between the cepstral vectors of the two speakers ; 2) using a new spectral synthesis process with excitation and phase extracted from the target training space encoded as a KD-tree.

Experimental results of the proposed methods exhibit a great improvement in intelligibility, quality and naturalness of the converted speech signals when compared with those obtained by a baseline conversion method. The extraction of excitation and phase from the target training space, allows the preservation of target speaker's identity.

Our last contribution of this thesis concerns the implementation of a novel speaking-aid system for enhancing esophageal speech (ES). The method adopted in this thesis aims to improve the quality of esophageal speech using a combination of a voice conversion technique and a time dilation algorithm. In the proposed system, a Deep Neural Network (DNN) is used as a nonlinear mapping function for vocal tract vectors conversion. Then the converted frames are used to determine realistic excitation and phase vectors from the target training space using a frame selection algorithm. We demonstrate that that our proposed method provides considerable improvement in intelligibility and naturalness of the converted esophageal stimuli.

Keywords Voice conversion, Gaussian Mixture Model, K-means classification, local mapping functions, deep neural network, cepstrum, KD-tree, cascaded cepstrum predictors, training space, excitation, phase, time dilation algorithm.

Table des matières

Liste des abréviations et notations

AE Auto-Encodeur

ALT Artificial Larynx Transducer

ANN Artificial Neural Networks

CD Cepstral Distance

DBN Deep Belief Networks

DFW Dynamic Frequency Warping

DKPL Dynamic Kernel Partial Least Squares

DNN Deep Neural Network

DTW Dynamic Time Warping

EL Electro-Larynx

EM Expectation-Maximisation

ES Esophageal Speech

FD-PSOLA Frequency-domain Pitch-Synchronous Overlap-Add

FFT Fast Fourier Transform

GBRBM Gaussian-Bernoulli Restricted Boltzmann Machine

GLA Griffin-Lim Algorithm

GMM Gaussian Mixture Model

GMM-GMF Gaussian Mixture Model Global Mapping Function

GMM-LMF Gaussian Mixture Model Local Mapping Function

GPU Graphics Processing Unit

GSRTISI-LA Gnann et Spiertz Real Time Iterative Spectrum Inversion with Look-Ahead

HMM Hidden Markov Model

HNM Harmonic plus Noise Model

IFFT Inverse Fast Fourier Transform

JD-GMM Joint Density Gaussian Mixture Model

LMR Linear Multiple Regression

LP-PSOLA Linear-Prediction Pitch-Synchronous Overlap-Add

LPC Linear Predictive Coding

LPCC Linear Predictive Coding Coefficients

LSF Line Spectral Frequency

LTI Linear Time Invariant

MGC Mel-generalized cepstral

ML Maximum Likelihood

MOS Mean Opinion Score

MSE Mean Squared Error

MVR Multi-Variables Regression

NN Neural Network

OLA Overlap-add

PLS Partial Least Squares

PSOLA Pitch-Synchronous Overlap-Add

QV Quantification Vectorielle

RBM Restricted Boltzmann Machine

RTISI Real-Time Iterative Spectrogram Inversion

RTISI-LA Real-Time Iterative Spectrogram Inversion with Look-Ahead

SER Signal-to-Error Ratio

SS Spectral Subtraction

STRAIGHT Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum

TD-PSOLA Time-Domain Pitch-Synchronous Overlap-Add

TE Tracheoesophageal Speech

TFCT Transformée de Fourier à Court Terme

TTS Text-to-Speech

VC Voice Conversion

VCC Voice Conversion Challenge

VQ Vector Quantization

Liste des figures

Liste des tableaux

Introduction

La parole est l'un des moyens les plus importants de communication entre humains. C'est un moyen efficace pour nous exprimer et pour partager rapidement des informations, que ce soit dans une conversation en face à face, par téléphone, à l'écoute de la radio ou par tout autre moyen. La parole en tant que forme de communication n'est pas seulement un moyen pour transmettre un message, mais est aussi un moyen pour transmettre des indications sur l'humeur du locuteur, son sexe, son âge, etc... Elle contient ainsi une combinaison de ces indications qui nous permet d'obtenir l'identité de la personne qui énonce le discours.

Le terme Transformation Vocale désigne les différentes modifications que subit le son produit par un locuteur source (Stylianou, 2009). Les modifications que nous appliquons visent à modifier le signal vocal de manière à conserver son contenu linguistique tout en modifiant les autres propriétés du signal. Ces changements permettent de transformer par exemple la voix d'une personne jeune en celle d'un personne âgée, la voix d'un homme en celle d'une femme ou même peuvent permettre la modification de l'émotion perçue (Kawanami et al., 2003).

Dans cette thèse, l'accent est mis sur un sous-ensemble de la transformation vocale, à savoir celui de la conversion vocale (VC). La conversion de voix est une technique de transformation d'un signal de parole d'un locuteur source, de manière à ce qu'il semble, à l'écoute, être prononcé par un locuteur cible (Mohammadi et Kain, 2017).

Au cours de ces dernières décennies, divers travaux de recherche en conversion vocale ont été proposés, cependant, nous n'avons toujours pas atteint un système idéal de conversion de la voix.

La construction des corpus de données, variés et volumineux, qui nécessite des ressources humaines et techniques importantes, est le problème majeur auquel se heurte tout système de synthèse vocale. C'est dans ce contexte que la conversion de la voix apparaît comme une solution permettant de résoudre le problème de préparation de nouveaux corpus de parole avec la transformation de corpus de référence vers une nouvelle identité vocale. La conversion vocale est donc une solution simple, flexible et efficace qui permet l'obtention d'une grande diversité d'identités vocales pour les systèmes de synthèse de la parole.

Les méthodes de conversion vocale ont progressé rapidement au cours de la dernière décennie. Ces technologies peuvent s'avérer très utiles dans différents champs

applicatifs comme la synthèse Text-to-Speech (TTS) (Kain, 2001), la synthèse des émotions (Kawanami et al., 2003), les systèmes d'aide à la communication (Nakamura et al., 2006) ou encore l'aide médicale dans le cadre de l'amélioration de la voix œsophagienne (Doi et al., 2014).

Les systèmes de conversion vocale permettent aussi de contrôler les systèmes de reconnaissance automatique du locuteur.

Dans la littérature, plusieurs techniques de conversion des voix ont été développées, parmi lesquelles nous citons : par régression linéaire multiple (Valbret, 1993) ; par quantification vectorielle (Abe et al., 1990) ; par déformation fréquentielle dynamique DFW (Dynamic Frequency Warping) (Valbret et al., 1992) ; par les réseaux de neurones artificiels (ANN ou Artificial Neural Networks) (Narendranath et al., 1995).

La technologie de conversion vocale ne concerne pas seulement les voix laryngées, mais aussi les voix œsophagiennes et on parle alors d'un système d'amélioration de la voix œsophagienne. Par exemple, dans (Doi et al., 2014; Bi et Qi, 1997; Ben Othmane et al., 2017b, 2018b) et (Tanaka et al., 2014), des systèmes d'amélioration de la voix alaryngée ont été développés dans le but de transformer la voix d'un locuteur source (alaryngée) en celle d'un locuteur cible (laryngée).

Ces techniques sont basées sur l'utilisation d'un algorithme de conversion suivi par une étape de re-synthèse vocale pour générer la voix convertie. Ces approches offrent des résultats acceptables mais encore insuffisants pour la correction de la voix œsophagienne car il est très difficile de compenser les différences acoustiques entre la voix alaryngée et la voix laryngée. Ceci est dû à la difficulté de générer des signaux excitatifs réalistes (similaires à ceux produits par les vibrations des cordes vocales).

Depuis une quinzaine d'années, l'étude de la parole alaryngée n'intéresse pas seulement la recherche clinique mais aussi les laboratoires de recherche en traitement automatique de la parole et du signal. L'objectif de la correction de la voix œsophagienne est de rehausser celle-ci dans le but de faciliter la communication d'une personne laryngectomisée.

Cette thèse vise à relever ce défi. C'est pour cette raison que nous nous proposons d'appliquer une nouvelle technique de conversion vocale en tenant compte de la particularité de l'appareil vocal des locuteurs alaryngés.

Contributions

Le sujet principal de cette thèse porte sur la conversion de la voix. Notre première contribution concerne une nouvelle méthodologie pour l'estimation de la fonction de transformation entre deux ensembles d'enveloppes spectrales. L'approche proposée est fondée sur une mise en cascade d'un réseau de neurones profonds DNN et d'un modèle de mélange gaussien GMM pour construire des modèles DNN-GMM et GMM-DNN-GMM afin d'obtenir une fonction de transformation performante entre les paramètres acoustiques de la voix source et ceux de la voix cible. La deuxième contribution est

relative à la prédiction de l'excitation cepstrale et des coefficients de phase à partir de l'espace d'apprentissage cible préalablement codé sous la forme d'un arbre binaire.

En second lieu, nous avons accordé une attention particulière à l'applicabilité des approches proposées aux voix alaryngées (les voix œsophagiennes). Nous proposons dans cette contribution d'améliorer la qualité et l'intelligibilité de la voix œsophagienne à l'aide de notre technique de conversion proposée en tenant compte de la particularité du système phonatoire d'un locuteur œsophagien. Le système de conversion de la voix laryngée a été ensuite adapté à la voix œsophagienne permettant ainsi d'élaborer une technique performante de correction de ce type de voix.

Notre troisième contribution réside dans la réalisation d'un système de rehaussement des voix œsophagiennes à l'aide d'un algorithme de dilatation cepstrale pour la correction des distorsions présentes dans les vecteurs cepstraux de la voix œsophagienne.

Organisation de la thèse

La présente thèse est composée de trois chapitres. Le premier chapitre présente des généralités sur le signal de la parole.

Nous y présentons le mécanisme physiologique de production de la voix laryngée ainsi que quelques descripteurs de voix de base et des méthodes de modélisation et de synthèse de la parole.

Le deuxième chapitre présente les principes d'un système de conversion de voix, ainsi qu'une description des méthodes et des techniques de conversion présentées dans la littérature. Nous détaillons dans ce chapitre la mise en œuvre de nos systèmes de conversion de la voix laryngée. Nous évaluons les différents systèmes à l'aide des tests objectifs et subjectifs.

Dans le troisième chapitre, nous nous focalisons sur les techniques de correction de la voix œsophagienne. Nous présentons une brève description anatomique et physiologique de l'appareil vocal d'un patient laryngectomisé ainsi que les différents types de voix alaryngées et les causes des distorsions de ce type de signaux de parole. Nous détaillons ensuite la mise en œuvre de notre système de rehaussement qui est capable d'améliorer la qualité et l'intelligibilité de la parole œsophagienne.

Pour terminer, nous dressons une conclusion du travail effectué, et présentons quelques perspectives et travaux futurs.

Chapitre 1

Anatomie du système phonatoire : Mécanismes de production de la parole

Sommaire

1.1 Introduction

La parole est le mode le plus important de la communication humaine. Elle est étudiée par le champ très vaste du traitement du signal. La production de la parole de l'être humain est un mécanisme très complexe. Les deux caractéristiques fondamentales de la parole sont les caractéristiques articulatoires et prosodiques :

— Une caractéristique articulatoire apparaît physiquement, dans la parole, comme une variation de la pression de l'air causée et émise par le système articulatoire.

— Une caractéristique prosodique est relative au rythme et à la mélodie de la parole. La caractéristique prosodique est formée par trois paramètres acoustiques qui dépendent du signal vocal et qui sont : la fréquence fondamentale F0, l'énergie et le spectre. Ces trois paramètres acoustiques sont liés à leurs grandeurs perceptuelles : le pitch, l'intensité et le timbre.

Afin de mieux comprendre les phénomènes qui sous-tendent les techniques de paramétrisation utilisées pour les systèmes de conversion de la voix, nous commençons par présenter le processus de production de la parole et nous étudierons les mécanismes mis en jeu lors de la phonation.

1.1.1 Le mécanisme de la phonation

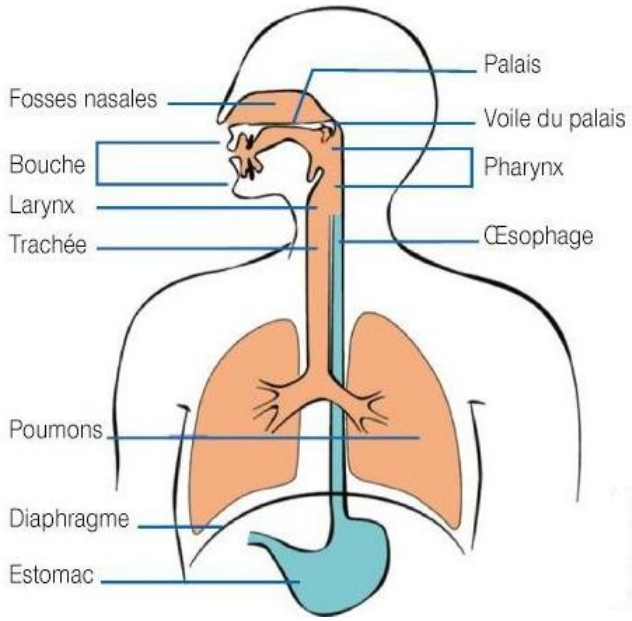


FIGURE 1.1 – *Anatomie de la partie supérieur de l'appareil phonatoire (DriX, 2004)*

Une vue schématique de l'appareil phonatoire est proposée dans la figure 1.1. L'appareil phonatoire comprend classiquement trois composantes : l'appareil respiratoire, les cordes vocales et la cavité bucco-pharyngale.

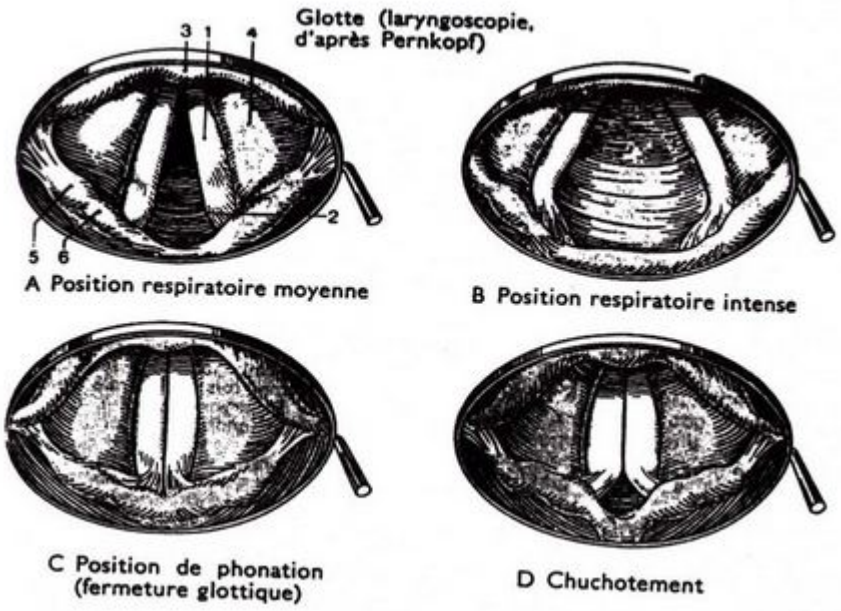


FIGURE 1.2 – *Coupe latérale du larynx, montrant quatre positions différentes des cordes vocales : a) glotte ouverte ; b) glotte largement ouverte ; c) glotte fermée, d) glotte entrouverte, (Pernkopf, 1952)*

— L'appareil respiratoire agit comme un soufflet qui fournit de l'énergie sous forme d'un souffle d'air.
— Les cordes vocales fonctionnent comme un générateur de sons ; pour la production de la parole, les lèvres, la langue et les dents interviennent afin de transformer le souffle d'air en énergie sonore.
— La cavité bucco-pharyngale est formée par le pharynx et la bouche, et permet d'amplifier l'énergie sonore transmise par les cordes vocales.

1.1.2 Le larynx

Le larynx est le conduit musculo-cartilagineux situé dans la gorge, à la partie médiane entre le pharynx et la trachée. Cet organe joue un double rôle : il assure le passage de l'air vers les poumons lors de l'inspiration dans le sens cavité buccale → trachée → poumons (Le Huche et Allali, 2001), ou vers l'extérieur lors de l'expiration.

Il contient aussi une membrane qui assure la protection des voies aériennes lors du passage des aliments du pharynx vers l'œsophage, évitant ainsi l'arrivée de la nourriture dans les voies aériennes.

Le larynx joue aussi un rôle majeur dans la phonation. Il contient les cordes vocales qui permettent l'émission de sons produits lors de la phonation. L'espace entre elles est appelé glotte.

Lorsque la glotte est ouverte, elle permet la respiration ; elle produit le chuchotement lorsqu'elle est entrouverte ; cependant lorsqu'elle est complètement fermée, elle produit la phonation (Figure 1.2). Au cours de la respiration, les cordes vocales sont écartées ce qui permet le passage de l'air. Pendant la phonation, les cordes vocales s'ouvrent et se ferment rapidement. Lorsqu'elles sont légèrement accolées (ce qui permet de libérer une partie du souffle d'air) elles produisent les sons chuchotés.

Fonctionnement du larynx

Le larynx est l'organe indispensable à la formation de la voix. Il fonctionne comme un robinet qui permet de bloquer l'accès à la trachée, en la connectant au pharynx, et par conséquent permet le passage de l'air pulmonaire pour la production vocale chantée ou parlée. Le larynx transforme le souffle en son laryngé (ou glottique). Le larynx renferme les cordes vocales qui sont formées par deux petits muscles recouverts d'une muqueuse très souple qui constitue la partie vibrante, qui s'accolent par influx nerveux et se décollent au passage de l'air. Le larynx n'est pas immobile : il change de position de haut en bas quand on parle. Il se dresse pour les sons aigus et descend pour les graves. Le mouvement des cordes vocales permet de faire une distinction entre les articulations sonores et sourdes. Les sonores impliquent la vibration des cordes vocales par contre au cours des articulations sourdes, elles ne vibrent pas.

C'est à ce niveau que nous pouvons déterminer la fréquence de battement des cordes vocales laquelle détermine la fréquence fondamentale du son et les variations de hauteur (la hauteur de la parole est fixée par la fréquence des mouvements des cordes vocales). L'intensité du son laryngien est liée directement à la pression de l'air. Plus la pression est grande, plus le son est intense. Au niveau des cordes vocales, le son a un timbre uniforme. Il ne change pas lorsque le locuteur essaye de produire des timbres différents. Mais ce sont les lèvres, le voile du palais, la langue qui en transformant la forme du conduit vocal transforment le timbre.

1.2 Modélisation et modification de la parole

Pour apporter des modifications à la voix source, et c'est l'objectif principal du système de conversion vocale, nous devons d'abord savoir comment fonctionne le système de production de la parole (Stylianou, 2009). Un point essentiel du processus de production de la parole consiste donc à reconnaitre comment la parole est produite sur le plan physique. Nous pouvons ensuite utiliser ces informations pour créer un modèle mathématique ou une représentation avec laquelle nous pourrons travailler.

Les pressions d'air présentes dans les régions du larynx situées au-dessus et au-dessous de la glotte sont appelées pressions supra-glottiques et sous-glottiques. La pression sous-glottique des poumons provoque un flux d'air traversant le larynx et le pharynx. Les cordes vocales permettent d'ouvrir et de fermer la glotte (Baer et al.,

1983). Les ouvertures et les fermetures glottales génèrent des impulsions glottales (Baken, 1992). De même, l'effet Bernoulli (Bernoulli, 1738), provoqué ici par des constrictions du larynx (Ladefoged, 1996), génère des turbulences supplémentaires.

Ainsi, les impulsions glottales constituent la partie déterministe et les turbulences d'air, présentent la partie stochastique d'excitation glottale. Ces deux flux passent par le conduit vocal et en particulier au travers des lèvres et des narines (Gopi, 2014). La position des articulateurs (lèvres, mâchoire, langue, etc.) modifie la forme du conduit vocal, affectant ainsi l'air sortant des poumons lors de sa libération et permet de modifier la voix de manière à répartir l'énergie déployée sur des modes vibratoires correspondant aux différents sons d'un langage.

C'est ainsi que différents sons sont produits (Kain, 2001). Différentes fréquences de résonance dans notre langage sont associées à divers formes prises par le conduit vocal. Ces fréquences de résonance sont appelées formants. Les formants peuvent être considérés comme les blocs de construction de notre discours. Ils sont importants pour produire les différents sons de base que nous produisons, appelés phonèmes (Kain, 2001).

Les phonèmes peuvent différer d'une langue à l'autre, ce qui peut engendrer des difficultés lorsque l'on essaye de parler une autre langue, car nous ne sommes pas habitués à entendre ou à prononcer des phonèmes spécifiques à une nouvelle langue. La modification acoustique de la parole peut être obtenue par trois manières fondamentales en manipulant indépendamment sa vitesse, sa fréquence fondamentale ou ses formants. Les deux premières catégories correspondent au niveau prosodique tandis que la troisième correspond au niveau spectral.

— La modification de l'échelle temporelle change la durée de la parole tout en préservant sa fréquence fondamentale et ses propriétés spectrales.
— La modification de la hauteur change la fréquence fondamentale d'un signal de parole tout en préservant la durée d'origine et les propriétés spectrales.
— La modification spectrale modifie la structure des formants tout en préservant la fréquence fondamentale et la durée initiale du signal.

Pour effectuer une conversion de la voix, des méthodes d'analyse/synthèse sont nécessaires pour donner à la parole une représentation paramétrique et pour pouvoir synthétiser la parole à partir d'une représentation paramétrique modifiée. De nombreuses méthodes d'analyse/synthèse ont été proposées dans la littérature, certaines dans le domaine temporel (Hamon et al., 1989; Moulines et Charpentier, 1990; Moulines et Laroche, 1995), certaines dans le domaine fréquentiel (Moorer, 1978; Flanagan et Golden, 1966) et d'autres dans le domaine temps-fréquence (George, 1991; George et Smith, 1997). Nous présentons ensuite brièvement certaines des approches les plus importantes et les plus pertinentes.

1.3 Analyse

Le modèle de parole utilisé pour l'analyse des signaux d'entrée et la reconstruction des signaux modifiés constitue un élément important dans la conception d'un système de conversion vocale. Un modèle de parole adapté à la conversion de voix devrait avoir les caractéristiques suivantes (Eslava et Bilbao, 2008) :

— Il doit permettre une reconstruction de haute-fidélité du signal à partir des paramètres du modèle.
— Il doit permettre la modification des paramètres prosodiques de la parole tels que la fréquence, la durée et l'intensité.
— Il doit permettre les modifications spectrales flexibles sans dégrader la qualité de la voix synthétisée.

Il existe une relation étroite entre les algorithmes de conversion vocale et le système de synthèse. Par contre, leur interaction correcte est importante lors de la transformation des paramètres acoustiques. En fait, le processus d'analyse-synthèse peut introduire des artefacts audibles dans le signal converti. Les modèles de parole les plus couramment utilisés pour la synthèse et la conversion de voix sont présentés ci-après.

1.3.1 Analyse spectrale

Un module d'extraction de caractéristiques utilise la sortie du module d'analyse de la parole en tant qu'entrée pour extraire des représentations de caractéristiques acoustiques. La sortie du module d'analyse de la parole contient généralement des composants spectraux et prosodiques. La dimension des représentations prosodiques, telles que le F0, l'intonation, la durée et l'intensité, est généralement faible et les caractéristiques prosodiques sont faciles à modéliser. Cependant, les représentations spectrales, telles que l'enveloppe spectrale, contiennent une quantité importante d'informations et leur dimension est généralement beaucoup plus élevée. Pour avoir une modélisation robuste de la représentation spectrale, l'extraction de caractéristiques a pour objectif de trouver des représentations de faible dimension à partir des spectres.

Les caractéristiques spectrales sont extraites en suivant les critères suivants : a) être capables de bien représenter l'identité vocale ; b) pouvoir être reconverties en enveloppes spectrales ; c) avoir de bonnes propriétés d'interpolation et permettre une modification flexible.

Ci-dessous, nous décrivons les modélisations spectrales les plus utilisées.

1.3.2 Analyse cepstrale

Dans le cadre de la modélisation source-filtre, l'analyse cepstrale a été largement utilisée. En utilisant la théorie des systèmes linéaires invariants (LTI) (Rabiner et Schafer, 1978), la production de la parole s peut être interprétée par un train d'impulsions e

excitant un système LTI avec une réponse impulsionnelle vt.

$$s(t) = vt(t) * e(t) \tag{1.1}$$

où $*$ représente l'opération de convolution.

La Transformée de Fourier Rapide (FFT : Fast Fourier Transform) est utilisée pour connaitre l'évolution du spectre de la parole.

$$F(k) = \sum_{n=0}^{(N-1)} s(n) exp(-j\frac{2\pi}{N}kn) \tag{1.2}$$

où N représente la longueur de la fenêtre d'analyse, $s(n)$ le signal d'entrée et $F(k)$ le $k^{\text{ème}}$ coefficient spectral complexe. La transformée de Fourier convient aux signaux périodiques, et comme le signal de parole est purement non stationnaire, l'utilisation d'un fenêtrage est impérative. Parmi les fenêtres existantes (rectangulaires, triangulaires, Hanning et Hamming), la fenêtre la plus utilisée est la fenêtre de Hamming (ou Hanning) normalisée représentée par la formule ci-dessous :

$$H(n) = \begin{cases} \frac{2\sqrt{\frac{L}{N}}}{\sqrt{4a^2+2b^2}}(a + b.cos(\frac{\pi(2n+1)}{N}),\ 0 \leq n < N \\ 0,\ otherwise \end{cases} \tag{1.3}$$

La transformée de Fourier, du signal vocal $s(n)$ sera représentée comme suit :

$$F_{mL}(k) = \sum_{n=mL}^{(mL+N-1)} s(n)H(n-mL)exp(-j\frac{2\pi}{N}k(n-mL)) \tag{1.4}$$

où N est la longueur en échantillons de la fenêtre d'analyse, L est le décalage temporel en échantillons entre deux trames consécutives analysées, a = 0.54 et b = -0.46. H représente la fenêtre de Hamming (ou Hanning si a=0.5 et b=-0.5) appliquée au signal $s(n)$. Afin de réaliser une analyse spectrale d'un signal de parole les étapes du traitement homomorphique (Rabiner et Schafer, 1978; Oppenheim et Schafer, 1968) peuvent être utilisées pour séparer les contributions du signal excitatif et du conduit vocal.

1. La Transformée de Fourier Rapide (FFT) : la transformée de Fourier permet le passage de la convolution (formule 1.1) vers une multiplication (formule 1.5) pour obtenir les signaux spectraux.

   $$S(f) = Vt(f)E(f) \tag{1.5}$$

   Où $S(f)$, $Vt(f)$ et $E(f)$ sont respectivement les transformées de Fourier des fonctions $S(t)$ $Vt(t)$ et $E(t)$.
2. (Oppenheim et Schafer, 1968) Afin de séparer les effets du conduit vocal et du signal excitatif, le logarihtme est appliqué à la transformée de Fourier du spectre d'amplitude. En effet, le logarithme d'un produit est la somme des logarithmes de ses facteurs. Ainsi le spectre logarithmique du signal vocal est donné par la formule suivante :

   $$log(S(f)) = log(Vt(f).E(f)) = log(Vt(f)) + log(E(f)) \tag{1.6}$$

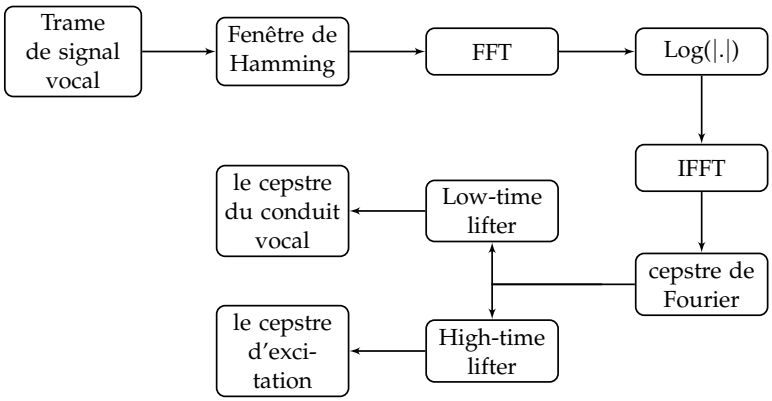


FIGURE 1.3 – *Schéma fonctionnel d'extraction des paramètres cepstraux*

3. Le cepstre : le cepstre $c(t)$ est obtenu par la transformation inverse (IFFT Inverse Fast Fourier Transform) appliqué au spectre logrithmique.

$$c(t) = IFFT(log(S(f))) = IFFT(log(Vt(f))) + IFFT(log(E(f))) \tag{1.7}$$

Le cepstre du signal vocal est formé par le cepstre de l'excitation et le cepstre du conduit vocal.

Le cepstre peut-être décomposé en deux parties : les composants cepstraux qui varient lentement avec la fréquence déterminée par un facteur de temps faible $F_l(n)$ et les composants cepstraux qui varient rapidement avec la fréquence déterminée par $F_h(n)$:

$$F_l(n) = \begin{cases} 1, & 0 \leq n \leq P \\ 0, & P < n \leq \dfrac{N}{2} \end{cases} \tag{1.8}$$

$$C_{vt} = F_l(n)c(n) \tag{1.9}$$

$$F_h(n) = \begin{cases} 1, & P < n \leq \dfrac{N}{2} \\ 0, & elsewhere \end{cases} \tag{1.10}$$

$$C_{exc} = F_h(n)c(n) \tag{1.11}$$

Où C_{vt} est le cepstre du conduit vocal ; C_{exc} est le cepstre de l'excitation glottale ; N est la taille de la FFT ; P est le nombre de coefficients du cepstre du conduit vocal lorsque le premier coefficient cepstral $c0$ est écarté.

1.3.3 Les coefficients LSF (Line Spectral Frequency)

Les coefficients LSF ont de bonnes propriétés d'interpolation et permettent une bonne représentation des formants (Itakura, 1975). En plus, ils ont été appliqués avec

succès au codage et à la synthèse de la parole. Cependant, il existe des problèmes avec les coefficients LSF en conversion de voix. En effet, si la fonction de conversion n'est pas stable, il est difficile de garantir que les coefficients LSF convertis se situent dans la plage $[0, \pi]$ par ordre croissant. Cela peut rendre le filtre de synthèse instable. Malgré ce problème, les coefficients LSF ont été largement utilisés dans la conversion de voix (Helander et al., 2008a; Erro et al., 2010b).

1.3.4 Estimation des descripteurs de voix de base

Le pitch

Le pitch de la voix humaine est la fréquence fondamentale moyenne perçue par l'oreille : c'est un paramètre fondamental dans la production, l'analyse et la perception de la parole. Cette fréquence donne des informations d'intonation de la parole et aussi des informations concernant le locuteur. Il permet de différencier la voix d'une femme de celle d'un homme. Son évolution détermine la mélodie de la parole et dépend de l'état physique et mental du locuteur. Les variations du pitch, nous donnent des informations sur l'émotivité d'un locuteur. C'est grâce à lui que l'on peut faire la distinction entre une affirmation, une interrogation ou un ordre. Le pitch varie autour de 120 Hz pour les hommes, de 240 Hz pour les femmes et de 350 Hz pour les enfants (Pépiot, 2013).

La fréquence fondamentale F0

La fréquence fondamentale est la fréquence de vibration des cordes vocales. Dans le domaine temporel, $F0_t$ est la période d'un signal voisé à un instant t. F0 représente un indicateur émotionnel et elle est considérée comme étant la quantité à estimer par un algorithme de détection du pitch.

Pour le signal de parole, sa fréquence fondamentale est la fréquence du cycle d'ouverture et fermeture des cordes vocales.

Selon plusieurs chercheurs, la différence entre les fréquences fondamentales moyennes des hommes et des femmes s'explique par la physiologie des plis vocaux. D'après (Kahane, 1978), les plis vocaux sont plus courts et moins épais chez les femmes : ils produisent un effort phonatoire plus important et vibrent à une fréquence plus élevée. Les plis vocaux des hommes sont en moyenne 60 % plus longs que ceux des femmes, comme l'illustre la figure 1.4. Dans la littérature, il existe de nombreuses méthodes pour estimer la fréquence fondamentale F0 à partir d'un signal audio que ce soit dans le domaine temporel (Gold et Rabiner, 1969; Bagshaw et al., 1993) ou dans le domaine temporel fréquentiel (Schroeder, 1968; Noll, 1970; Bahja et al., 2015, 2016). La détection de F0 est difficile à cause des problèmes suivants :

— Le suivi de F0 en temps réel est difficile.
— La présence d'harmoniques faussent souvent la détection du F0.
— La mise en œuvre des algorithmes de décision voisé/non voisé est complexe.

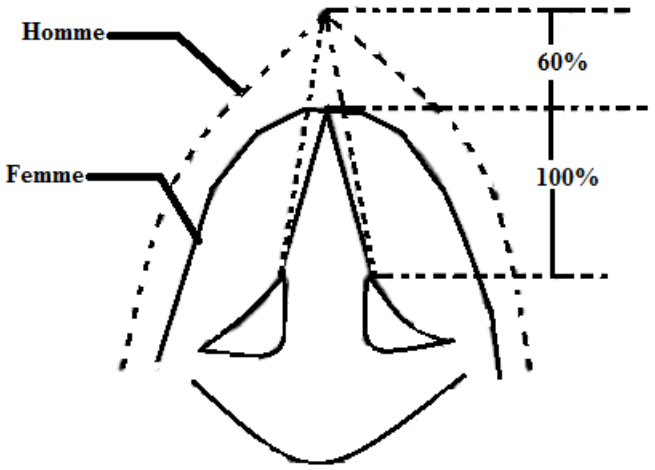


FIGURE 1.4 – *Coupe horizontale du larynx d'un homme et d'une femme, les pourcentages dans la partie droite illustrent la différence de taille des plis vocaux (Kahane, 1978)*

1.4 Modélisation et Synthèse

1.4.1 TD-PSOLA (Time-Domain Pitch-Synchronous Overlap-Add)

La méthode TD-PSOLA (Time-Domain Pitch-Synchronous Overlap-Add) (Moulines et Charpentier, 1990) est une technique de synthèse très populaire qui permet des modifications prosodiques et fournit une voix synthétisée de haute qualité. Elle fonctionne en échantillonnant des parties fenêtrées du signal d'origine, puis en les synthétisant avec la technique addition-chevauchement. Cependant, comme la modification de la parole est effectuée directement à partir des échantillons, le procédé manque de contrôle sur les enveloppes spectrales, ce qui le rend inapproprié à la conversion vocale. La méthode PSOLA (Pitch-Synchronous Overlap and Add) utilise des fenêtres qui se chevauchent et agit directement sur la forme d'onde par trame, permettant ainsi des manipulations souples de la durée, de la fréquence et des formants. Pour modifier l'échelle de temps, par exemple, les trames sont soit répétées, soit supprimées, les marques de fréquences restent inchangées. Une modification de la fréquence est obtenue en ajustant l'espacement entre les marques du fondamental. Pour la manipulation des enveloppes spectrales, les variantes FD-PSOLA et LP-PSOLA (Moulines et Verhelst, 1995) peuvent être utilisées.

Malgré un son de haute qualité, le procédé est moins approprié pour les modifications fines (Nguyen, 2009).

1.4.2 Modèles sinusoïdaux

Dans le modèle sinusoïdal, la forme d'onde de la parole est représentée localement comme une somme de sinusoïdes dont les paramètres varient avec le temps. Le modèle sinusoïdal convient à tous types de transformations de la voix pour plusieurs raisons :

— Il permet l'obtention d'une bonne qualité audio du son reconstruit et la modification prosodique.
— Les caractéristiques du modèle conviennent à la synthèse de la parole concaténative, car elles permettent de lisser les discontinuités spectrales à la limite de la transition entre deux unités adjacentes.
— Les paramètres du modèle contiennent des informations sur la forme d'onde et le spectre. Ils peuvent être utilisés pour obtenir des estimations des enveloppes spectrales d'amplitude et de phase. Ils permettent ainsi une manipulation spectrale et une conversion de voix flexibles.
— Le modèle sinusoïdal est compatible avec la plupart des méthodes de conversion vocale et donc il a été adopté par de nombreux systèmes.

On peut distinguer le modèle proposé par McAulay and Quatieri (McAulay et Quatieri, 1986; Quatieri et McAulay, 1986) et le modèle ABS/OLA (George, 1991; George et Smith, 1997) implémenté par George et Smith. Ces deux méthodes permettent de représenter la parole sous forme d'une somme de sinusoïdes variant dans le temps et dont les paramètres d'amplitude, de fréquence et de phase sont estimés à partir de la transformée de Fourier à court terme en utilisant un algorithme de sélection de crêtes.

modèle harmonique plus bruit (HNM) Les modèles harmoniques sont un cas particulier de modèles sinusoïdaux. Le modèle harmonique suppose qu'un signal de parole est représenté comme une composante harmonique additionnée à une composante de bruit qui sont séparées dans le domaine des fréquences par une fréquence de coupure F_c, appelée fréquence maximale de voisement (Stylianou, 2001). La composante harmonique est modélisée par la somme des sinusoïdes harmoniques jusqu'à la fréquence maximale de voisement. Au-delà de cette fréquence, le spectre est modélisé par un filtre excité par un bruit blanc gaussien.

1.4.3 Modèle source-filtre

Lorsque nous utilisons un modèle source-filtre la parole est obtenue par la convolution d'un signal excitatif avec la réponse impulsionnelle d'un filtre (Mohammadi et Kain, 2017). Le signal source (ou d'excitation) représente l'air sortant par les cordes vocales et le filtre représente le conduit vocal (voir la figure 1.5). Le signal source et le filtre sont considérés comme indépendants : nous pouvons les modifier indépendamment l'un de l'autre. Comme les cordes vocales changent constamment avec la position des articulateurs, le filtre est considéré comme un filtre variant dans le temps. Un signal d'excitation voisé peut être modélisé comme un train d'impulsions dans lequel la distance entre les impulsions varie avec la fréquence fondamentale du locuteur. Par contre, le signal non voisé peut être modélisé comme un signal de bruit. En réalité, il s'agit d'une simplification dans la mesure où certains sons peuvent ne pas être classés comme étant purement voisés ou non voisés. Cependant, cette simplification est utile pour permettre une modélisation simple d'un signal source par le modèle source-filtre. Les modèles source-filtre sont flexibles et permettent une modification aisée d'un signal

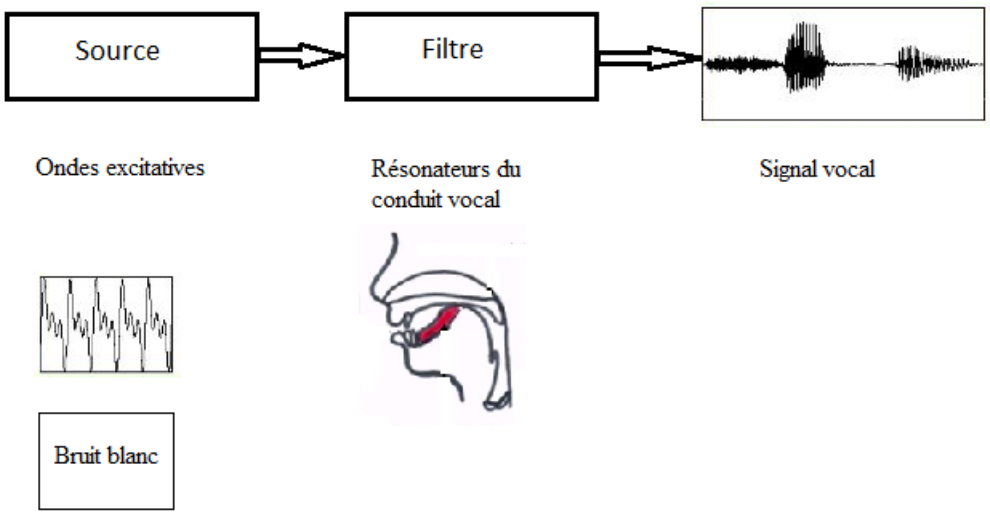


FIGURE *1.5* – *Modèle source-filtre*

de parole (Mohammadi et Kain, 2017), et donc peuvent s'avérer utiles pour la conversion vocale.

STRAIGHT

Le modèle STRAIGHT ou "Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum" (Kawahara et al., 2001, 1999) est basé sur le modèle source-filtre [37]. STRAIGHT est un vocodeur de haute qualité, qui décompose le signal de parole en trois composants : le spectrogramme, les paramètres apériodiques et le contour de la fréquence F0. Le modèle STRAIGHT permet une manipulation de la parole flexible, car les trois composants sont indépendants.

Ce modèle a été largement utilisé dans les applications de conversion vocale (Helander et al., 2010; Toda et al., 2007; Desai et al., 2009; Helander et al., 2012). En outre, parmi les 17 équipes qui ont participé au "Voice Conversion Challenge 2016 (VCC2016)" pas moins de 11 d'entre elles ont utilisé STRAIGHT dans le cadre de leurs systèmes de conversion de la voix (Toda et al., 2016).

Reconstruction d'un signal audio à partir de ses spectres d'amplitude de Fourier modifiés

Le premier algorithme pour la reconstruction du signal à partir de ses spectres d'amplitude de Fourier a été introduit par Griffin et Lim (Griffin et Lim, 1984) (Griffin-Lim

Algorithm (GLA)) il y a plus de 30 ans : il est fondé sur une méthode itérative pour estimer un signal audio réel. L'un des inconvénients de la méthode Griffin-Lim est qu'elle nécessite généralement de nombreuses itérations pour obtenir des signaux audio de bonne qualité. De plus cet algorithme est intrinsèquement non temps-réel. De plus ; il est relativement lourd en temps de calcul. Une version en temps réel de GLA a été proposée par (Beauregard et al., 2005) (Real-Time Iterative Spectrogram Inversion RTISI). L'algorithme est toujours itératif, mais le signal est reconstruit trame par trame à l'aide d'une initialisation intelligente de la phase, de sorte que seules quelques itérations sont nécessaires pour obtenir un bon résultat. Par la suite , RTISI a été étendu par (Zhu et al., 2006) en RTISI-LA (Real-Time Iterative Spectrogram Inversion with Look-Ahead) puis légèrement modifié avec l'approche de de Gnann et Spiertz (GSRTISI-LA) (Gnann, 2014), avec l'ajout d'un procédé d'estimation de la phase.

1.5 Conclusion

Ce chapitre a un rôle introductif au domaine de la conversion vocale. Pour développer un système de conversion de voix, il est nécessaire de bien comprendre les mécanismes de production de la voix et il faut avoir une bonne connaissance des paramètres acoustiques caractérisant l'identité du locuteur. Au cours de ce chapitre, nous avons décrit l'anatomie du système de production de la parole. La section 1.1 présente la manière dont le système de production de la voix humaine génère la parole, ce qui est tout à fait nécessaire pour comprendre les moyens mis en œuvre pour construire un système de conversion vocale. Certains descripteurs de voix de base sont présentés dans les sections 1.3 et 1.3.4. La modélisation et la synthèse de la parole ont été introduites dans la section 1.4. Le chapitre suivant présente les systèmes de conversion que nous proposons ainsi que leurs évaluations objectives et subjectives.

Chapitre 2

Conversion de la voix

Sommaire

FIGURE 2.1 – *La conversion de la voix.*

2.1 Introduction

Ce chapitre concerne la mise en œuvre d'un système de conversion de voix ainsi qu'une description des méthodes et des techniques de conversion présentées dans la littérature. Ce chapitre comprend deux parties. La première introduit le concept et les étapes à mener pour effectuer la conversion de voix. La deuxième partie explique les schémas proposés et les résultats obtenus.

2.2 Principes d'un système de conversion de voix

La conversion vocale (VC) a pour objectif de transformer les caractéristiques du signal de parole de la voix d'un locuteur source de telle sorte qu'il soit perçu et reconnu comme s'il avait été prononcé par un locuteur cible. En d'autres termes, un système de conversion de voix modifie les caractéristiques du signal de parole du locuteur, telles que la forme spectrale, les formants, la fréquence fondamentale F0, l'intonation, l'intensité et la durée, de manière à changer l'identité du locuteur tout en conservant les informations linguistiques. Le module principal d'un système de conversion vocale est la fonction de conversion dont la tâche est de transformer les caractéristiques du locuteur source en celles du locuteur cible. La fonction de conversion peut être formulée comme suit :

$$Y = F(X) \tag{2.1}$$

où $X \in R^{d\times 1}$ et $Y \in R^{d\times 1}$ sont respectivement les caractéristiques source et cible. d est la dimension des vecteurs acoustiques et $F(.)$ est la fonction de conversion permettant de transformer les vecteurs acoustiques source en vecteurs cible.

La conversion vocale peut être décrite comme étant le processus de conversion permettant de transformer l'identité vocale d'un locuteur source en celle d'un locuteur cible.

De nos jours, les systèmes de conversion de la voix cherchent à calculer des fonctions de transformation permettant la conversion de la voix d'un locuteur source en celle d'un locuteur cible. Ces systèmes préservent généralement l'intonation expressive et la prosodie du locuteur source.

La figure 2.2 illustre un système de conversion de voix typique comprenant deux phases : la phase d'apprentissage et la phase de conversion. La phase d'apprentissage nécessite généralement des corpus parallèles (constitués de phrases parallèles) qui sont préparées pour un locuteur source X et un locuteur cible Y.

Ces données sont ensuite analysées par un modèle de production de la parole pour calculer des caractéristiques acoustiques indépendantes tels que le spectre, l'excitation et la fréquence F0. Ces données sont ensuite transmises à un module d'extraction de caractéristiques qui génère des vecteurs de faibles dimensions tels que les coefficients cepstraux ou les coefficients de prédiction linéaires prédictifs (LPCC). Pour chaque énoncé parallèle, des techniques d'alignement de trames, tel que l'alignement temporel dynamique (DTW) (Stylianou et al., 1998), sont utilisées pour apparier deux à deux les vecteurs acoustiques source X et cible Y.

$$X = \{x_1, x_2, ...x_n.....x_K\} \tag{2.2}$$

$$Y = \{y_1, y_2, ...y_n.....y_K\} \tag{2.3}$$

où $X \in R^{d \times K}$ et $Y \in R^{d \times K}$ sont respectivement les données appariées source et cible. K est le nombre de trames et x_k et y_k sont respectivement les trames source et cible d'indice k.

La fonction de conversion optimale $F(.)$ est estimée à partir des vecteurs appariés. Une fonction de conversion exprime la relation entre la voix source et la voix cible en fonction de l'ensemble des paramètres choisis pour définir les caractéristiques de la voix cible. Une transformation de haute qualité est nécessaire pour préserver la qualité de la parole d'origine. Cette fonction $F(.)$ peut être par la suite appliquée pour transformer une nouvelle séquence de vecteurs acoustiques source. Enfin, les séquences de paramètres transformés sont transmises à un module de reconstruction de la parole dont le but est de générer un signal de parole contenant le même discours linguistique qui a été prononcé par le locuteur source avec un style d'expression similaire, et qui doit être perçu comme s'il avait été prononcé par le locuteur cible. Les modules d'analyse et de reconstruction de la parole sont importants dans un système de conversion vocale. Comme montré dans la figure 2.2, ces deux modules sont fortement corrélés l'un à l'autre. Si le module d'analyse permet de représenter le signal comme un ensemble de composants acoustiques indépendants les uns des autres afin de permettre des modifications flexibles, le module de reconstruction doit recréer le signal à partir de la même représentation. La flexibilité des vecteurs acoustiques est importante dans un système de conversion vocale, car les représentations introduites dans le module de reconstruction de la parole sont celles qui ont été modifiées dans le processus de conversion. Le but principal de ces modules est de reconstruire un signal de haute qualité.

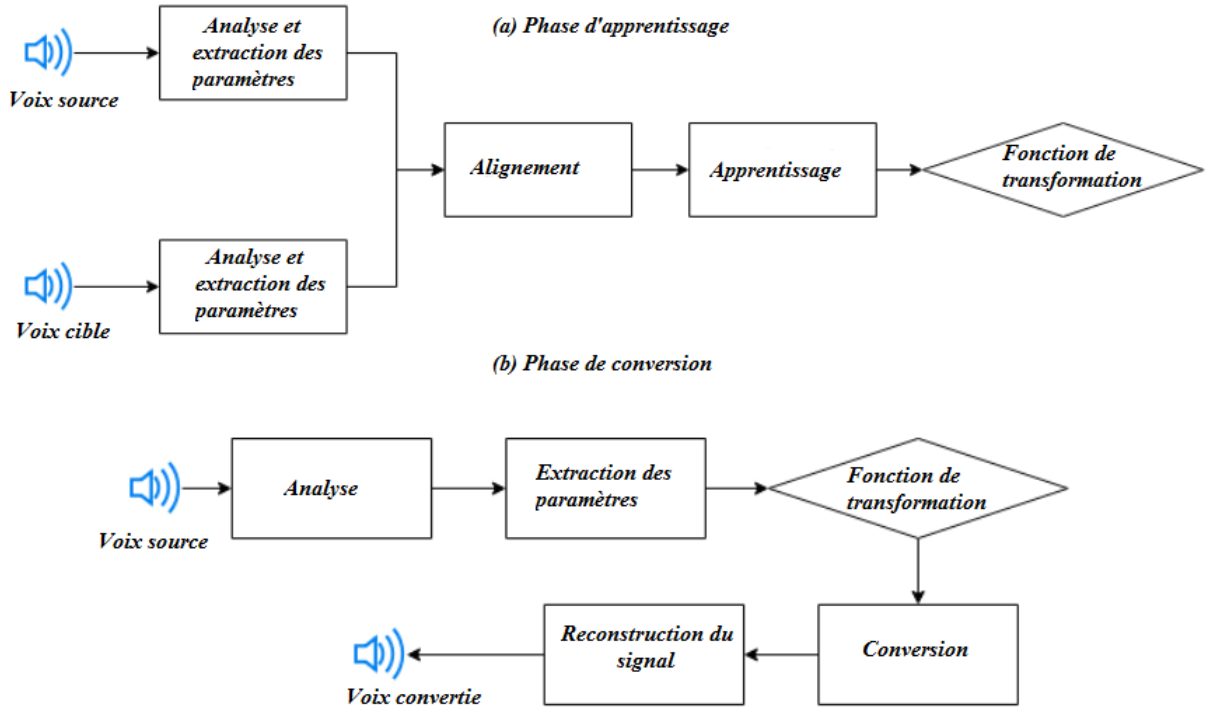


FIGURE 2.2 – *Architecture d'un système de conversion de voix.*

2.3 Applications de la conversion vocale

La conversion de la voix a de nombreuses applications intéressantes. Les objectifs sont prometteurs pour les différents domaines où la voix joue un rôle important, tels que les jeux vidéo, les vidéos, les films, l'animation et le doublage. D'autres domaines du traitement audio général, tels que la production musicale, le multimédia ou la traduction parole-parole, trouvent aussi un intérêt certain dans l'utilisation des techniques de conversion vocale.

La technologie relativement nouvelle a fait l'objet d'une attention croissante au sein de la communauté des chercheurs en parole au cours des dernières années en raison des améliorations récentes de la qualité de la synthèse et du score élevé de conversion de l'identité du locuteur source. La transformation en voix spécifiques peut trouver une utilisation dans :

1. **Recréation de voix** : La recréation de voix de personnes humaines décédées sur la base d'anciens enregistrements.
2. **Doublage de la voix** : La conversion vocale peut être utilisée dans les films ou les jeux vidéo afin de transformer la voix d'un utilisateur dans l'identité vocale d'une autre personne. Cela permettrait, par exemple, d'éviter aux producteurs de films ou de jeux vidéos de payer des frais élevés à une célébrité, pour prononcer toutes les phrases. Ainsi un utilisateur du jeu vidéo pourrait agir vocalement

comme un acteur/actrice connu(e).

3. **Conversion de voix multilingue** :
   — Doublage du film : L'identité vocale d'un acteur/actrice célèbre dans sa langue maternelle, en tant que langue source, pourrait être transformée dans une autre identité d'une langue cible. Cela n'est pas simple à réaliser car la langue cible peut présenter un contenu linguistique et contextuel différent et peut contenir une couverture phonétique non complète par rapport à la langue source.
   — Traduction parole-parole : Cette technique pourrait préserver l'identité vocale d'un locuteur dans un discours enregistré et traduit dans une autre langue.

4. **Extension de corpus Text-to-Speech (TTS)** :

   De nouvelles voix pour un système de synthèse vocale TTS peuvent être créées à partir d'un corpus existant sans qu'il soit nécessaire d'enregistrer et d'étiqueter la voix d'un autre locuteur pour une nouvelle base de données TTS. Actuellement, la création d'une nouvelle voix est relativement coûteuse et nécessite des enregistrements de plusieurs heures. Les systèmes de la conversion vocale nécessitent une quantité relativement faible d'enregistrements. De nouvelles voix pourraient être obtenues par l'utilisation d'une base de données peu volumineuse de la voix d'un locuteur cible. La création d'un corpus de parole expressive à partir d'un corpus caractérisé par un style de parole normal est possible avec un système de conversion vocale.

5. **Amélioration des systèmes de transformation vocale** :
   — La transformation souhaitée peut d'abord être obtenue approximativement à l'aide de la conversion vocale. Par la suite les techniques de transformation vocale peuvent être appliquées. Une transformation de voix d'homme en celle d'une femme peut être obtenue en utilisant d'abord une technique de conversion vocale à l'aide d'une corpus de voix cible féminine. La modification de la voix convertie peut ensuite être appliquée au moyen d'algorithmes de transformation vocale. La conversion vocale doit être ajustée pour permettre l'obtention d'une qualité audio élevée.
   — La méthode opposée consiste à appliquer d'abord la transformation vocale puis à utiliser la conversion vocale. L'algorithme de transformation vocale souhaité est préalablement appliqué suivi par la conversion vocale pour affiner le signal converti en fonction d'un caractère de voix prédéfini.

6. **Pathologie vocale** :
   — Amélioration de la parole pour les voix alaryngées : les voix de personnes souffrant de troubles de la voix peuvent être converties en parole laryngée, à l'aide d'un système de conversion vocale, afin de retrouver une qualité de voix naturelle.
   — Formation vocale : une interface de formation vocale pour les patients souffrant de troubles de la voix peut être réalisée au moyen d'un système de conversion vocale.

2.4 Techniques de base d'une conversion vocale

2.4.1 Corpus de parole parallèles et non parallèles

Les systèmes de conversion vocale diffèrent généralement par la manière dont les données de locuteurs sont fournies et traitées. Les systèmes utilisant des corpus parallèles nécessitent des enregistrements de mêmes phrases pour les locuteurs source et cible. Ces phrases sont généralement phonétiquement équilibrées. Un corpus de formation non parallèle nécessite des informations phonétiques et linguistiques supplémentaires afin de regrouper des segments similaires et définir les catégories phonétiques (Machado et Queiroz, 2010). Les données parallèles peuvent être générées à partir des données non parallèles en regroupant des trames ou des segments identiques. De même, la sélection d'unités peut être utilisée pour apparier des phonèmes, des diphones, des syllabes ou même des mots sources et cibles similaires. L'adaptation du modèle à des données non parallèles consiste à regrouper des segments similaires pour l'apprentissage du modèle de conversion.

2.4.2 Alignement

En général, toutes les techniques de conversion vocales passent par une étape d'alignement des données d'apprentissage. Nous pouvons citer deux approches d'alignement : parallèle et non-parallèle.

Alignement parallèle

Dans certaines situations où l'ensemble de données est constitué de paires d'énoncés de deux locuteurs différents avec le même contenu linguistique (les phrases du locuteur source et les phrases du locuteur cible sont identiques), on dit qu'on dispose de données parallèles. Lorsque les deux fichiers parallèles source-cible sont analysés, le nombre des trames source/cible ne sont pas identiques généralement, et il est extrêmement improbable que les deux locuteurs émettent les mêmes phonèmes dans les mêmes temps. Comme les durées des énoncés parallèles sont souvent différentes, un alignement temporel est utilisé pour apparier les vecteurs acoustiques des voix source et cible. Cette phase consiste à associer un vecteur source à un vecteur cible qui se correspondent. Après l'alignement, x_t et y_t sont respectivement les vecteurs acoustiques source et cible pour le temps t.

Alignement manuel : Certains auteurs comme (Mizuno et Abe, 1995), afin d'améliorer le système de conversion, alignent les formants de manière manuelle. Mais cette technique est peu efficace car le temps nécessaire pour réaliser cet alignement manuel est trop important et non pratique dans le cadre d'un apprentissage automatique.

Alignement par DTW (Dynamic Time Warping) : Lorsqu'on dispose de données parallèles, l'alignement par DTW est la technique la plus utilisée dans les systèmes de conversion vocale. La DTW permet de trouver le chemin optimal qui minimise la distance cepstrale entre les trames des locuteurs source et cible. Cette technique peut être appliquée à des phrases complètes (Stylianou et al., 1998) ou même à des diphones (Kain et Macon, 2001).

Alignement par HMM : En 1999, (Arslan, 1999) a aligné les vecteurs cepstraux via la technique HMM. Le principe est fondé sur la sélection des phrases prononcées par les deux locuteurs source et cible. Pour chaque phrase, les coefficients cepstraux sont extraits et les régions de silence supprimées, pour appliquer le HMM. Le nombre d'états de chaque HMM par phrase est proportionnel à la durée de cette phrase. La meilleure séquence d'états par phrase est estimée par l'algorithme Viterbi (Viterbi, 1967). Les coefficients LSF moyen, sont calculés par les vecteurs de chaque état correspondant à chacun des locuteurs. Ces coefficients sont collectés pour définir le dictionnaire des trames alignées source-cible. L'année 2009 a été marquée par les démonstrations effectuées par (Godoy et al., 2009), qui démontrent que cet alignement est influencé par le problème de "one-to-many". On rencontre ce problème lorsqu'une trame source est associée à plusieurs trames cible de caractéristiques spectrales différentes. Dans ce cas, il est impossible d'associer une trame source à une seule trame cible.

Alignement non parallèle

Lorsque les données parallèles ne sont pas disponibles, DTW n'est pas applicable. Plusieurs techniques d'alignement de trames non parallèles ont été proposées pour résoudre ce problème. L'une des techniques les plus simples est la méthode d'alignement par trame, dans laquelle chaque trame source est associée à la trame cible la plus proche en réduisant la distance euclidienne. Cette approche ne nécessite aucune information linguistique ou contextuelle et a été utilisée comme base de référence dans plusieurs études (Ney et al., 2004; Erro et al., 2010a).

2.4.3 Fonctions de transformation

Dans la littérature, plusieurs techniques de conversion de voix ont été proposées, bien qu'il s'agisse d'un domaine d'étude relativement récent. Dans cette section, nous analysons certaines des techniques de base afin de donner un bref aperçu au lecteur des différentes possibilités disponibles pour réaliser un système de conversion vocale.

Conversion de voix par quantification vectorielle (VQ)

La conversion de voix par quantification vectorielle (Abe et al., 1990) est la première technique proposée. Cette méthode consiste à projeter les paramètres acoustiques d'un

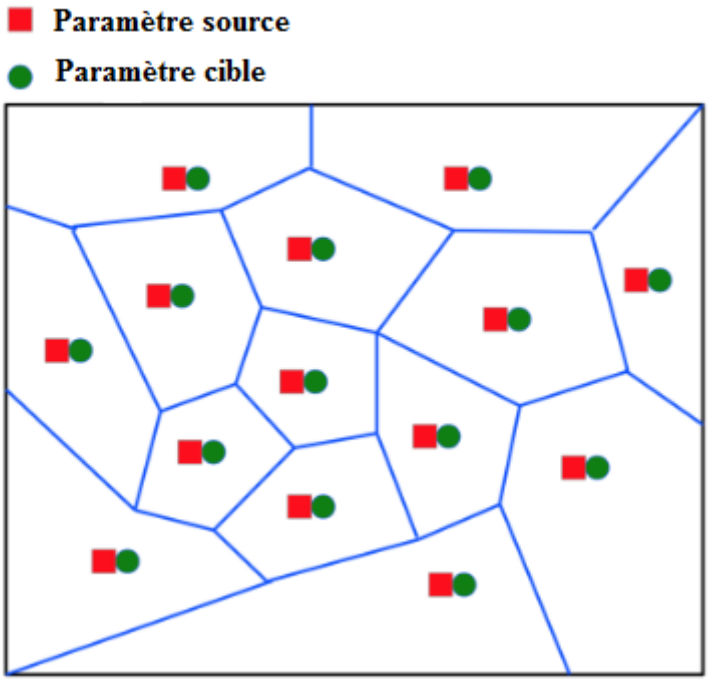


FIGURE 2.3 – *Exemple d'une quantification vectorielle (Wu, 2015).*

espace vectoriel multidimensionnel, dans un sous-espace discret de plus petite dimension constitué de classes. La quantification vectorielle (VQ) représente un moyen simple et direct pour prédire le vecteur converti. VQ comprend une phase d'apprentissage et une phase de test, qui sont similaires à ceux d'un système de conversion vocale typique illustré dans la figure 2.2. Au cours de la phase d'apprentissage, la création des classes constituées de vecteurs source et cible joints est effectuée de manière à construire un dictionnaire, comme illustré à la figure 2.3. Durant le test, chaque vecteur source est associé à un vecteur cible grâce à la liste de correspondances réalisée par le dictionnaire, qui sera ensuite sélectionné comme le vecteur converti. L'implémentation classique de (Abe et al., 1990) utilise l'alignement DTW pour effectuer les correspondances entre les centroïdes source et cible, qui sont ensuite accumulées pour former un histogramme. Cet histogramme sera utilisé par la suite comme fonction de pondération. Le plus gros problème de cette technique est due à la classification qui produit des incohérences au niveau des trames du signal converti. En effet elle n'assure qu'une représentation discrète de la transformation.

Conversion de voix par mélange de gaussiennes (GMM)

Les méthodes basées sur le modèle de mélange de gaussiennes (GMM) sont les méthodes de base actuellement utilisées dans la plupart des recherches effectuées sur la

conversion de la voix. La technique a été proposée dans (Toda et al., 2007; Kain et Macon, 1998; Stylianou et al., 1998) pour remédier aux défauts de la méthode de quantification vectorielle. Des recherches antérieures ont montré que les fonctions de conversion linéaires produisaient de bons résultats. Cependant, le fait d'avoir une seule fonction globale de transformation limite les performances du système.

Par conséquent, les méthodes de conversion basées sur les GMMs modélisent les données avec un modèle de mélange gaussien et recherchent des transformations linéaires locales pour chaque gaussienne utilisée. À cet égard, deux approches principales sont adoptées : la méthode standard initialement proposée par Styliannou qui consiste à modéliser les paramètres source avec un GMM (Stylianou et al., 1998) et une méthode améliorée (JD-GMM : Joint Density GMM) basée sur l'estimation de la probabilité conjointe source/cible (au lieu du modèle source) (Toda et al., 2007; Kain et Macon, 1998).

Le modèle de mélange gaussien de densité source : La méthode statistique basée sur GMM (Stylianou, 1996) divise l'espace acoustique du locuteur source en classes acoustiques disjointes. La classification probabiliste GMM permet d'associer à un vecteur acoustique une certaine probabilité d'appartenir à chacune des classes acoustiques modélisées.

Chaque composant gaussien est représenté par une distribution normale gaussienne $N(x, \mu, \Sigma)$. Les paramètres initiaux GMM (α, μ, Σ) sont estimés via : l'erreur quadratique moyenne (MSE) (Kain et Macon, 1998), l'algorithme espérance-maximisation (EM) (Bishop, 2006) et l'algorithme du maximum de vraisemblance (ML) (Dempster et al., 1977).

La distribution de probabilité d'un vecteur source x est définie par la fonction suivante :

$$p(x) = \sum_{i=1}^{n} \alpha_i N_i(x, \mu_i, \Sigma_i) \tag{2.4}$$

Après la classification par GMM, la fonction de conversion source/cible est représentée sous la forme de la régression linéaire suivante :

$$F(x) = \sum_{i=1}^{n} p(i|x)[\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(x - \mu_i^x)] \tag{2.5}$$

Où :

— n est le nombre de classes acoustiques.
— $p(i|x)$ est la probabilité d'observer la classe i sachant x (Bishop, 2006).

$$p(i|x) = \frac{\alpha_i N_i(x, \mu_i, \Sigma_i)}{\sum_{j=1}^{n} \alpha_j N_j(x, \mu_j, \Sigma_j)} \tag{2.6}$$

— μ_i^x est le centroïde de la classe source i.
— μ_i^y est le centroïde de la classe cible i.

— Σ_i^{xx}est la matrice de covariance de la classe source i.
— Σ_i^{yx} est la matrice de covariance croisée entre les vecteurs y de la classe cible i et les vecteurs x de la classe source i.

La matrice de covariance croisée Σ_i^{yx} et le vecteur moyen μ_i^y de la gaussienne i sont déterminés en minimisant la distance quadratique suivante entre les vecteurs convertis et les vecteurs cible :

$$E = \sum_{k=1}^{m} ||y_k - F(x_k)||^2 \tag{2.7}$$

x_k et y_k sont respectivement les vecteurs source et cible et m est le nombre de vecteurs source/cible appariés.

Le modèle de mélange gaussien conjoint (JD-GMM ou Joint Density GMM) : Le modèle de mélange gaussien de densité source a été étendu par (Kain et Macon, 1998) à un modèle de mélange gaussien conjoint. Il permet de créer un modèle source/cible. Cependant avant la classification, il faut concaténer la séquence des vecteurs acoustiques correspondant à la voix du locuteur source $X_i = [X_i^1 ... X_i^P]$ avec la séquence des vecteurs acoustiques correspondant à la voix du locuteur cible $Y_i = [Y_i^1 ... Y_i^P]$ (précédemment mis en correspondance par l'alignement DTW) en un seul vecteur étendu $z = [x, y]^T$ afin de déterminer les paramètres GMM initiaux ($\alpha, \mu_x, \mu_y, \Sigma_{xx}, \Sigma_{yx}$) pour la probabilité conjointe $p(z)$ (Helander et al., 2008b).

$$p(z) = p(x, y) = \sum_{i=1}^{n} \alpha_i N_i(z, \mu_i, \Sigma_i) \tag{2.8}$$

Chaque gaussienne est représentée par un vecteur moyen μ, une matrice de covariance Σ et un poids de mélange α, avec $\sum_{i=1}^{n} \alpha_i = 1$ et $\alpha_i \geq 0$ pour tout i .

$$\mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix} et\ \Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix}$$

Une fois la classification par GMM effectuée, les paramètres du modèle peuvent être utilisés pour mettre en œuvre la fonction de transformation.

La fonction de transformation $F(x)$ qui minimise l'erreur quadratique moyenne (MSE) entre les vecteurs convertis et cible est définie comme suit :

$$F(x) = E[y|x] = \int y p(y|x) dy = \sum_{i=1}^{n} p(i|x)[\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(x - \mu_i^x)] \tag{2.9}$$

Cependant, les méthodes basées sur GMM rencontrent toujours des problèmes non résolus. La moyenne et la covariance de chaque composant gaussien sont calculées à l'aide de tous les vecteurs d'apprentissage, ce qui conduit à un sur-lissage des paramètres de la voix convertie. La solution proposée dans (Toda et al., 2007) consiste à calculer la variance globale relative aux vecteurs acoustiques convertis. Malgré le fait

que cette approche offre une réduction des erreurs au cours de la conversion, la qualité de la voix synthétisée est dégradée, car plusieurs informations indispensables à la reconstitution de la voix sont perdues.

Approche par réseaux de neurones

Avec le succès des réseaux de neurones (NN) dans le domaine de l'apprentissage automatique, de nombreux chercheurs ont tenté d'appliquer cette technique au problème de la conversion vocale. Ces réseaux sont exceptionnellement efficaces pour apprendre des modèles non linéaires, qui jusque-là n'avaient pas été explorés dans le contexte de la conversion vocale. La forme la plus simple du réseau de neurones est appelée réseau de neurones artificiels (ANN). Cependant, plus récemment, un nouveau champ d'étude est apparu au sein du domaine d'apprentissage automatique : le modèle d'apprentissage profond (Deep Learning). La principale différence par rapport au modèle ANN classique est que les réseaux de neurones profonds (DNN) sont généralement construits à partir de plusieurs couches cachées. Dans les sous-sections suivantes, nous explorerons les approches basées sur les réseaux de neurones, en commençant par les premières approches ANN et en terminant par les techniques plus récentes d'apprentissage profond.

Les réseaux de neurones artificiels : Les réseaux de neurones artificiels (ANN) représentent un outil puissant pour la modélisation des relations complexes non linéaires entre un vecteur d'entrée et un vecteur de sortie. Les modèles ANN sont constitués par des couches, chaque couche contenant un certain nombre de neurones. Il y a principalement trois types de couches : la couche d'entrée qui est l'endroit où les vecteurs d'entrée sont placés ; la couche de sortie qui est l'emplacement où nous obtenons la sortie du réseau et qui au niveau de l'apprentissage réceptionne les vecteurs de sortie. Enfin, nous avons les couches dites cachées, qui sont toutes des couches situées entre les couches d'entrée et de sortie. Dans la figure 2.4, nous considérons un exemple de réseau de neurones. Comme expliqué auparavant, le réseau de neurones est constitué de la couche d'entrée X_Q, de la couche de sortie Y_M et de nombreuses couches cachées. Les réseaux de neurones avec plus d'une couche cachée sont appelés réseaux de neurones profonds. Chaque nœud reçoit l'entrée pondérée des connexions qu'il a avec la couche qui le précède :

$$y_k = \sum_i \omega_i x_i \tag{2.10}$$

Les réseaux de neurones ont été proposés dans le domaine de la conversion vocale presque à la même période de l'avènement des méthodes basées sur les GMMs, mais ils n'ont pas eu un grand succès à cause du réglage difficile de l'architecture optimale du réseau. L'une des premières approches basée sur des réseaux de neurones artificiels appliqués à la conversion de voix a mis en œuvre un simple réseau de neurones à propagation directe (feedforward) pour transformer les formants de la voix source en ceux de la voix cible (Narendranath et al., 1995) et un vocodeur LPC pour resynthétiser la

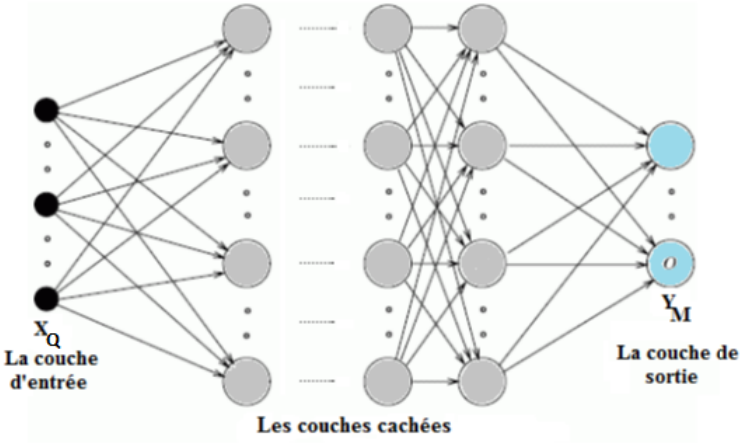


FIGURE 2.4 – *Réseau de neurones multicouches avec Q entrées et M sorties.*

voix convertie. Par la suite, une autre méthode basée sur les caractéristiques spectrales LPC a été introduite (Watanabe et al., 2002), mais cette fois en utilisant un réseau de fonctions à base radiale de trois couches avec une base gaussienne pour transformer les caractéristiques de conduit vocal. Une étude comparative présentée dans (Bandoin et Stylianou, 1996) a montré que les systèmes basés sur les réseaux de neurones étaient plus performants que les systèmes basés sur des modèles GMM. Plus tard, une méthode inspirée des techniques de conversion vocale basées sur JD-GMM, en utilisant des machines de Boltzmann restreintes (RBM) comme modèles de densité de probabilité pour représenter les distributions conjointes de caractéristiques de parole source et cible a été présentée dans (Chen et al., 2013). Le RBM a été initialement introduit en tant que modèle graphique non orienté bipartite qui apprend un modèle généré à partir de données observées. Après l'introduction des RBMs, des machines Gaussienne-Bernoulli (GBRBM) (Cho et al., 2011) ont été proposées.

Le réseau de neurones profond : Le réseau de neurones profond (DNN) est un nouveau domaine de recherche qui a connu une énorme croissance en popularité ces dernières années. Les DNNs utilisent une cascade de couches d'unités de traitement non linéaires pour l'extraction et la transformation des caractéristiques (Deng et al., 2014). Pendant de nombreuses années, la communauté a estimé qu'il était impossible d'entrainer plusieurs réseaux de neurones empilés jusqu'à ce que (Hinton et al., 2006) proposent un algorithme de pré-entraînement efficace appelé "Deep Belief Networks" (DBN) (Hinton et al., 2006). Récemment, les réseaux de neurones profonds (DNNs) ont montré des améliorations de performances dans les domaines de la reconnaissance (Hinton et al., 2012) et de la synthèse de la parole (Ze et al., 2013; Lu et al., 2013; Ling et al., 2013). Des DNNs à quatre couches ont été proposés pour les systèmes de conversion vocale mais aucune différence significative n'a pu être exhibée entre l'utilisation d'un GMM et d'un DNN (Rao et al., 2007). Plus récemment, les DNNs à trois couches ont amélioré leur précision par rapport aux GMMs lorsqu'ils ont été entrainés avec 40 phrases d'ap-

prentissage (Desai et al., 2010). Les deux approches précédentes utilisent des DNNs avec une initialisation aléatoire des poids. Cependant, il a été démontré dans la littérature que les DNNs converge plus rapidement et vers une solution plus performante si leurs paramètres initiaux sont définis via un pré-entraînement au lieu d'une initialisation aléatoire (Erhan et al., 2010). Les méthodes de pré-entraînement utilisent des techniques non supervisées telles que les autoencodeurs (AE) (Vincent et al., 2010; Hinton et Salakhutdinov, 2006). Des DNNs pré-formés ont également été appliqués à la conversion vocale dans une étude récente (Nakashika et al., 2013), dans laquelle des RBMs superposées ont été utilisées pour construire des modèles acoustiques pour chaque locuteur. La plupart des résultats expérimentaux récents avec une architecture profonde sont obtenus avec des modèles qui peuvent être transformés en réseaux de neurones supervisés profonds, mais avec des schémas d'initialisation ou de formation différents des réseaux de neurones classiques "feedforward" (Rumelhart et al., 1986). Des travaux antérieurs (Bengio et al., 2007) ont montré que même une procédure par couches purement supervisées mais profondes fournissait de meilleurs résultats.

Les méthodes basées sur les moindres carrés partiels

Les approches basées sur GMM classiques nécessitent généralement le calcul de matrices de covariance ce qui implique des calculs mathématiques lourds et conduit souvent au problème de surapprentissage (overfitting) en raison du nombre élevé de degrés de liberté disponibles pour la conversion.

Les matrices de covariance complètes fournissent une bonne représentation des données avec un faible nombre de modèles gaussiens, mais nécessitent le calcul d'une grande quantité d'informations. Cependant, une faible quantité de données d'apprentissage est nécessaire pour une mise en œuvre pratique d'un système de conversion vocale.

De plus, il est souhaitable d'avoir un nombre réduit de degrés de liberté au niveau des matrices de covariance pour contrôler le sur-apprentissage des données. D'autre part, le calcul d'une version simplifiée de la matrice de covariance, comme une version diagonale de celle-ci, se traduit par l'utilisation d'un grand nombre de gaussiennes afin d'obtenir une représentation précise.

Pour remédier à ces problèmes, une approche proposée par (Helander et al., 2010) utilise les moindres carrés partiels (PLS), qui traitent spécifiquement de la corrélation croisée entre les prédicteurs et les valeurs prédites. PLS combine les concepts de l'analyse en composantes principales et de la régression linéaire multivariée (MVR), qui peuvent être considérés comme une solution à moyen terme pour le problème de calcul des matrices de covariance.

Cette approche suppose que les données source et cible peuvent être projetées dans un espace latent dans lequel la relation entre les paramètres (source et cible) est établie via une transformation linéaire. En résolvant le modèle de régression obtenu à partir de ces hypothèses, il est possible de convertir les paramètres source en nouveaux paramètres convertis, qui devraient s'approcher des paramètres cible. La sélection appro-

priée de certains des paramètres impliqués contribue à prévenir le sur-apprentissage en réduisant le nombre de degrés de liberté disponible pour le modèle.

Afin d'obtenir une fonction de conversion globale efficace et d'éviter le sur-apprentissage des données, Helander et al. font appel à un GMM pour obtenir la division initiale des données avant d'extraire la fonction de conversion obtenue par la régression partielle des moindres carrés. Ceci vise à réaliser de multiples transformations linéaires locales, en considérant la faible probabilité que les données soient bien modélisées par une seule transformation linéaire.

Les résultats expérimentaux ont confirmé l'efficacité de la méthode basée sur la régression PLS par rapport aux méthodes conventionnelles JD-GMM.

La méthode de conversion de la voix basée sur les PLS que nous venons de présenter convertit chaque trame de parole indépendamment, sans prendre en compte les corrélations temporelles entre les trames et en s'appuyant uniquement sur des transformations linéaires. Pour remédier à ce problème, la conversion de la voix par les moindres carrés dynamiques du noyau (dynamic kernel partial least squares DKPLS) a été proposée dans (Helander et al., 2012).

Cette approche suppose que les paramètres source et cible ont une relation non linéaire. La non-linéarité est implémentée par une fonction du noyau gaussien, utilisée pour projeter les paramètres source dans un espace noyau à haute dimension. La méthode est basée sur une transformation du noyau des paramètres source afin de permettre la modélisation non linéaire et la concaténation des trames précédentes et suivantes dans le but de modéliser la dynamique des données.

De plus, cette mise en œuvre convertit entièrement les caractéristiques spectrales, F0 et les paramètres apériodiques.

Bien que l'efficacité de cette méthode ait été démontrée par la mise en œuvre des DKPLS par rapport aux méthodes fondées sur JD-GMM, elle présente néanmoins certains inconvénients. En premier lieu, la méthode proposée ignore les contraintes temporelles au niveau des paramètres cibles ; deuxièmement, elle manque de souplesse en ce qui concerne les dépendances temporelles à long terme qui ne sont gérées que par l'empilement du noyau des vecteurs source.

2.5 Étapes de construction du système de conversion de voix proposé

L'algorithme proposé est réalisé en deux phases, la phase d'apprentissage et la phase de transformation, (voir la figure 2.5). Les étapes d'analyse de la parole transforment les signaux de la voix source et cible en une représentation cepstrale. Après cela, l'ensemble des vecteurs cepstraux de la voix source et cible sont alignés avec la méthode DTW pour associer les segments qui ont un contenu phonétique similaire. Ces entités alignées sont ensuite utilisées pour estimer la fonction de transformation. Cette fonction

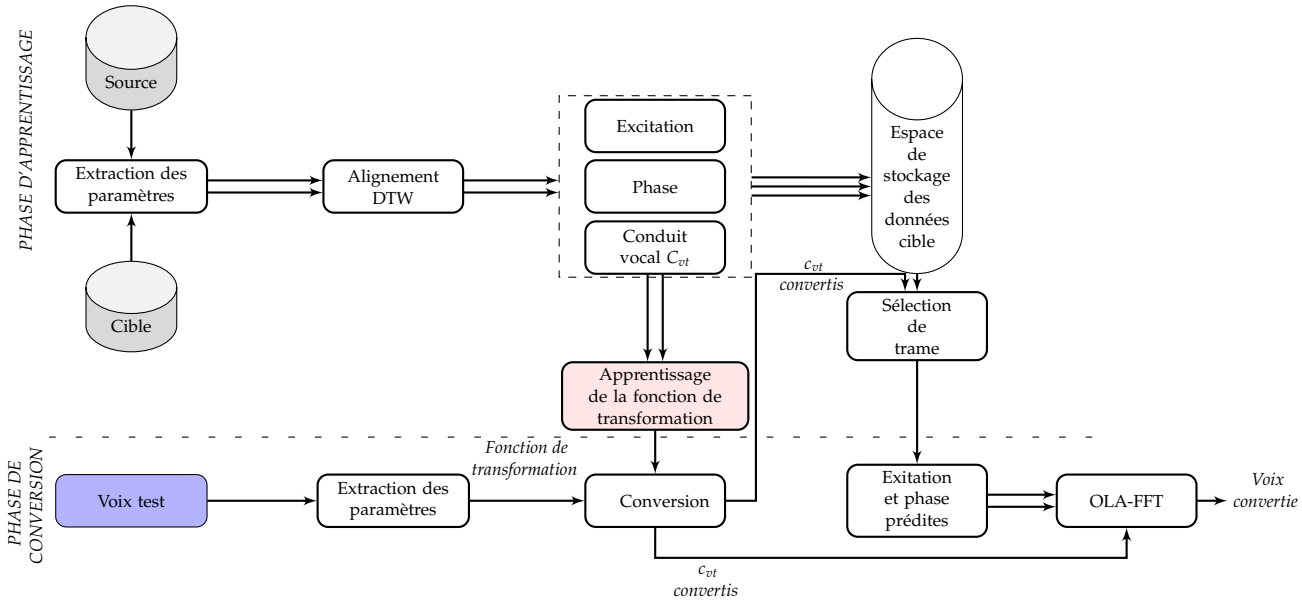


FIGURE 2.5 – *Principales étapes du système de conversion de voix proposé .*

de transformation est utilisée par la suite dans la phase de conversion pour convertir de nouveaux signaux source (appelés phrases test). Les paramètres convertis sont ensuite utilisés pour extraire l'excitation cepstrale et la phase à partir de l'espace d'apprentissage cible afin de synthétiser la forme d'onde de la voix convertie. Dans cette section, nous décrivons chaque étape de notre approche.

2.5.1 Analyse et extraction des vecteurs acoustiques

La phase d'analyse joue un rôle important dans notre système de conversion vocale. L'objet de ce processus est d'extraire les vecteurs acoustiques de la voix source et cible. Initialement, deux corpus parallèles sont disponibles, le premier du locuteur source et le second du locuteur cible. L'extraction des paramètres acoustiques des deux voix source et cible a été réalisée par une analyse cepstrale (Oppenheim et Schafer, 1968) pour ses avantages.

En effet les cepstres sont des quantités logarithmiques impulsionnelles dans le temps donc : ils sont moyennables ; ils peuvent faire l'objet d'une distance euclidienne ou d'une classification vectorielle ; ils sont utilisés en conversion vocale pour la prédiction de la fréquence F0, etc... De plus en utilisant les cepstres, il est facile de revenir dans le domaine fréquentiel.

Le signal de parole est échantillonné à 16 kHz. Par la suite l'extraction des vecteurs cepstraux (d'excitation et du conduit vocal) est réalisé par les étapes suivantes : tout d'abord, une fenêtre de Hamming normalisée $H(n)$ (Griffin et Lim, 1984) (voir la formule 2.11) de longueur 512 est utilisée pour obtenir les signaux temporels à court terme

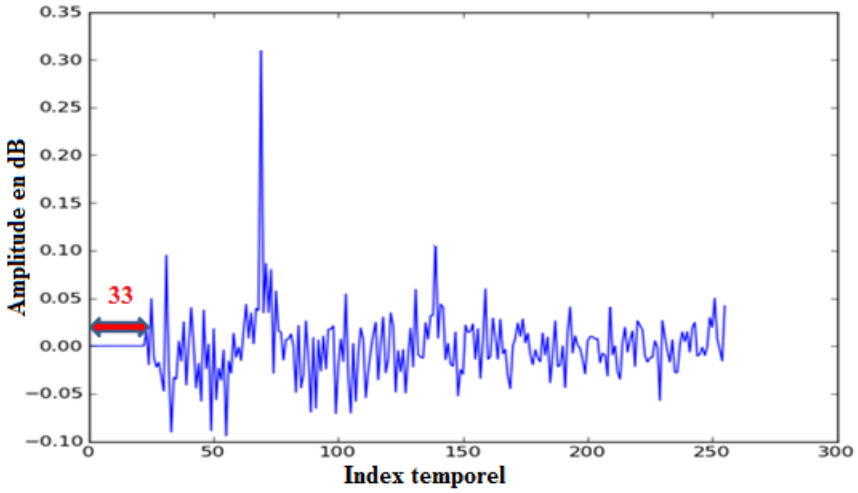


FIGURE 2.6 – *Illustration d'un signal cepstral*

à partir desquelles les coefficients cepstraux seront extraits :

$$H(n) = \begin{cases} \frac{2\sqrt{\frac{L}{N}}}{\sqrt{4a^2+2b^2}}(a + b.cos(\frac{\pi(2n+1)}{N}), 0 \leq n < N \\ 0, otherwise \end{cases} \tag{2.11}$$

où N est la longueur en échantillons de la fenêtre d'analyse, L est le décalage temporel en échantillons entre deux trames consécutives analysées, a = 0.54 et b = -0.46. Par la suite une transformation de Fourier rapide (FFT) est appliquée sur la trame analysée, suivie par le calcul du module sur ce signal. Le spectre logarithmique de Fourier est obtenu en calculant le logarithme sur ce nouveau signal.

La transformée de Fourier rapide inverse (IFFT) du spectre logarithmique, permet de trouver le cepstre logarithmique réel lié à la trame analysée. En mettant les premiers coefficients à zéro nous obtenons le cepstre d'excitation (voir la figure 2.6). Dans notre étude, les premiers coefficients représentent le cepstre du conduit vocal constitué par les 33 premiers coefficients.

Le cepstre réel, n'utilise que le module du spectre du signal, il perd donc la partie de l'information contenue dans la phase. Nous ne pouvons pas reconstruire parfaitement le signal de départ à partir du cepstre. Pour cela nous avons déterminé la phase. Le spectre de phase a été déterminé par l'équation 2.12

$$Phase = tan^{-1}(\frac{S_{im}}{S_{re}}) \tag{2.12}$$

Où S_{im} et S_{re} sont respectivement la partie imaginaire et réelle du spectre de Fourier. Enfin, le contenu linguistique du signal de parole est codé par l'ensemble de coefficients suivants :

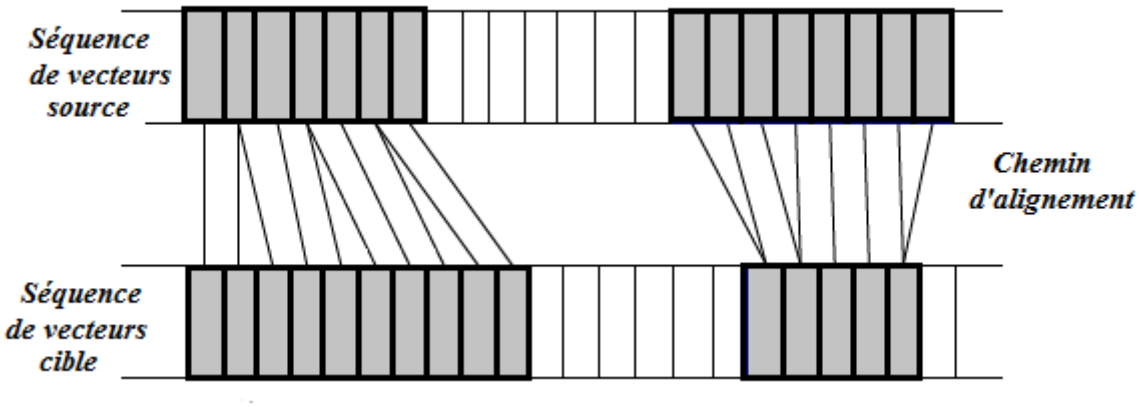


FIGURE 2.7 – *Alignement temporel par la programmation dynamique (DTW) entre les vecteurs source et cible.*

— Le premier coefficient cepstral c_0
— Le vecteur cepstral relatif au conduit vocal $C_{vt} = [c_1...c_P]$
— Le vecteur cepstral relatif au signal d'excitation $C_{exc} = [c_{P+1}...c_{N/2}]$
— La phase $Ph = [ph_0...ph_{N/2}]$

Pour convertir la voix, il est indispensable de trouver des méthodes de transformation pour les vecteurs cepstraux relatifs au conduit vocal, à l'excitation et à la phase.

2.5.2 Alignement multi-passes

L'alignement est réalisé par l'algorithme DTW (Dynamic Time Warping). Cet alignement permet d'effectuer une synchronisation des échelles de temps de deux séquences de vecteur source C_Q^S et cible C_M^C afin de trouver un chemin optimal.
Ce chemin construit une correspondance entre les deux séquences de vecteurs source et cible (Di Martino, 1984) (voir figure 2.7). Le résultat de cette phase est une liste de vecteurs appariés.

La méthode d'alignement DTW suppose qu'un même phonème prononcé par deux locuteurs présente des caractéristiques similaires. Cependant, cette hypothèse n'est pas toujours vraie et pourrait produire des alignements sous-optimaux. Soit X=[x_1,x_2,...,x_U] la séquence de trames de la voix source et Y=[y_1,y_2,...,y_V] est la séquence de trames de la voix cible.

Nous proposons dans la première itération, d'appliquer l'alignement DTW entre les vecteurs source X_U et cible Y_V.

Nous voulons obtenir le chemin d'alignement (A, B), où A=[a_1,a_2,...,a_W] et B=[b_1,b_2,...,b_W] sont des séquences d'indices utilisées pour faire correspondre la séquence source initiale X avec celle de la cible initiale Y. Ensuite, les vecteurs appariés X=[X_{a_1},X_{a_2},...,X_{a_W}] et Y=[Y_{b_1},Y_{b_2},...,Y_{b_W}] sont utilisés pour calculer la fonction de transformation. Une fois la

fonction de transformation établie, elle est utilisée pour transformer les vecteurs source en vecteurs convertis Z.

À partir de la seconde itération, l'alignement est réalisé entre les vecteurs convertis Z et les vecteurs cible Y_V En appliquant DTW sur la séquence convertie Z et la séquence cible Y, nous obtenons un nouvel alignement.

Pour affiner la liste de correspondances, le processus d'alignement et de conversion est répété jusqu'à ce que le nombre total d'itérations soit atteint (voir la figure 2.8).

Les fonctions de transformation obtenues à cette étape de pré-traitement ne sont pas toutes utilisées. Seul le meilleur alignement (le dernier) est conservé pour la suite du processus.

2.5.3 Apprentissage

Dans cette étape, les séquences alignées sont utilisées pour calculer une fonction de transformation $y = F(x)$. Seuls les P premiers coefficients cepstraux de signal vocal source $x_i = [x_i^1 ... x_i^P]$ et les P premiers coefficients cepstraux de signal vocal cible $y_i = [y_i^1 ... y_i^P]$ sont utilisés pour calculer la fonction de transformation F. Deux approches ont été utilisées dans ce travail : une méthode à base de mélange gaussien (GMM) et deux modèles de transformation basés sur la mise en cascade de modèles de mélange gaussien GMM et d'un réseau neuronal profond (DNN).

Amélioration des performances computationnelles des systèmes de conversion vocale basés sur JD-GMM

Pour déterminer la fonction de transformation nous avons implémenté le modèle de mélange gaussien conjoint décrit dans la section 2.4.3 (Ben Othmane et al., 2018c). Afin de déterminer les paramètres GMM initiaux, une classification est effectuée via l'algorithme de classification des K-moyennes (en anglais K-means). L'algorithme des K-moyennes a été utilisé pour sa simplicité, sa rapidité d'exécution et par le fait qu'il permet de déterminer un nombre quelconque de classes contrairement aux autres algorithmes de classification.

Après la classification initiale, nous proposons de partitionner chaque classe en M sous-classes en appliquant à chaque classe l'algorithme de classification des K-moyennes.

Pour chaque classe, une fonction de transformation peut être déterminée. Le choix de la fonction locale F_{i^*} qui doit être appliquée à un vecteur source x est déterminé par la probabilité conditionnelle maximale que x appartienne à la classe i^*.

$$i^* = argmax_i P(C_i, x) \tag{2.13}$$

où

$$P(C_i, x) = \sum_k \alpha_k^i N_k^i(x, \mu_k^i, \Sigma_k^i) \tag{2.14}$$

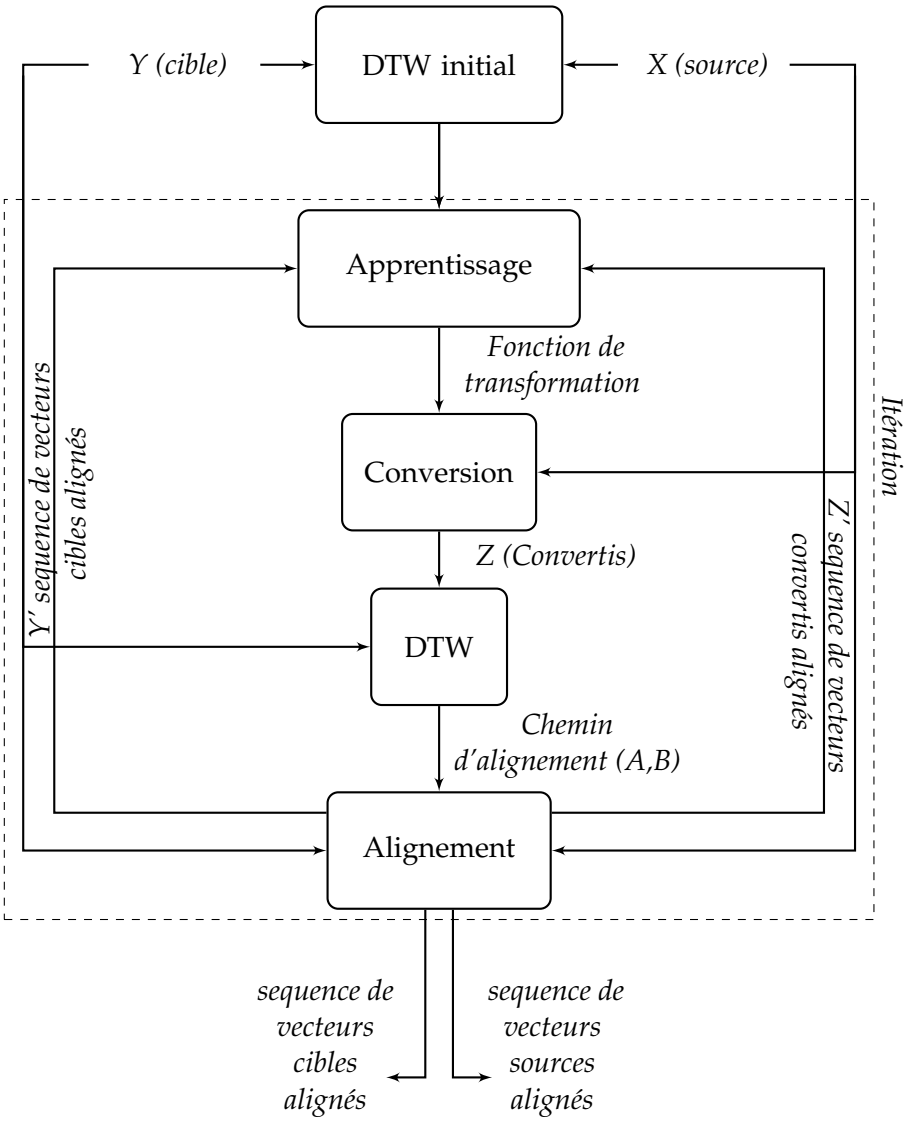


FIGURE 2.8 – *Schéma fonctionnel de l'alignement obtenu par plusieurs passes.*

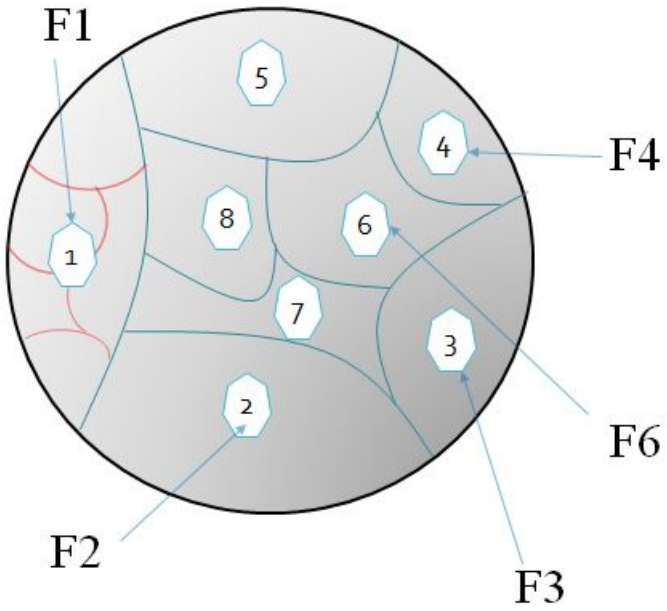


FIGURE 2.9 – *Classes et sous-classes obtenues par classification vectorielle*

Σ_k^i et α_k^i désignent respectivement la $k^{\text{ème}}$ matrice de covariance de la $k^{\text{ème}}$ distribution normale N_k^i et le $k^{\text{ème}}$ poids du mélange gaussien de la classe i.

La fonction de conversion $F(x) = F_{i^*}(x)$ est utilisée pour calculer la fonction de transformation locale (GMM-LMF : Local Mapping function) :

$$F_{i^*}(x) = \sum_{\hat{\imath}=1}^{M} p(\hat{\imath}|x)(\mu_{\hat{\imath}}^y + \Sigma_{\hat{\imath}}^{yx}(\Sigma_{\hat{\imath}}^{xx})^{-1}(x - \mu_{\hat{\imath}}^x)) \tag{2.15}$$

où $p(\hat{\imath}|x)$ représente la probabilité postérieure que x soit généré par le composant $\hat{\imath}$:

$$p(\hat{\imath}|x) = \frac{\alpha_{\hat{\imath}}^{i^*} N_{\hat{\imath}}^{i^*}(x, \mu_{\hat{\imath}}^{i^*}, \Sigma_{\hat{\imath}}^{i^*})}{\sum_{k=1}^{M} \alpha_k^{i^*} N_k^{i^*}(x, \mu_k^{i^*}, \Sigma_k^{i^*})} \tag{2.16}$$

M désigne le nombre des sous-classes ; $\mu_{\hat{\imath}}^x$ and $\mu_{\hat{\imath}}^y$ représentent respectivement les vecteurs moyens source et cible de la sous-classe $\hat{\imath}$ de classe i^*.

Cette méthode permet de trouver une fonction de transformation plus précise tout en réduisant le temps de calcul.

Pour cette méthode, nous proposons dans la première itération d'appliquer un alignement entre les vecteurs cepstraux source X_U et cible Y_V. À partir de la seconde itération, l'alignement est effectué entre les vecteurs convertis Z et les vecteurs cibles Y_V afin d'affiner le chemin d'alignement temporel.

Réseaux de neurones profonds

Les réseaux de neurones profonds (DNN) sont également utilisés pour modéliser la relation $y = F(x)$, ce qui permet de transformer les caractéristiques de la voix d'un locuteur source en celles d'un locuteur cible.

Le réseaux reposant sur l'apprentissage profond est la méthode la plus couramment utilisée dans le domaine de l'apprentissage automatique et a accumulé les succès ces derniers temps. Les vecteurs cepstraux de la voix source et de la voix cible sont extraits et présentés respectivement en tant que vecteurs d'entrée et de sortie.

Les performances du modèle DNN dépendent de l'architecture du réseau, des nœuds, des poids, du biais et des fonctions d'activation associées : tous ces paramètres sont adaptés en fonction de la tâche de conversion vocale afin d'assurer les meilleures performances du réseau. Les modèles DNN sont appliqués dans nos systèmes de conversion vocale pour leur efficacité.

Les réseaux neuronaux profonds avec un grand nombre de paramètres sont très performants. Cependant, certains problèmes d'optimisation sont difficiles à résoudre. De plus, le sur-apprentissage est un problème important. Étant donné que les grands réseaux nécessitent un temps de calcul important pour l'apprentissage, il est difficile de faire face à ce problème.

Le décrochage (dropout) est une technique de régularisation et résout actuellement ce problème. Le décrochage a été proposé par Srivastava et al. (Srivastava et al., 2014).

Cette technique de régularisation, évite des co-adaptations complexes sur les données d'apprentissages. C'est un moyen très efficace pour réaliser un moyennage du modèle de calcul avec des réseaux de neurones profonds.

Le décrochage consiste à supprimer de manière temporaire et aléatoire des neurones (les neurones cachés et visibles) dans un réseau de neurones profonds (voir la figure 2.10).

Un réseau de neurones de type "feedforward" profond avec L couches cachées peut être décrit comme :

$$\begin{cases} z_i^{(l+1)} = W_i^{(l+1)} y^l + b_i^{(l+1)}, l \in 1, ..., L \\ y_i^{(l+1)} = f(z_i^{(l+1)}) \end{cases}$$

$z^{(l)}$ est le vecteur d'entrée dans la couche l et $y^{(l)}$ le vecteur de sortie, $W^{(l+1)}$ and $b^{(l)}$ représentent respectivement les poids et les biais de la couche l. f est la fonction d'activation.

Nous proposons d'utiliser l'unité linéaire rectifiée ReLU (Nair et Hinton, 2010) comme fonction d'activation pour sa simplicité et ses bonnes performances. En cas de décro-

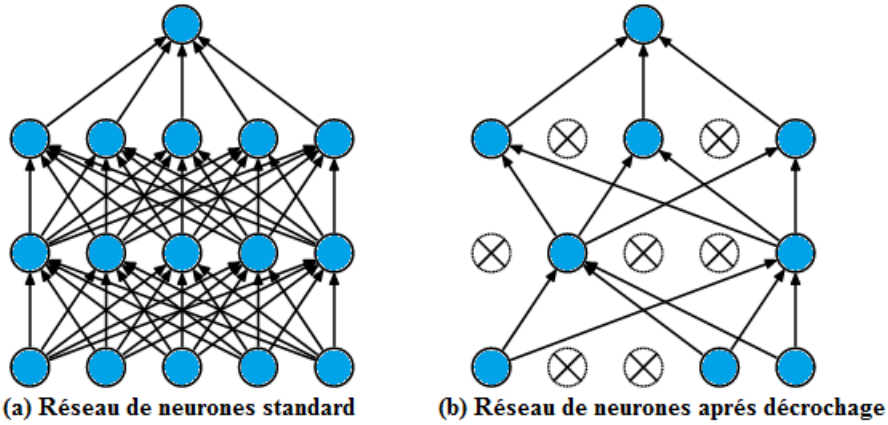


FIGURE 2.10 – *Modèle de réseau de neurones décroché. À gauche : un réseau de neurones standard. À droite : exemple de réseau réduit obtenu en appliquant un décrochage sur le réseau de gauche.*

chage, la fonction "feedforward" devient :

$$r_j^{(l)} \simeq Bernoulli(p) \tag{2.17}$$

$$\tilde{y}^{(l)} = r^l * y^l \tag{2.18}$$

$$y_i^{(l+1)} = f(z_i^{(l+1)}) \tag{2.19}$$

$$z_i^{(l+1)} = W_i^{(l+1)} \tilde{y}^{(l)} + b_i^{(l+1)} \tag{2.20}$$

où $r_j^{(l)}$ est appelé un vecteur de variables aléatoires de Bernoulli, dont chacune a la probabilité p d'être égale à 1. Pour chaque couche, $r_j^{(l)}$ est échantillonné, puis multiplié élément par élément avec les sorties $y^{(l)}$, afin de créer les sorties amincies $\tilde{y}^{(l)}$ (Srivastava et al., 2014).

Le décrochage peut améliorer considérablement les performances du processus d'apprentissage. Après la phase d'apprentissage, le modèle DNN est utilisé pour convertir un vecteur d'entrée source en un vecteur cible. Ces paramètres de sortie sont utilisés pour créer le signal vocal converti.

Modèles de transformation basés sur la mise en cascade de deux modèles GMM et DNN

L'idée de la mise en cascade de deux modèles DNN et GMM permet d'améliorer les vecteurs cepstraux convertis. Cette nouvelle technique améliore la conversion vocale. Nous sommes parvenus à trouver les combinaisons des deux fonctions de transformation (GMM et DNN) les plus performantes, tout en évitant le problème de surapprentissage.

Modèle DNN-GMM

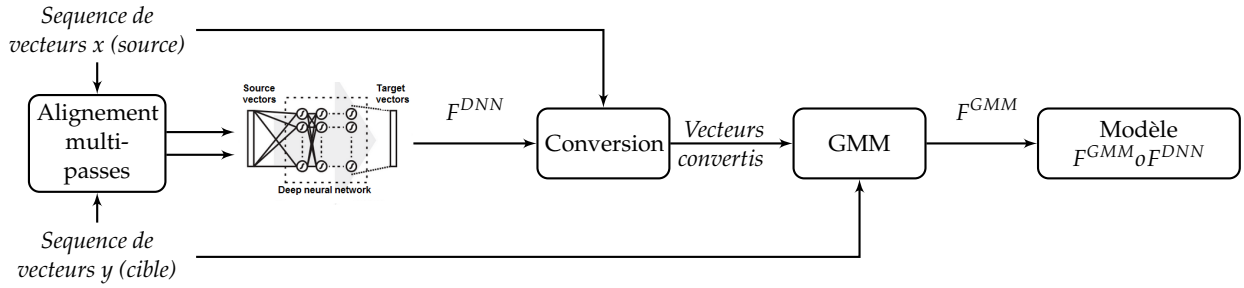


FIGURE 2.11 – *Phase d'apprentissage pour le modèle DNN-GMM.*

Dans cette section, nous proposons de mettre en cascade les fonctions de transformation DNN et GMM afin d'estimer la relation entre les vecteurs cepstraux source et cible, comme indiqué dans la figure 2.11.

Les principales étapes basées sur la mise en cascade d'un réseau de neurones profond et d'un modèle de mélange gaussien sont décrites comme suit :

1. Entraîner le DNN pour obtenir les vecteurs cepstraux du conduit vocal convertis z.
2. Convertir les vecteurs cepstraux du conduit vocal source en utilisant la fonction de conversion basée sur le DNN pour chaque vecteur source x :

$$z = F^{DNN}(x) \tag{2.21}$$

3. Effectuer l'apprentissage de modèle GMM avec les vecteurs cepstraux du conduit vocal convertis z et les vecteurs cepstraux du conduit vocal cible y.

4. Convertir les vecteurs cepstraux du conduit vocal convertis en utilisant la fonction de conversion basée sur les GMM donnée par la formule 2.15 pour chaque vecteur converti z :

$$F^{GMM}(z) = \sum_{k=1}^{m} p(k|z)(\mu_k^y + \Sigma_k^{yz}(\Sigma_k^{zz})^{-1}(z - \mu_k^z)) \tag{2.22}$$

Pour cette méthode, nous proposons dans la première itération d'appliquer un alignement entre les vecteurs cepstraux source X_U et cible Y_V. À partir de la seconde itération, l'alignement est effectué entre les vecteurs convertis Z avec le modèle DNN-GMM et les vecteurs cibles Y_V afin d'affiner le chemin d'alignement temporel.

Modèle GMM-DNN-GMM

Cette section présente une description générale du modèle GMM-DNN-GMM représenté dans la figure 2.12. Nous procédons de la même manière que lors de la construction du modèle DNN-GMM.

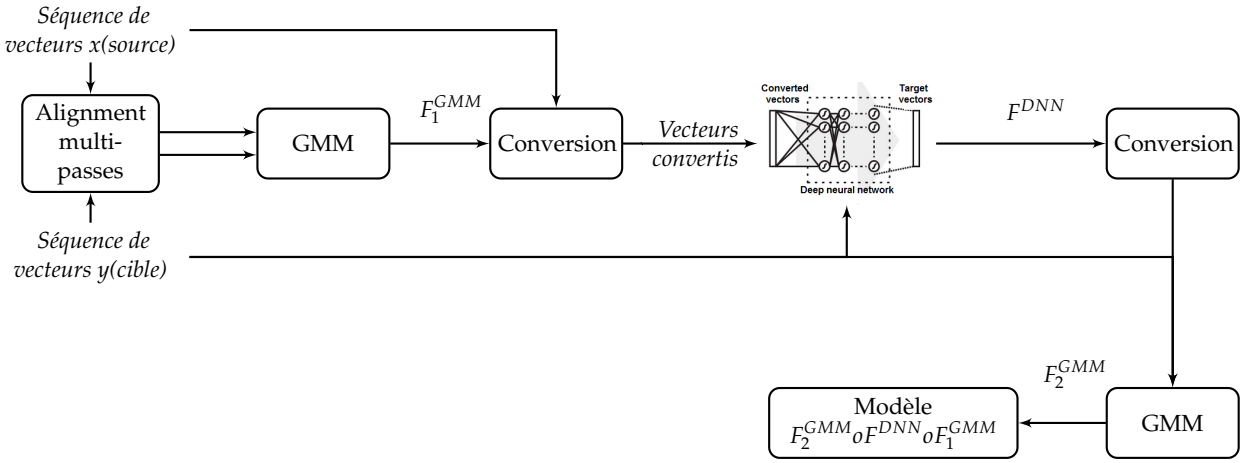


FIGURE 2.12 – *Phase d'apprentissage pour le modèle GMM-DNN-GMM*

1. Après l'alignement, un GMM est utilisé pour modéliser la fonction de transformation $x' = F_1^{GMM}(x)$ entre les vecteurs cepstraux source et cible.
2. Après cela, la fonction de conversion basée sur GMM et donnée par la formule 2.15 est utilisée pour convertir chaque vecteur cepstral de la voix sources x :

$$F_1^{GMM}(x) = \sum_{k=1}^{m} p(k|x)(\mu_k^y + \Sigma_k^{yx}(\Sigma_k^{xx})^{-1}(x - \mu_k^x)) \tag{2.23}$$

3. Le modèle DNN est entraîné avec les vecteurs cepstraux convertis et cible pour prédire le modèle F^{DNN}.

$$z = F^{DNN}(x') = F^{DNN}(F_1^{GMM}(x)) \tag{2.24}$$

4. Après cela, F^{DNN} est utilisé pour transformer la séquence de vecteurs cepstraux x'.
   Ensuite, en utilisant GMM, nous obtenons une nouvelle fonction de transformation :

$$y = F_2^{GMM}(z) = F_2^{GMM}(F^{DNN}(x')) \tag{2.25}$$

   entre les vecteurs cepstraux convertis $z = F^{DNN}(x')$ et ceux de la cible y.

$$F_2^{GMM}(z) = \sum_{k=1}^{m} p(k|z)(\mu_k^y + \Sigma_k^{yz}(\Sigma_k^{zz})^{-1}(z - \mu_k^z)) \tag{2.26}$$

Enfin, la fonction de transformation devient :

$$y = F_2^{GMM} o F^{DNN} o F_1^{GMM}(x) \tag{2.27}$$

Pour affiner le chemin d'alignement temporel, dans la première itération, un alignement est appliqué entre les vecteurs cepstraux source X_U et cible Y_V. À partir de la seconde itération, l'alignement est effectué entre les vecteurs convertis Z (avec le modèle GMM-DNN) et les vecteurs cible Y_V .

2.5.4 Conversion

Après avoir défini les fonctions de transformation, nous abordons dans la suite la phase de conversion. Les coefficients cepstraux sont extraits du signal vocal source. Ensuite, seul les P premiers coefficients $x_i = [x_i^1 ... x_i^P]$ sont convertis par les méthodes décrites précédemment (GMM, DNN, DNN-GMM et GMM-DNN-GMM).

2.5.5 Prédiction de l'excitation cepstrale et des coefficients de phase

Pour estimer l'excitation, les approches classiques prédisent la fréquence F0 et génèrent une onde excitative périodique dans le cas où la trame temporelle est voisée et une onde de bruit dans le cas où la trame temporelle est non voisée. Cette onde excitative excite un modèle autorégressif calculé à partir du vecteur cepstral transformé. Cette approche présente plusieurs inconvénients. En effet la prédiction de la fréquence fondamentale F0 et la prédiction du voisement ne peuvent pas être parfaites et l'onde excitative temporelle générée n'est pas réaliste car son spectre ne correspond pas généralement à un spectre excitatif réel.

Pour générer une excitation réaliste, nous supposons qu'il existe une forte corrélation entre le vecteur cepstral du conduit vocal et le vecteur cepstral excitatif. Pour cela nous avons proposé d'extraire l'excitation à partir de l'espace d'apprentissage cible. Nous proposons également d'extraire les coefficients de phase à partir de l'espace d'apprentissage cible afin d'obtenir une voix convertie plus naturelle. Le processus de prédiction de l'excitation cepstrale et des coefficients de phase est illustré dans la figure 2.13. Tout d'abord, une étape de concaténation est appliquée aux paramètres suivants :

— Les vecteurs cepstraux du conduit vocal cible.
— Les vecteurs cepstraux du conduit vocal convertis.
— L'excitation cepstrale.
— La phase.

Chaque NB_{frame} paquets sont concaténés afin de former une trame. Ensuite, les trames cible sont codées sous la forme d'un arbre binaire KD-tree (Arya, 1996). Par la suite, seuls les vecteurs cepstraux convertis relatifs au conduit vocal sont utilisés pour estimer l'excitation et la phase.

L'arbre binaire est interrogé par les trames converties pour une recherche rapide du plus proche voisin. L'arbre nous fournit en sortie un indexe I, qui peut être utilisé pour trouver rapidement les trames les plus proches de chaque trame transformée.

Cet index I est ensuite utilisé pour trouver l'excitation cepstrale et la phase associée à la trame convertie.

Ensuite, chaque trame convertie est concaténée avec son excitation. Dans ce travail, la prédiction de l'excitation et de la phase avec un arbre KD-tree nous a permis de générer des spectres réalistes.

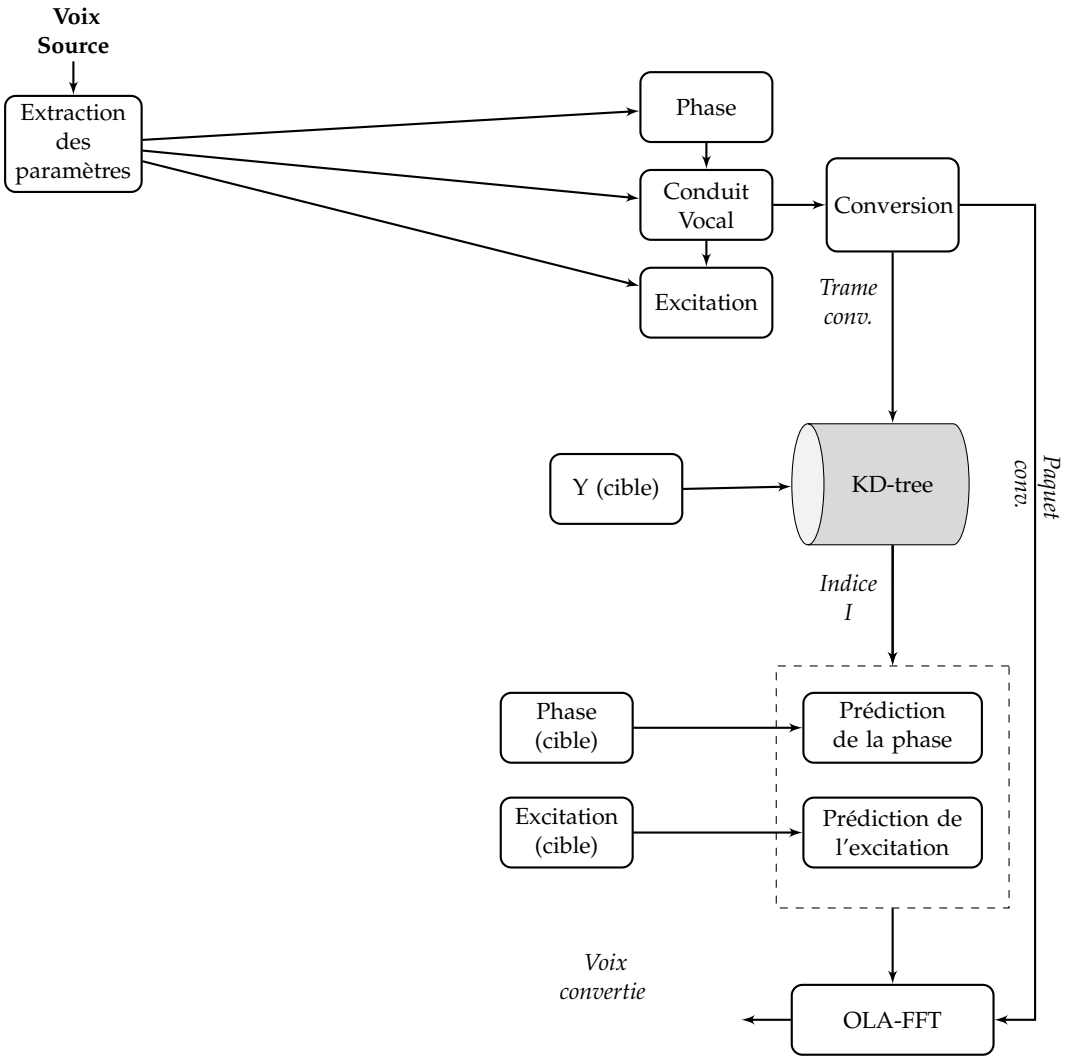


FIGURE 2.13 – *Schéma fonctionnel relatif à l'extraction de l'excitation et de la phase.*

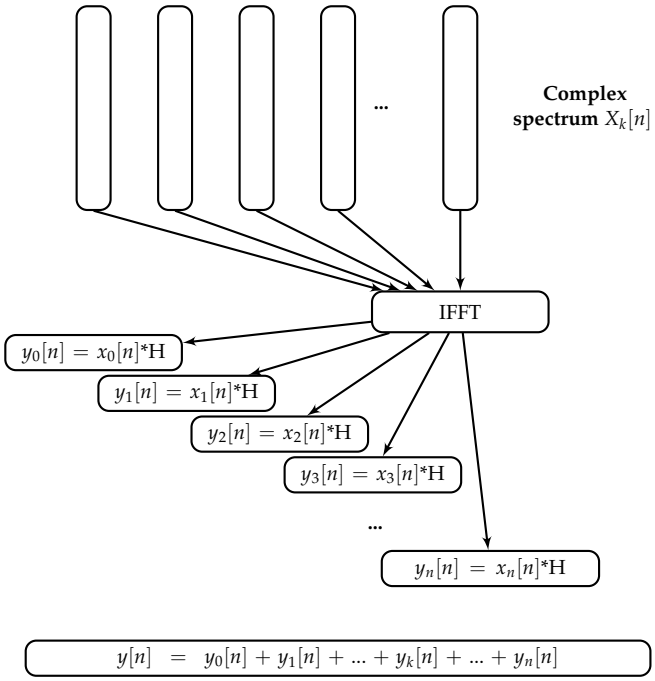


FIGURE 2.14 – *OLA-FFT*

2.5.6 Synthèse vocale

Il existe différentes méthodes pour reconstruire le signal de parole. Le signal vocal peut-être synthétisé à partir des spectres d'amplitude en utilisant le synthétiseur Griffin & Lim (Griffin et Lim, 1984), RTISI (Beauregard et al., 2005), RTISI-LA (Zhu et al., 2006), etc. Puisque la phase, dans notre étude, a été estimée, les spectres d'amplitude et de phase sont utilisés pour créer des spectres complexes. La voix convertie peut-être alors reconstruite par la méthode addition-recouvrement à court terme OLA-FFT.
Cette technique de synthèse consiste à appliquer une transformée de Fourier rapide inverse IFFT aux spectres complexes, voir la figure 2.14.

2.6 Résultats expérimentaux

Dans ce travail, nous utilisons la base de données CMU ARCTIC (Kominek et Black, 2004). Cette base de données est composée des phrases enregistrées par sept locuteurs. Chaque locuteur a enregistré un ensemble de 1132 phrases phonétiquement équilibrées. Cette base de données comprend les énoncés des locuteurs suivants : SLT(femme américaine), CLB (femme américaine), BDL (homme américain), RMS (homme américain), JMK (homme canadien), AWB (homme écossais), KSP (homme indien). Seuls les énoncés des locuteurs américains (2 locuteurs masculins et 2 locuteurs féminins) ont été pris en compte dans cette étude. Pour la phase d'apprentissage, nous utilisons des corpus parallèles, en utilisant le même ensemble de phrases prononcées par les locu-

teurs source et cible. Les fichiers audio ".wav" ont été obtenus à l'aide d'une fréquence d'échantillonnage de 16 kHz avec chaque échantillon codé sur 16 bits. Le tableau 2.1 résume les conditions expérimentales : La structure des expériences est la suivante.

TABLE 2.1 – *Conditions experimentales.*

Nombre de phrases d'apprentissage	100/locuteur
Nombre de phrases de test	20/locuteur
Fenêtre d'analyse	Hamming normalisée
Taille de la fenêtre d'analyse	512
Durée de la fenêtre	32 ms
Nombre de classes pour le 1^{er} classifieur	8
Nombre de sous-classes pour le $2^{ème}$ classifieur	8
NB_{frame}	20
Nombre des coefficients cepstraux du conduit vocal	32
Nombre des coefficients cepstraux d'excitation	224

Dans le premier ensemble d'expériences, nous avons évalué la performance de modèle proposé GMM-LMF. Dans le deuxième ensemble d'expériences, nous avons comparé l'effet de l'alignement. Pour l'alignement multi-passes, nous avons utilisé les deux modèles de transformation DNN-GMM et GMM-DNN.

Après avoir déterminé le meilleur alignement, nous avons utilisé quatre modèles d'apprentissage pour estimer la fonction de conversion des vecteurs cepstraux du conduit vocal.

- Les systèmes de référence :
  1. Conversion vocale basée sur un réseau de neurones profond.
  2. Conversion de voix basée sur un modèle de mélange gaussien.
- Systèmes proposés : Nos deux approches proposées sont :

  1. DNN-GMM : la fonction de transformation est déterminée en combinant un réseau de neurones profonds suivi d'un modèle de mélange gaussien. Cette fonction de transformation est utilisée dans DTW itératif et dans la conversion.

  2. (GMM-DNN)-GMM : une combinaison de GMM et de DNN est utilisée pour l'alignement multi-passes. Nous utilisons ensuite cet alignement pour calculer la dernière fonction de transformation GMM.

2.6.1 Évaluation objective

Pour évaluer les performances des systèmes proposés, nous utilisons le rapport signal sur erreur (SER) entre les spectres cible et convertis. Celui-ci est estimé par la for-

mule suivante 2.28 :

$$SER = -10 * log_{10} \frac{\sum_k ||y_k - \hat{y}_k||^2}{\sum_k ||y_k||^2} \tag{2.28}$$

où y_k et $\hat{y}_k$ sont respectivement les vecteurs cepstraux cibles et convertis.
Nous avons aussi calculé la distance cepstrale CD. Nous avons donc calculé la distance cepstrale entre les trames converties et cibles :

$$CD(\hat{y}, y)\ [dB] = \frac{10}{log_{10}} \sqrt{\sum_k (c_k(\hat{y}) - c_k(y))^2} \tag{2.29}$$

Où $c_k(\hat{y})$ et $c_k(y)$ représentent respectivement les $k^{\text{ème}}$ coefficients cepstraux du vecteur converti et du vecteur cible.

Évaluation de la méthode GMM-LMF

Pour évaluer la performance de la fonction locale proposée GMM-LMF, nous avons effectué une comparaison entre la méthode classique GMM-GMF (avec une fonction de transformation globale) et la méthode proposée GMM-LMF (avec une fonction de transformation locale) les résultats sont présentés dans le tableau 2.2. Nous pouvons ob-

TABLE 2.2 – *SER*

Test Objectif	GMM-GMF	GMM-LMF
F⇒H	14.13	14.21
H⇒F	12.38	12.89
F⇒F	13.18	13.38
H⇒H	14.08	14.27

server que la fonction de transformation locale est plus performante que la fonction de transformation globale. Il est à noter que les deux méthodes utilisent la même stratégie pour l'extraction de l'excitation et de la phase. Le tableau 2.3 présente une comparaison entre les temps de calcul du processus d'apprentissage pour GMM-GMF et GMM-LMF. Il est clair que le temps de calcul de la phase d'apprentissage a été considérablement réduit par rapport à la méthode de référence GMM-GMF, pour tous les types de conversion. Le tableau 2.4 indique le temps de calcul moyen de la phase de conversion pour

TABLE 2.3 – *Durées d'apprentissage des fonctions de la transformation globale et locale, respectivement GMM-GMF et GMM-LMF*

Objective test	GMM-GMF	GMM-LMF
F⇒H	28.87	15.81
H⇒F	48.71	14.91
F⇒F	35.78	19.83
H⇒H	53.24	19.11

chaque trame. Après avoir déterminé le temps de calcul de la phase de conversion pour

les deux méthodes, nous pouvons constater que la méthode proposée réduit également le temps de calcul de la phase de conversion. Selon les résultats obtenus, nous pouvons noter que la méthode proposée améliore la précision de la conversion tout en réduisant les temps de calcul pour la phase d'apprentissage et de transformation.

TABLE 2.4 – *Temps de calcul (ms) pour la conversion de vecteurs cepstraux du conduit vocal à l'aide de GMM-GMF et GMM-LMF*

Objective test	GMM-GMF	GMM-LMF
F⇒H	1.64	0.96
H⇒F	1.65	0.91
F⇒F	1.61	0.89
H⇒H	1.78	1.12

Évaluation des méthodes DNN-GMM et GMM-DNN-GMM

D'abord, nous comparons l'effet de l'alignement. Dans l'alignement multi-passes, 10 itérations sont calculées. Nous avons remarqué qu'après 3 alignements, aucune amélioration n'est constatée.
D'après le tableau 2.5, l'alignement itératif est plus performant que lorsque une seule

TABLE 2.5 – *DTW vs DTW multi-passes.*

Méthodes	DTW		DTW multi-passes	
	apprentissage	test	apprentissage	test
DNN-GMM	16.20	13.54	**16.94**	13.61
GMM-DNN	15.74	13.49	**16.83**	13.58

itération DTW est réalisée, quelle que soit la fonction de transformation utilisée. Le meilleur résultat avec un seul DTW permet d'obtenir un SER égal à 16.20. Mais, lorsque l'alignement itératif est utilisé, le meilleur résultat permet d'obtenir est un SER égal à 16.94.

TABLE 2.6 – *Performances objectives des approches proposées DNN-GMM et GMM-DNN-GMM par rapport aux approches classiques à base de méthodes GMM et DNN (CD).*

Type de conversion	F1 à H1	H1 à F2
GMM	5.71	5.14
DNN	5.36	4.73
GMM-DNN-GMM	5.03	4.56
DNN-GMM	4.30	4.41

Dans la construction d'un système de conversion vocale basé sur DNN, la tâche la plus importante est de trouver une architecture du DNN optimale. Nous avons calculé les SER, avec différentes fonctions d'activation et avec différents nombres de neurones dans les couches cachées. La comparaison est présentée dans le tableau 2.7, deux parmi

cinq architectures de DNN avec la fonction d'activation ReLU offrent les meilleures performances et les meilleurs SER (SER=16.73 et 16.63). Le tableau 2.8 présente les

TABLE 2.7 – *Effet du choix de la fonction d'activation*

Hidden	ReLu		Tanh	
	Apprentissage	Test	Apprentissage	Test
256,256	16.27	14.25	14.22	14.22
512,512	**16.63**	14.27	14.23	14.02
64,256,256,64	16.00	14.26	14.27	14.22
512,256,256,512	**16.73**	14.26	14.34	14.23
512,512,512,512	16.58	14.09	14.35	14.34

TABLE 2.8 – *Performances objectives des approches proposées DNN-GMM et GMM-DNN-GMM par rapport aux approches classiques à base de méthodes GMM et DNN (SER).*

Type de conversion	Méthode de conversion	SER	
		Apprentissage	Test
F1 à H1	GMM	15.05	14.43
	DNN	15.67	14.18
	DNN-GMM	**17.61**	14.48
	GMM-DNN-GMM	**16.56**	14.49
H1 à F2	GMM	14.58	13.21
	DNN	15.27	12.08
	DNN-GMM	**16.94**	13.61
	GMM-DNN-GMM	**16.83**	13.58
F1 à F2	GMM	14.87	13.78
	DNN	15.75	12.81
	DNN-GMM	**16.71**	14.20
	GMM-DNN-GMM	**15.93**	13.69
H1 à H2	GMM	15.14	14.58
	DNN	16.59	14.28
	DNN-GMM	**17.25**	14.69
	GMM-DNN-GMM	**16.40**	14.70

valeurs de SER pour les méthodes proposées et les méthodes de référence.

Nous pouvons conclure que les deux méthodes DNN-GMM et GMM-DNN-GMM sont plus performantes que les méthodes de transformation GMM et DNN. Nous pouvons observer que les valeurs de SER pour les méthodes proposées sont meilleures que celles obtenues avec les systèmes de référence (GMM et DNN) pour tous les types de conversion inter-genre et intra-genre. Le tableau 2.6 présente les distances cepstrales CD pour les méthodes proposées et les méthodes de référence.

Nous pouvons remarquer que les distances cepstrales obtenues avec les deux méthodes proposées sont inférieures à celles obtenues avec les méthodes de référence.

2.6.2 Évaluation subjective

Les performances des différents systèmes proposés ont été évaluées par des tests subjectifs. Les tests subjectifs couramment utilisés dans le domaine de la conversion vocale sont les tests MOS (Mean Opinion Score) et XAB.

Le test MOS : C'est un test subjectif qui permet d'évaluer la qualité de la voix convertie. Plusieurs travaux de recherche ont utilisé ce type de test pour évaluer leurs travaux (Kain et Macon, 1998; Toda et al., 2007). Les sujets jugent par une note la qualité de la voix convertie. Avant de commencer le test MOS nous présentons à chaque sujet les niveaux de qualité possibles indiqués dans le tableau 2.9. Le score moyen de test MOS est utilisé pour juger la qualité de la voix convertie.

TABLE 2.9 – *Note graduelle à 5 niveaux concernant le test MOS.*

Note	Qualité
1	**Mauvaise**
2	**Médiocre**
3	**Correcte**
4	**Bonne**
5	**Excellente**

Test XAB : XAB est un test de choix dans lequel les auditeurs classent la voix convertie X comme plus proche de la voix source ou cible, qui sont les échantillons A et B dans un ordre aléatoire.

Évaluation de la méthode GMM-LMF

Dans nos expériences, nous avons utilisé 100 paires de phrases pour la phase d'apprentissage et 20 autres pour la phase de test. La qualité et le naturel de la parole convertie sont évalués par dix auditeurs. Les auditeurs ont comparé au total 64 phrases choisies au hasard et composées de quatre phrases test différentes provenant de huit types de conversion pour chacune des deux approches.

En considérant les MOS moyens (voir la figure 2.15), il s'avère que les participants préfèrent la conversion GMM-LMF. Comme le montre la figure 2.15, le système proposé GMM-LMF fournit les scores les plus élevés (MOS=3.16) par rapport à la méthode de référence (GMM-GMF)(MOS=2.94). On peut conclure que la performance du système de conversion est directement liée à la fonction de conversion utilisée.

Dans notre système de conversion vocale la performance de la fonction de transformation locale a un effet sur l'estimation de l'excitation glottique. En effet plus la conversion est meilleure plus la prédiction de l'excitation est performante. Lors de tests d'écoute XAB, les participants ont reconnu l'identité de trois phrases sélectionnées au

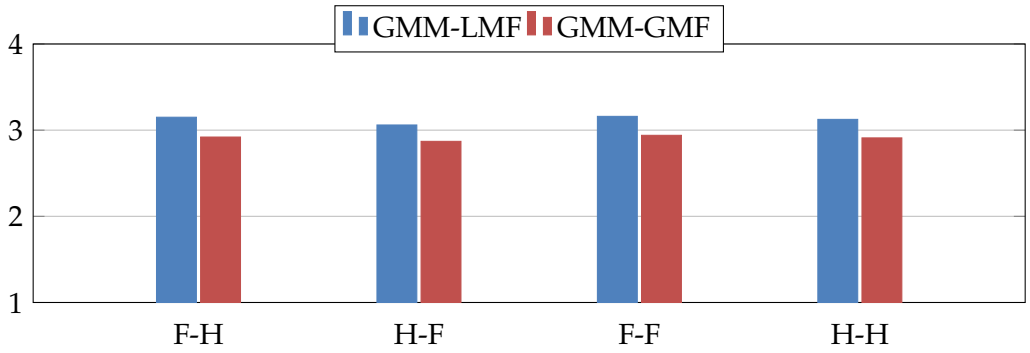


FIGURE 2.15 – *MOS obtenus par la conversion GMM-LMF et la conversion conjointe GMM-GMF.*

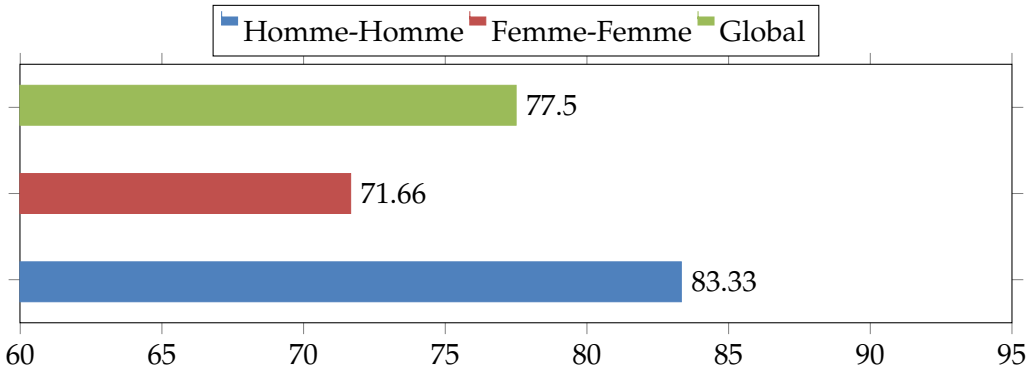


FIGURE 2.16 – *Résultats du test XAB*

hasard parmi un ensemble de 20 échantillons de test pour chaque paire de conversions. Le score final est obtenu en pourcentage pour les voix converties X qui sont reconnues comme étant le locuteur cible. Les résultats du test XAB sont présentés à la figure 2.16. Le taux de reconnaissance chez les hommes est plus élevé que chez les femmes.

Il faut noter que les résultats du test XAB sont directement liés à la méthode d'extraction de l'excitation et de la phase : cette méthode nous permet d'imiter de manière significative les caractéristiques de la voix cible.

Nous avons conclu que l'excitation glottique est le facteur le plus déterminant pour l'identification du locuteur par les auditeurs.

Évaluation des méthodes DNN-GMM et GMM-DNN-GMM

Pour évaluer les performances de conversion de la voix, nous avons effectué des tests d'écoute. Ces tests reposent sur la collecte d'opinions humaines et sont liés à la perception humaine. Nous avons décidé d'utiliser un DNN avec quatre couches cachées avec [512,256,256,512] neurones pour le modèle GMM-DNN-GMM et deux couches cachées, chaque couche comportant 512 neurones pour le modèle DNN-GMM. Pour

GMM, nous avons utilisé pour les trois modèles huit classes et huit sous-classes. Afin de démontrer que les modèles de transformation DNN-GMM et GMM-DNN-GMM peuvent être appliquées avec succès à différentes bases de données, nous avons collecté les scores MOS pour la conversion de voix réalisés avec 8 paires différentes de locuteurs : quatre intra-genres et quatre inter-genres.

– BDL (Homme) - CLB (Femme).
– RMS (Homme) - SLT (Femme).
– CLB (Femme) - BDL (Homme)
– SLT (Femme) - RMS (Homme).
– CLB (Femme) - SLT (Femme).
– SLT (Femme) - CLB (Femme).
– RMS (Homme) - BDL (Homme).
– BDL (Homme) - RMS (Homme).

Dans nos expériences, nous avons utilisé 100 phrases qui définissent le corpus d'apprentissage et 50 autres phrases définissent le corpus de test.

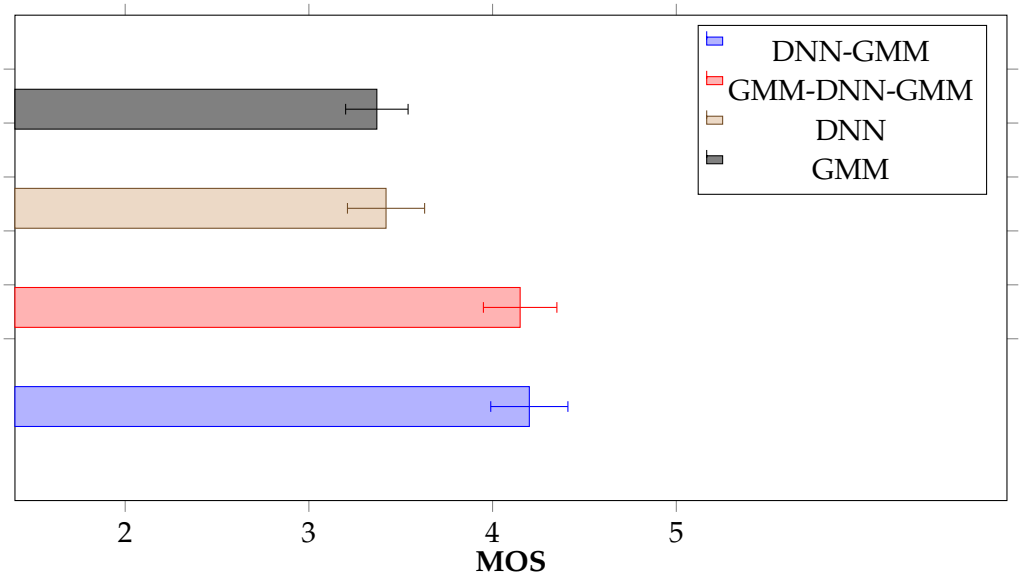
Test MOS : 8 phrases ont été choisies au hasard parmi les phrases tests pour chaque paire de conversion. Le nombre d'auditeurs est 13. Chaque sujet a évalué la qualité de 64 échantillons (issus de 8 phrases test différentes et de 8 paires différentes de locuteurs) pour chacune des quatre approches. Ainsi, chaque auditeur a évalué la qualité de 320 phrases.

Dans ce test, les auditeurs sont invités à juger la voix convertie en termes de qualité et d'intelligibilité. La moyenne de chaque système est montrée dans la figure 2.17. En considérant les MOS moyens, il apparaît clairement que dans tous les cas, les deux systèmes proposés DNN-GMM et GMM-DNN-GMM surpassent les systèmes de référence DNN et GMM. Nous pouvons voir que les résultats obtenus en utilisant les deux modèles proposés sont très proches. Cela indique que les deux modèles basés sur la mise en cascade de GMM et DNN permettent d'améliorer la fonction de conversion.

Dans nos systèmes de conversion, nous évitons d'estimer et de transformer le pitch. Nous estimons l'excitation et la phase à partir de l'espace d'apprentissage cible afin d'obtenir une meilleure qualité de la parole convertie tout en maintenant l'identité du locuteur cible inchangée. Les performances de la fonction de transformation utilisée pour convertir les caractéristiques du conduit vocal jouent un rôle important dans notre système de conversion vocale et ont un effet sur la recherche de l'excitation et de la phase à partir de l'espace d'apprentissage cible. Les détails des scores obtenus pour chaque paire de locuteurs sont illustrés dans la figure 2.18 (a) et (b). DNN-GMM et GMM-DNN-GMM fournissent les scores de préférence les plus élevés pour toutes les paires de conversion.

Test XAB : Dans le test XAB, une paire de voix source et cible a été présentée aux auditeurs après avoir présenté la voix convertie en tant que référence.

FIGURE 2.17 – *MOS obtenus à partir des huit paires de conversion pour les deux méthodes proposées et les méthodes de référence avec des intervalles de confiance de 95 %.*



Le test XAB a été réalisé afin de déterminer si la voix convertie ressemblait au locuteur cible ou au locuteur source (les notes sont : 1 point pour le choix du locuteur cible et 0 pour le choix du locuteur source).

Dans l'expérience d'écoute, nous présentons aux participants trois phrases (A, B et X) et les participants choisissent laquelle de la voix A et B est la plus proche de X. Généralement, X est un exemple de voix convertie, A et B sont choisis au hasard entre les voix source et cible. Nous avons décidé d'évaluer uniquement les paires de conversion intra-genre (Homme-Homme et Femme-Femme). En effet, selon une expérience d'écoute préliminaire, toutes les transformations entre hommes et femmes étaient parfaitement reconnues.

Dans un test d'écoute, dix phrases ont été sélectionnées de manière aléatoire parmi un ensemble de 50 phrases de test pour chaque paire de conversion. Les participants ont reconnu l'identité de 40 échantillons.

Notre objectif principal dans cette étude consiste à évaluer la qualité de l'onde d'excitation afin de prouver que la prédiction de l'excitation et de la phase à l'aide d'un arbre KD-tree permet d'obtenir un degré élevé de similitude avec le locuteur cible.

Les résultats de la reconnaissance sont montrés à la figure 2.19. Les taux de reconnaissance pour les femmes sont inférieurs aux taux de reconnaissance pour les hommes. Des détails pour toutes les paires concernant les conversions entre femmes et entre hommes sont présentés dans la figure 2.20.

L'extraction de l'excitation et de la phase dans l'espace d'apprentissage cible joue un rôle important dans les résultats de la reconnaissance. Il est clair que les auditeurs ont

FIGURE 2.18 – *Résultats du test MOS pour les méthodes proposées et les méthodes de base pour toutes les paires de conversion avec des intervalles de confiance de 95 % .*

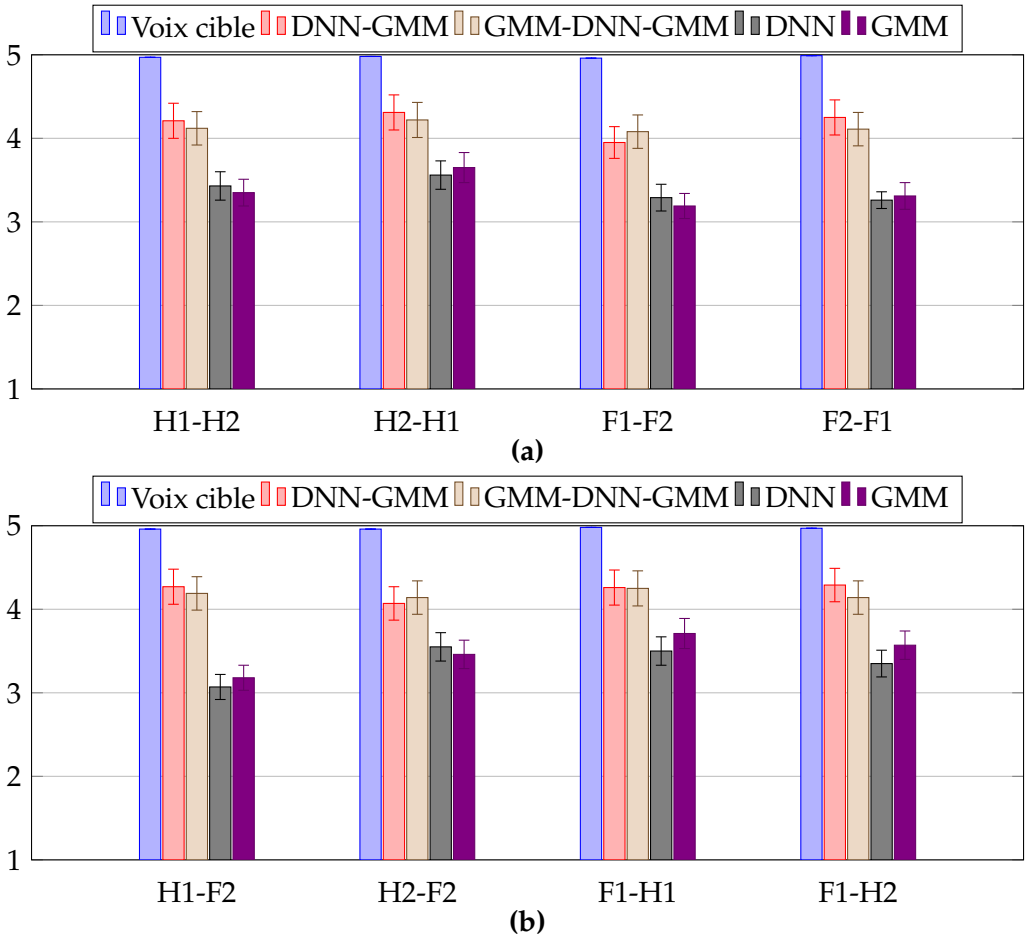


FIGURE 2.19 – *Taux de reconnaissance moyens du locuteur cible obtenus par le test XAB avec des intervalles de confiance de 95 %.*

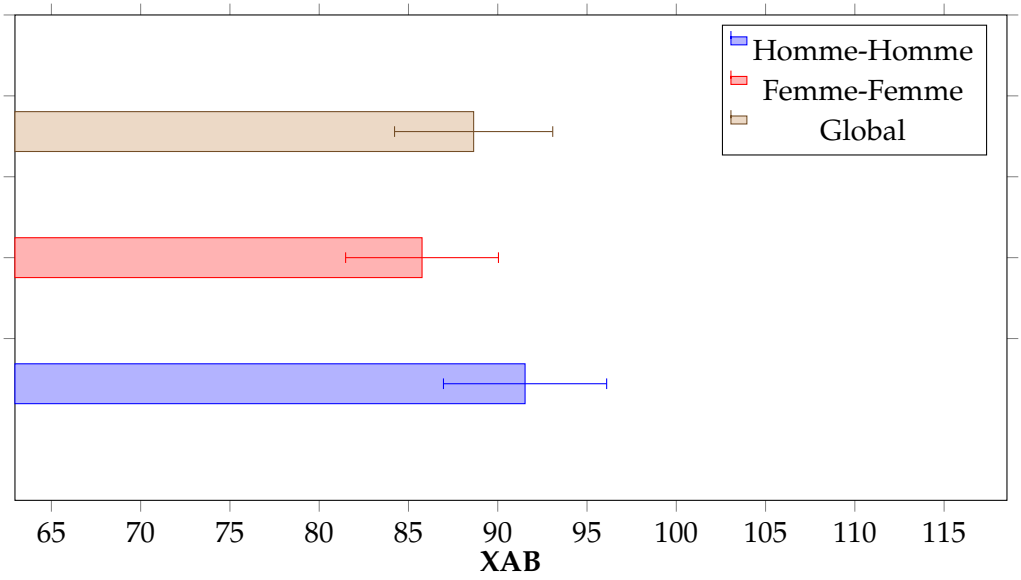
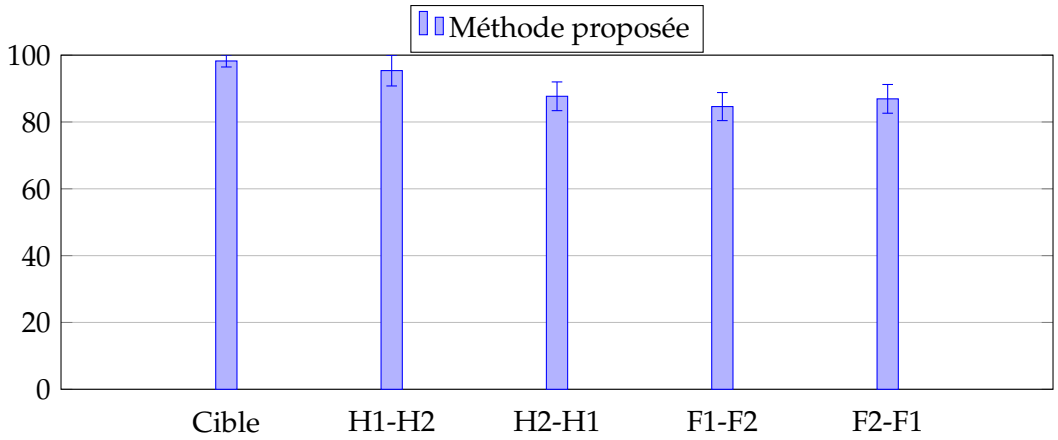


FIGURE 2.20 – *Résultats du test XAB pour toutes les paires de locuteurs avec des intervalles de confiance de 95 %*



réussi à faire la distinction entre la voix cible et la voix source dans la plupart des tests XAB.

Nous prouvons par conséquent que nos méthodes permettent de préserver l'identité du locuteur cible.

2.7 Conclusion

Nous avons décrit dans ce chapitre les étapes de construction de nos systèmes de conversion vocale basés sur l'extraction de l'excitation et de la phase à partir de la base d'apprentissage cible via un arbre binaire à k dimensions (KD-tee). C'est une méthode simple qui permet d'obtenir une voix convertie plus naturelle et plus intelligible, tout en préservant l'identité de la voix cible.

Les résultats du test XAB prouvent clairement que l'extraction de l'excitation et de la phase à partir de l'espace d'apprentissage cible est une technique efficace.

Nous avons présenté aussi un modèle JD-GMM amélioré avec une détermination des fonctions de transformation locale permettant d'améliorer les résultats obtenus avec la technologie JD-GMM classique. Deux autres techniques basées sur la mise en cascade de deux modèles GMM et DNN, tels que les modèles DNN-GMM et GMM-DNN-GMM, permettant de mettre en œuvre une fonction de transformation robuste et performante, ont été proposées.

Chapitre 3

Le rehaussement de la voix œsophagienne

Sommaire

3.1 Introduction

La parole est le langage articulé humain. Elle est destinée à communiquer les pensées, les émotions, les plaisirs, les peines et les joies. Cependant, il arrive que cette dernière soit rendue impossible. Par exemple, le patient laryngectomisé, perd la possibilité de produire une voix laryngée (normale). En effet la laryngectomie totale consiste en une ablation du larynx. Le larynx est un organe situé là où la trachée et l'œsophage se séparent. Le rôle important du larynx est de permettre l'aspiration et d'assurer la sécurité des voies respiratoires en guidant les aliments et les boissons vers l'estomac via l'œsophage et l'air vers les poumons par la trachée. Étant donné que les laryngectomisés n'ont plus de larynx, ils ne peuvent plus produire des sons laryngés. L'impossibilité de communiquer avec la parole engendre une infirmité redoutable durement ressentie par le patient et son entourage. Pour cela l'annonce de cette chirurgie doit être associée à un espoir de restauration de la voix naturelle. Le traitement de la maladie n'est plus suffisant, il faut en plus assurer la réinsertion sociale du patient. L'étude de la parole alaryngée a attiré l'attention des nombreux chercheurs dans des domaines multidisciplinaires (Dibazar et al., 2006; Pravena et al., 2012). Toutefois, l'effort majeur a été orienté plutôt vers l'analyse, l'évaluation, la classification et la reconnaissance de la parole pathologique. Par contre il existe peu de travaux sur l'amélioration de la qualité de ce genre de voix.

3.2 La voix pathologique

La voix pathologique, désigne la parole prononcée par des locuteurs atteints de dysfonctionnements ou de troubles de la voix. Ces troubles de la voix peuvent être une altération momentanée ou durable. Les perturbations ou troubles de la voix engendrés par ces pathologies se traduisent le plus souvent par une détérioration d'un ou de plusieurs paramètres acoustiques de la voix, soit, par ordre d'intensité, de la hauteur tonale, de la fréquence et du timbre. En général, ces pathologies se subdivisent en trois grandes catégories : les pathologies de la voix d'origine fonctionnelle, organique et cancéreuse.

Les pathologies d'origine fonctionnelle : mauvaise utilisation des organes de la phonation par exemple à cause de l'âge ou de l'état psychologique du patient.

Les pathologies d'origine organique : présence de lésions sur les cordes vocales, kystes, laryngite aiguë, etc..., qui sont souvent causées par une infection virale ou bactérienne.

Les pathologies d'origine cancéreuse : suite à une intervention chirurgicale partielle ou totale du larynx qui peut empêcher le sujet d'émettre une voix laryngée. Dans le cadre de ce travail, nous étudierons les dysfonctionnements de la voix liés aux pathologies d'origines cancéreuses.

3.2.1 Le cancer du larynx

Le cancer du larynx est la pathologie la plus fréquente parmi les cancers de la tête et du cou, même s'il s'agit d'une maladie mineure parmi tous les cancers.

Bien que ce soit une maladie handicapante, la détection précoce du cancer est comparativement plus facile que pour d'autres cancers. En effet les cancers du larynx qui se forment sur les cordes vocales (glotte) provoquent souvent un enrouement ou une modification du timbre de la voix appelée dysphonie et une difficulté respiratoire appelée dyspnée.

Le cancer du larynx prend naissance dans les cellules qui tapissent le larynx. Les hommes sont nettement plus touchés que les femmes par ce type de cancer favorisé par le tabagisme et la consommation d'alcool. La quasi-totalité des patients ont fumé. De plus l'alcoolisme concerne au moins un tiers des malades de sexe masculin. Le cancer du larynx est caractérisé par une tumeur sous la forme d'une ulcération anormale de la corde vocale. Parfois le cancer évolue et s'étend vers l'autre corde. À un certain niveau cette maladie provoque l'immobilisation des cordes vocales.

Le traitement consiste alors en une chimiothérapie ou/et une radiothérapie, associée à une ablation chirurgicale de la corde vocale selon la localisation de la tumeur. Les taux les plus élevés des cancers du larynx sont observés en Europe du Sud et au Brésil. La Tunisie est le pays d'Afrique le plus concerné par cette maladie. Ce pays se situe dans la moyenne haute mondiale. Il est très probable que la consommation de tabac brun, est le plus important facteur de risque. La tranche d'âge la plus touchée en Tunisie est celle des 62.6 ans avec 25 % des patients âgés de plus de 70 ans suivie de celle des 50 ans (13%) (Abdallah, 2012).

3.2.2 Laryngectomie totale

Une laryngectomie totale est un geste chirurgical consistant en l'ablation complète du larynx qui est effectué dans le cas de certaines tumeurs avancées du larynx/pharynx ou après l'échec de la radiothérapie.

Par conséquent, la trachée est alors détournée et attachée à la paroi antérieure de la base du cou (voir figure 3.1). Cette ouverture artificielle pratiquée dans la trachée qui se termine sur la face du cou, appelée trachéostome va permettre au patient de respirer. L'air pulmonaire passe donc exclusivement par le trachéostome et par conséquent ne peut pas monter vers la cavité buccale. Ce qui rend impossible la production de la phonation.

Après la chirurgie, certains patients peuvent abandonner toute tentative de communication orale à cause du bouleversement mental et physique engendré par l'acte chirurgical.

En effet les changements anatomiques privent temporairement le patient de sa voix. Dans la période postopératoire immédiate, le patient peut seulement utiliser sa voix

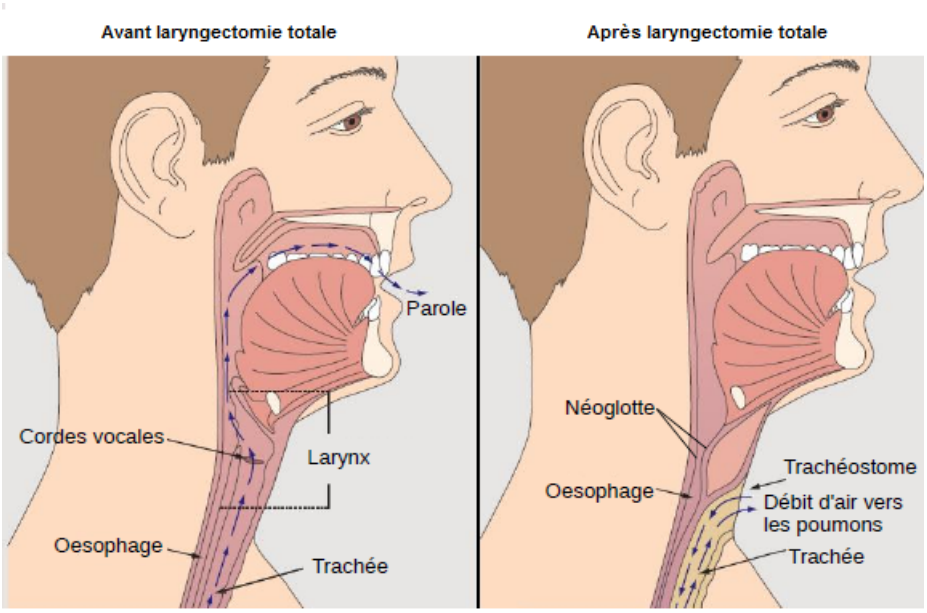


FIGURE 3.1 – *Appareil phonatoire d'une personne laryngectomisée avant (à gauche) et après (à droite) l'opération (Blagnys et Montgomery, 2008).*

chuchotée pour communiquer avec son entourage.

La perte de production de la parole peut entraîner une grande altération de la qualité de vie des patients. A cet égard , les laryngectomisés ont besoin d'autres processus pour produire des sons de parole sans larynx. Ces processus sont, par exemple, le transducteur de larynx artificiel (ALT), la prothèse trachéo-œsophagienne (TE) ou la parole œsophagienne.

3.2.3 Les voix de substitution

Suite à une laryngectomie totale, le patient doit apprendre à communiquer avec une voix nouvelle dite de substitution.

En effet, la suppression de la totalité du larynx prive ce patient de sa voix laryngée.

Les organes bucco-phonatoires ainsi que l'œsophage vont être utilisés par le patient pour l'apprentissage d'une nouvelle voix. Après l'opération, les laryngectomisés peuvent réaliser une restauration de leur voix en utilisant les méthodes de production de la parole alaryngée suivantes :

La voix électro-larynx : Un électro-larynx, également appelé larynx artificiel, est un moyen facile et abordable qui sert à parler après une laryngectomie totale. Ce dispositif permet de créer une vibration lorsqu'il est maintenu contre le menton ou sur la peau du visage (figure 3.3). Ces vibrations atteignent la bouche et sont par la suite modulées

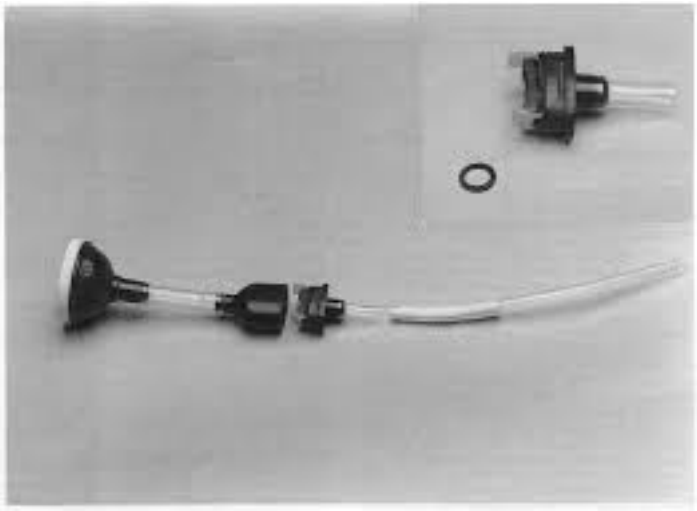


FIGURE 3.2 – *Le larynx artificiel pneumatique (Chalstrey et al., 1994)*

par les résonateurs supérieurs pour produire une voix synthétisée. Bien que cette technique ne requière presque aucun apprentissage et produit une voix intelligible, le son synthétisé reste "robotique".

La voix Pneumatique : Un deuxième type, le larynx artificiel pneumatique, peut être utilisé chez les patients. Le larynx artificiel pneumatique est utilisé en poussant le vibrateur vers le trachéostome et en tenant un sifflet dans la bouche (voir figure 3.2). Les vibrations sont transmises dans la cavité buccale (le lieu d'articulation) pour générer la voix.

Il en résulte que le larynx artificiel pneumatique permet aux patients de parler avec un langage naturel par rapport à un électro-larynx. Un larynx artificiel pneumatique intéressant a été développé (Chalstrey et al., 1994). Cependant, ce dispositif est de moins en moins utilisé de nos jours, même s'il semble utile, car les deux mains du locuteur sont utilisées pour produire la voix alaryngée.

La voix œsophagienne : Elle est l'une des modalités les plus utilisées pour communiquer après une laryngectomie. Après l'ablation totale du larynx, la partie supérieure de la trachée située immédiatement sous le larynx est attachée à une ouverture permanente (stomate) dans la gorge qui permet au patient de respirer. Pour communiquer avec leur entourage, le patient peut apprendre avec l'aide d'un orthophoniste une nouvelle façon de parler avec la voix œsophagienne.

Elle consiste essentiellement à injecter de l'air, et à le renvoyer à travers l'œsophage : l'air fait vibrer les surfaces de l'œsophage et du pharynx pour créer un son œsophagien (voir figure 3.4). L'avantage de la voix œsophagienne est qu'elle s'effectue sans outil (ne nécessite aucun appareil ou procédure) et ne nécessite pas l'utilisation des mains.

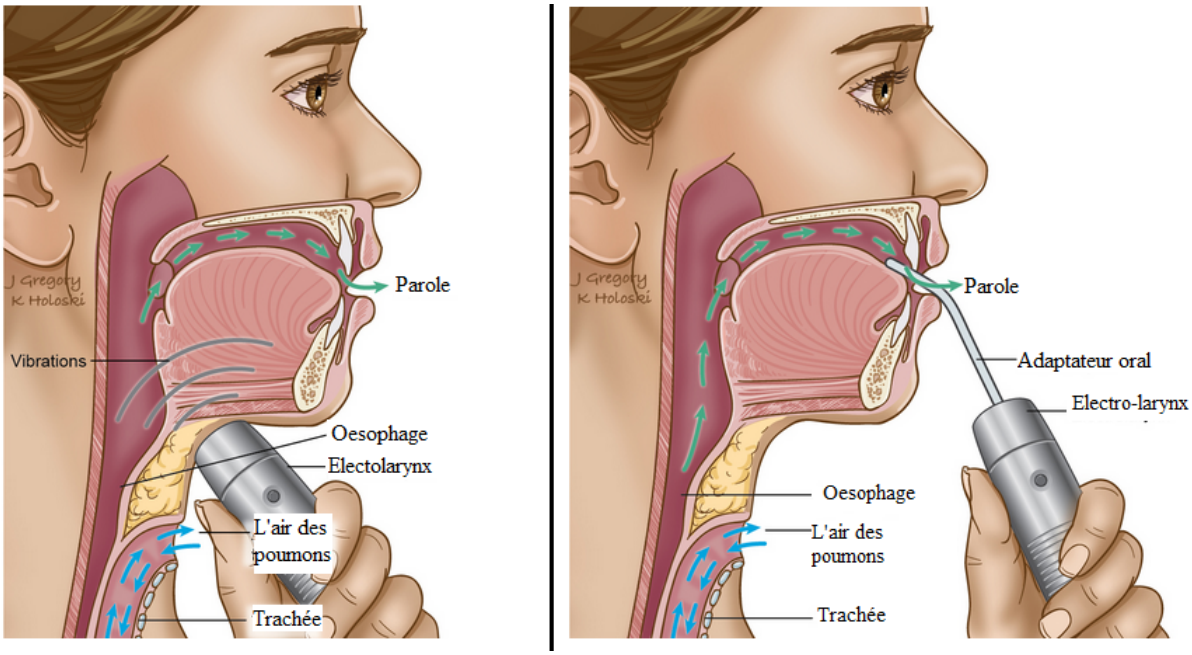


FIGURE 3.3 – *Parole électro-larynx obtenue à l'aide d'un électro-larynx maintenu contre le menton (à gauche) et sur la peau à droite (Mattice, 2015)*

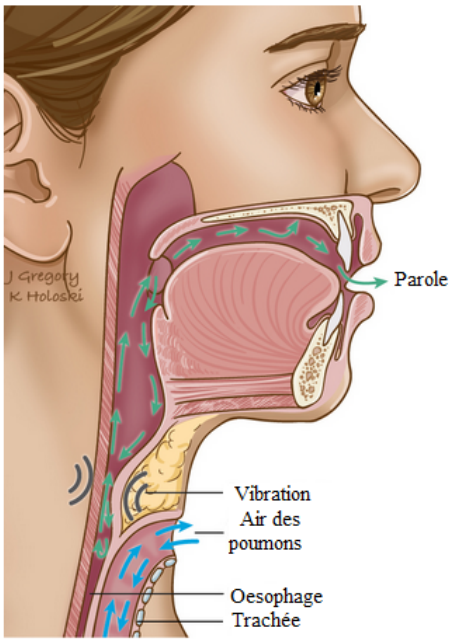


FIGURE 3.4 – *Parole œsophagienne (Mattice, 2015)*

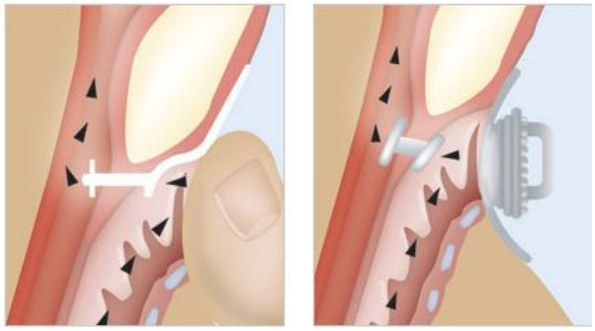


FIGURE 3.5 – *Parole trachéo-œsophagienne avec implant phonatoire (Diamond, 2011)*

La voix trachéo-œsophagienne : Elle consiste à utiliser l'air provenant des poumons et à le faire passer dans l'œsophage par une fistule (une petite prothèse de silicone) créée de façon chirurgicale entre la trachée et l'œsophage. L'air dévié dans l'œsophage par l'intermédiaire de l'implant produit des vibrations dans la néoglotte. La parole est enfin produite à travers l'articulation des cavités de résonance bucco-pharyngo-nasales.

Bien que la voix produite soit plus intelligible et la durée de phonation possible soit plus longue, les implants phonatoires nécessitent un nettoyage régulier et ont une durée de vie limitée. En plus, il faut obturer le trachéostome sur l'expiration pour parler, en appuyant avec les doigts sur la cassette ou le boîtier pour éviter les fuites d'air trachéales, voir (figure 3.5).

3.3 Caractéristiques acoustiques de la voix pathologique

Différents travaux de recherche ont tenté d'analyser les caractéristiques acoustiques de la voix alaryngée en utilisant les méthodes fondamentales du traitement du signal. Les voix de substitution ne peuvent pas être modifiées par des systèmes de conversion conçus pour la voix normale car leurs propriétés spécifiques sont très différentes de celles de la voix laryngée :

Le voisement Le voisement est très difficile à déterminer. Le changement du mécanisme phonatoire provoque des troubles de la parole alaryngée et des changements dans ses différentes caractéristiques acoustiques. De plus le pitch d'une voix de substitution est instable voire chaotique et a une fréquence fondamentale significativement inférieure à celle de la parole laryngée.

La voix électro-larynx La voix semble mécanique, a un timbre métallique, sans variations d'intensité ni de hauteur à cause du signal d'excitation monotone et périodique avec une fréquence fondamentale constante. Un des problèmes majeurs de l'électro-larynx est la radiation sonore, qui est un bruit indésirable produit par l'électro-larynx (Espy-Wilson et al., 1998). Pour remédier à ce problème (Liu et Ng, 2007) propose de modifier le signal d'excitation de l'électro-larynx en plaçant une pièce en mousse synthétique entre l'électro-larynx et le cou, ce qui per-

met de rendre la voix plus naturelle, mais affecte le niveau sonore. Pour pallier ce problème, (Liu et al., 2006a) propose d'utiliser la soustraction spectrale. Cependant, ce type de méthode de réduction du bruit ne peut pas s'adapter à différents volumes sonores. D'autres méthodes ont montré que le lissage du contour de la fréquence F0 affecte l'intelligibilité de la parole (Laures et Weismer, 1999; Laures et Bunton, 2003).

La voix œsophagienne et trachéo-œsophagienne le signal d'excitation produit par la néoglotte est souvent instable : en effet l'intensité de cette voix demeure faible avec une modulation réduite. Cependant, la fréquence fondamentale est chaotique et difficile à détecter. (Bellandese et al., 2001) démontrent qu'il existe une différence importante relative à la fréquence fondamentale F0 entre la voix alaryngée (la voix œsophagienne et trachéo-œsophagienne) et la voix laryngée. En effet, la fréquence F0 de la voix alaryngée est instable avec une fréquence faible. En outre, la voix alaryngée est particulièrement faible en intensité et contient un bruit particulièrement élevé. La voix alaryngée a une énergie vibratoire instable et relativement faible. En outre la quantité d'air obtenue par éructation est relativement faible quant on la compare à la quantité d'air pulmonaire expiré dans le cas de la voix laryngée.

3.4 Les recherches sur l'amélioration de la voix pathologique

La voix pathologique appelée alaryngée se caractérise par un bruit spécifique important. Elle est faible en intensité avec une fréquence fondamentale instable. La faible intelligibilité de cette voix est le problème principal dans les communications orales. Dans la littérature, diverses approches ont été proposées dans le but d'améliorer la qualité et l'intelligibilité de la voix alaryngée. Leur objectif principal est de rétablir les caractéristiques de la voix naturelle dans la mesure du possible.

3.4.1 Les recherches sur l'amélioration de la voix œsophagienne (ES)

La modification des caractéristiques acoustiques de la parole ES produite sur la base du traitement du signal a été largement étudiée. De nombreuses méthodes basées sur cette approche ont été proposées et leur efficacité a été rapportée.

Afin de produire une voix plus naturelle, (Sharifzadeh et al., 2010) a utilisé une prédiction linéaire à excitation par code (CELP) pour estimer les contours de la fréquence fondamentale F0 à partir d'une voix chuchotée. Cependant, il est encore très compliqué d'estimer des signaux excitatifs réalistes similaires à ceux de la voix laryngée. D'autres approches basées sur la réduction du bruit de fond en utilisant un filtrage en peigne (Hisada et Sawada, 2002), ou en utilisant la synthèse par formants (Matsui, 1997; Matsui et al., 2002) ont été proposées.

Bien que les approches citées ci-dessus démontrent une certaine efficacité dans l'amélioration de la parole ES, ces méthodes n'arrivent pas à compenser complètement les

différences acoustiques qui existent entre la voix alaryngée et la voix normale. Ceci est dû au fait que les caractéristiques acoustiques de la parole ES sont très différentes de celles de la parole normale.

D'autres auteurs ont proposé, pour améliorer la voix œsophagienne, de modifier les paramètres acoustiques du conduit vocal, en utilisant la méthode d'analyse LPC (García et al., 2002)), ceci combiné avec un algorithme de réduction du bruit (García et al., 2005). (Ali et Jebara, 2006) proposent, quant à elles, de déplacer les fréquences des formants vers les hautes fréquences. Dans (Mantilla-Caeiros et al., 2010), une technique de reconnaissance vocale est utilisé pour identifier les segments vocaux de la voix alaryngée afin de remplacer ceux-ci par leurs équivalents laryngés. "Adaptive Gain Equalizer" est une autre approche présentée par (Ishaq et Zapirain, 2013) qui permet d'améliorer la voix œsophagienne grâce à la modification de l'excitation glottique.

Récemment, deux approches statistiques basées sur une technique de conversion vocale ont été proposées dans (Doi et al., 2010b) et (Doi et al., 2010a) pour améliorer la voix œsophagienne. Les paramètres spectraux du conduit vocal et les paramètres excitatifs de la voix convertie sont estimés séparément à partir des paramètres spectraux du signal vocal œsophagien : ces méthodes sont fondées sur une technique de conversion statistique basée sur les GMMs. Bien que ces deux approches tentent d'améliorer les caractéristiques acoustiques de la voix œsophagienne, le processus de conversion utilisé est complexe et peut générer des erreurs d'estimation des paramètres. Ainsi, les sons convertis sont ne sont pas naturels en raison des paramètres non réalistes de l'excitation estimée.

La figure 3.6 montre des exemples de formes d'onde, les spectrogrammes, les contours F0 de la voix ES et de la voix laryngée (Ben Othmane et al., 2017a, 2018b). Ces caractéristiques acoustiques ont été extraites grâce à STRAIGHT (Kawahara et al., 1999). Nous pouvons voir que les caractéristiques acoustiques de la voix œsophagienne diffèrent considérablement de celles de la voix laryngée. La voix œsophagienne est perturbée par un bruit spécifique, que l'on peut facilement observer dans les zones de silence de la figure. Ce bruit est produit par le processus de génération de l'excitation glottique, c'est-à-dire par pompage/libération d'air dans l'œsophage et l'estomac .

L'intensité a été déterminée en mesurant la puissance spectrale des signaux normaux et œsophagiens. L'enveloppe de la forme d'onde et les composantes spectrales de la voix œsophagienne ne varient pas de manière aussi régulière que celles de la voix normale. Ces variations instables produisent les sons non naturels spécifiques à la voix œsophagienne (Ben Othmane et al., 2018a).

De plus, le pitch (la hauteur de voix perçue) de la voix œsophagienne est caractérisé par une fréquence plus basse et moins stable que celle de la voix normale. Par conséquent, un processus d'extraction du F0 habituel pour la voix laryngée échoue souvent lors de l'extraction du pitch de la voix œsophagienne.

Dans cette figure, nous pouvons voir que l'estimation de la fréquence F0 ne fonctionne pas. Il est difficile d'extraire les contours F0 de la voix œsophagienne car sa fréquence fondamentale est chaotique. Cependant, il est relativement facile de trouver

un contour F0 à partir d'une voix laryngée.

Ces caractéristiques de la voix œsophagienne entraînent une dégradation importante de la qualité des sons synthétisés. L'intelligibilité et le caractère naturel de la voix œsophagienne dépendent fortement de l'aptitude de chaque patient à produire la parole œsophagienne. Cependant, certains bruits spécifiques sont difficiles à éliminer car ils sont causés par le mécanisme inhérent de production de la parole œsophagienne. Par conséquent, l'amélioration de la voix ES basée sur la modification de ses caractéristiques acoustiques ne permet pas souvent d'obtenir le caractère naturel d'une voix laryngée. Pour améliorer la qualité de la parole ES, il est essentiel de développer des techniques plus sophistiquées permettant une modification plus importante et performante des paramètres de la voix œsophagienne.

3.4.2 Les recherches sur l'amélioration de la voix électro-larynx (EL)

Un des problèmes avec l'électro-larynx est la radiation sonore : c'est un bruit indésirable qui nuit à la communication. Pour apporter une solution à ce problème, une approche consistant à placer un capteur de pression d'air dans le trachéostome a été proposée.

Une autre approche pour améliorer la parole EL consiste à modifier les caractéristiques acoustiques de la parole EL. Dans cette approche, la méthode la plus populaire est la réduction de bruit par soustraction spectrale (SS) (Boll, 1979). Cette technique est basée sur l'hypothèse que les signaux de parole et le bruit additif ne sont pas corrélés. Dans cette méthode, la radiation sonore est considérablement réduite et la voix générée devient plus claire que la voix EL. Cependant, cette méthode n'améliore pas le naturel de la voix EL. (Tanaka et al., 2014) ont proposé une soustraction spectrale pour réduire le bruit et une méthode statistique de conversion de la voix pour prédire les paramètres excitatifs. D'autres approches basées sur la réduction du bruit de fond en utilisant le masquage auditif (Liu et al., 2006b) ont été présentées.

Récemment, une méthode d'amélioration basée sur la conversion vocale statistique a été proposée (Nakamura et al., 2009). Cette méthode, appelée "EL-to-Speech", transforme la parole EL en parole normale grâce à la conversion vocale. Dans le processus d'apprentissage, la relation entre les caractéristiques acoustiques de la parole EL et la parole normale est modélisée par un GMM. Dans la phase de conversion, une voix EL arbitrairement choisie est convertie en voix normale tout en maintenant les informations linguistiques inchangées. Cette méthode améliore considérablement le naturel de la voix EL. Cependant, le contour mélodique généré n'est pas naturel. (Doi et al., 2014) ont proposé de convertir la voix alaryngée avec l'approche de conversion vocale "one-to-Many Eigenvoice". Cette approche propose d'ajuster les vecteurs moyens par des poids de pondération déterminés au cours de la phase d'apprentissage afin de prendre en compte les différentes caractéristiques du locuteur laryngé et pour compenser le manque de données d'apprentissage.

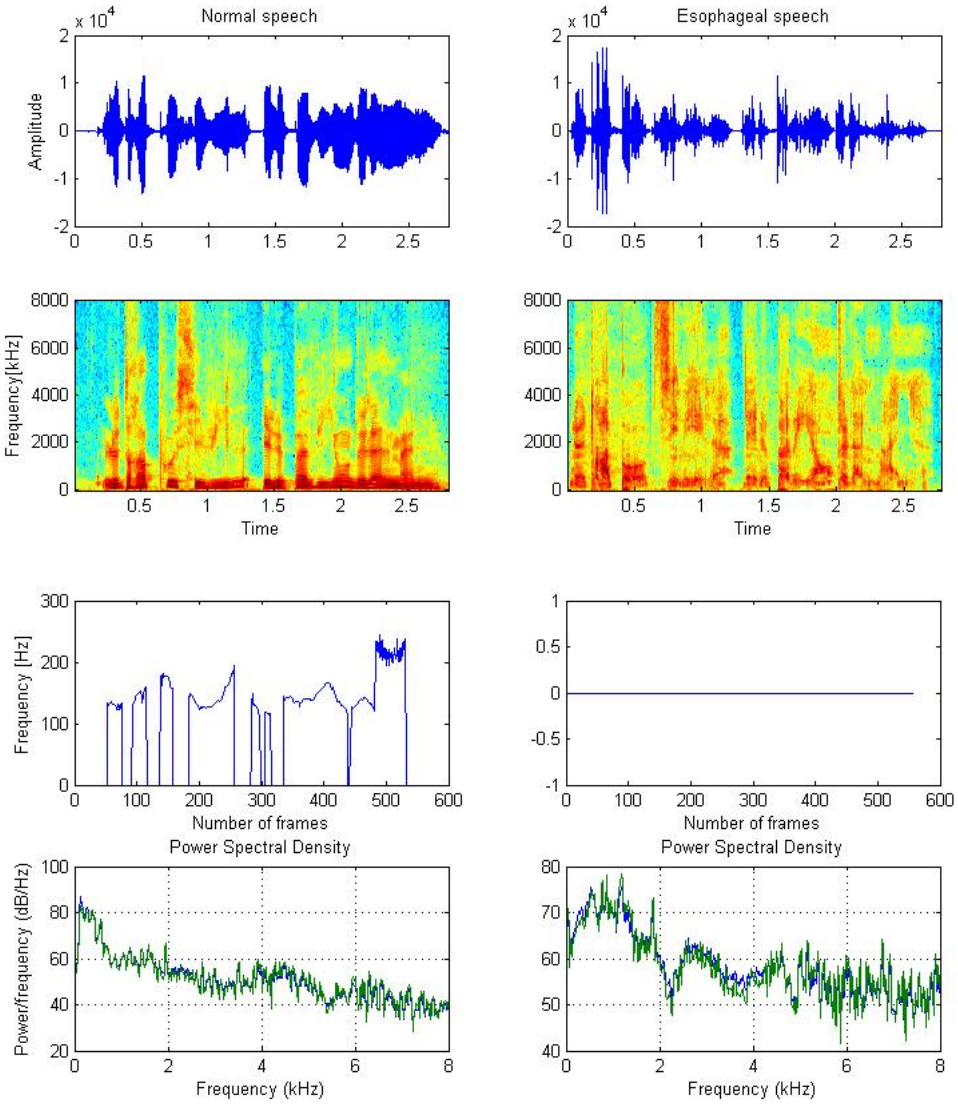


FIGURE 3.6 – *Exemples de formes d'onde, spectrogrammes, contours F0 et puissance spectrale des voix normale et œsophagienne*

3.4.3 Les recherches sur l'amélioration de la voix trachéo-œsophagienne (TE)

Pour améliorer la qualité de la voix trachéo-œsophagienne TE, Qi a tenté de remplacer la source vocale de la parole œsophagienne en utilisant une méthode LPC (Bi et Qi, 1997; Qi et Weinberg, 1995; Qi, 1990). (Qi et al., 1995) ont proposé d'utiliser une forme d'onde glottale artificielle et une fréquence F0 lissée pour améliorer la parole trachéo-œsophagienne. Afin de réduire le souffle et le caractère "grave" de la parole trachéo-oesophagienne originale, (Del Pozo et Young, 2006) ont utilisé une forme d'onde glottique synthétique combinée avec un modèle de réduction du jitter et shimmer. Par la suite del Pozo et Young (Del Pozo et Young, 2008) ont proposé d'estimer les nouvelles durées des phonèmes, de la parole trachéo-œsophagienne grâce à une prédiction par des arbres de régression construits à partir de données laryngées. Les techniques de conversion vocale ont été aussi proposées afin de transformer le plus possible les caractéristiques de la voix alaryngée en celles de la voix laryngée. Généralement, la technique de conversion vocale a été appliquée afin de modifier la voix d'un locuteur source (alaryngée) en celle d'un locuteur cible (laryngée). L'un des premiers systèmes de conversion utilisés pour rehausser la voix TE a été présenté par (Bi et Qi, 1997). Dans cette approche, les auteurs ont proposé d'utiliser la Régression Linéaire Multivariée (RLM) et la Quantification Vectorielles (QV) pour l'estimation de la fonction de transformation. Dans (Bi et Qi, 1997) deux algorithmes de conversion de la parole ont été proposés pour convertir la voix TE pour qu'elle soit perçue comme si elle était prononcée par un locuteur ayant une voix laryngée en utilisant un système de conversion basé sur la quantification vectorielle et un autre basé sur la régression multivariée linéaire (LMR).

3.5 Étapes de construction du système de correction de voix proposé

L'amélioration de la voix œsophagienne a fait l'objet de nombreuses études dont celle toute récente de (Doi et al., 2014). Cette section décrit le système d'aide à la parole basé sur une méthode de conversion vocale pour améliorer la voix ES. Les sons spécifiques et les variations acoustiques instables mentionnées à la section 3.4.1 sont efficacement atténués par le processus de conversion proposé. En outre, même s'il est difficile d'extraire directement certains paramètres tels que la fréquence fondamentale ou les informations voisées/non voisées à partir de la voix ES, notre système permet l'obtention de ces paramètres avec des propriétés similaires à celles de la voix laryngée.

3.5.1 Création de notre base de données

Les corpus de la parole œsophagienne sont moins nombreux par rapport à ceux de la voix normale. Généralement, les analyses portent sur quelques phrases enregistrées par des patients laryngectomisés pour des besoins ponctuels d'une étude. L'enregistrement des voix oesophagiennes et le stockage de données acoustiques sont souvent réalisés par du personnel non formé pour certains aspects techniques. Ce qui explique

souvent la perte fréquente des métadonnées comme par exemple le type de voix pathologique (type de pathologie, le contexte d'enregistrement : analyse acoustique, correction vocale, etc.).

Pour diverses raisons, il est apparu nécessaire de concevoir notre propre base de données française de la parole œsophagienne, notamment pour faire des analyses acoustiques sur plus de données afin de pouvoir trouver des solutions pour ce type de voix.

Les données recueillies comportaient 867 fichiers de données acoustiques. Nous avons enregistré 289 phrases de voix œsophagiennes phonétiquement équilibrées prononcées par deux hommes laryngectomisés français (PC et MH), qui ont effectué une rééducation vocale basée sur une technique d'éructation contrôlée durant plusieurs mois pour maîtriser leur voix de substitution.

Nous avons également enregistré les mêmes phrases prononcées par un locuteur français non-laryngectomisé (AL). Le signal audio a été échantillonné à 16 KHz et directement stocké dans un fichier wave sur un ordinateur. Par la suite ce signal a été segmenté manuellement en 289 fichiers audio. Chaque fichier contient une seule phrase prononcée par le locuteur.

L'objectif principal était de récolter une quantité importante de données phonétiques afin de faciliter l'implémentation de notre système de conversion vocale pour la voix œsophagienne.

Nous avons utilisé deux paires de corpus pour évaluer les performances de la méthode proposée à savoir :
— PC-> AL
— MH-> AL

3.5.2 Phase d'apprentissage

Dans la mise en œuvre d'un système de correction de voix œsophagienne, le principe est de transformer une voix œsophagienne (voix source) en une voix laryngée (voix cible).

Étant donnés la spécificité de la voix œsophagienne, nous proposons dans cette thèse d'appliquer une nouvelle technique de conversion vocale en tenant compte de la particularité de l'appareil vocal des patients qui ont subi une ablation de larynx. En effet, l'ablation des cordes vocales perturbe profondément le signal glottique et par conséquent la voix œsophagienne acquise par le patient laryngectomisé est difficile à comprendre, rauque et faible en intensité.

Il y a une différence d'énergie significative entre la voix œsophagienne et la voix normale. Pour résoudre ce problème, une étape de normalisation initiale a été réalisée afin d'ajuster l'énergie des énoncés du locuteur avant de commencer la phase d'apprentissage.

Pour la phase d'apprentissage, nous procédons exactement de la même façon que lors de l'apprentissage d'un système de conversion voir la figure 3.7.

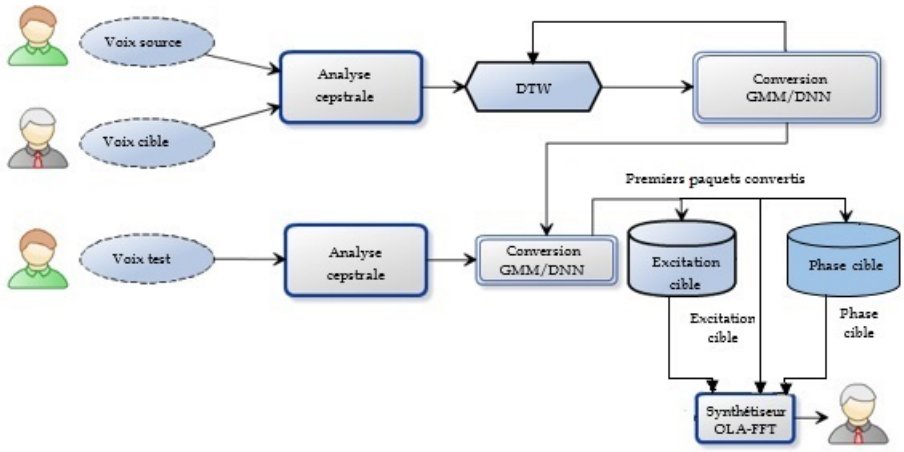


FIGURE 3.7 – *Système d'amélioration de la voix ES*

Deux corpus parallèles ont été utilisés : un concernant la voix œsophagienne comme source et l'autre concernant la voix normale comme cible préalablement alignés à l'aide d'un algorithme de programmation dynamique DTW.

Par la suite, un module d'apprentissage a été appliqué sur les premiers paquets cepstraux des voix source et cible afin de déterminer une fonction de transformation.

Cette fonction a été établie par deux approches différentes : la première, par réseaux de neurones profonds DNN (présenté dans la partie 2.5.3) et la deuxième, à l'aide du modèle de mélange gaussien GMM (présenté dans la partie 2.5.3) (Ben Othmane et al., 2017a, 2018b). La fonction établie permet de convertir les premiers paquets cepstraux.

3.5.3 Phase de test

Estimation de l'excitation et de la phase

Pour estimer l'excitation, les approches classiques transforment la fréquence fondamentale F0 ; après cela, la fréquence F0 convertie est utilisée pour créer une onde excitative qui excite un modèle autorégressif calculé à partir du vecteur cepstral transformé. Cette méthode a plusieurs inconvénients. Comme dans le cas de la parole œsophagienne l'information du F0 est chaotique et la classification des sons voisés/non voisés est particulièrement compliquée, il est difficile voire impossible de prédire la fréquence F0 pour ce type de voix. Bien que les erreurs de décision concernant les trames voisées/non voisées pour la parole œsophagienne ne sont pas très importantes, l'onde excitative estimée n'est pas réaliste et son spectre ne correspond pas nécessairement à un spectre d'excitation réel. Pour cette raison, nous ne pouvons pas appliquer STRAIGHT (Kawahara et al., 1999). Pour résoudre ce problème, nous essayons de prédire l'excitation et la phase sans qu'il soit nécessaire de transformer la fréquence F0. Nous proposons d'utiliser la technique précédemment détaillée dans 2.5.5 pour extraire l'excitation

et la phase de l'espace d'apprentissage cible (voix laryngée).

Réduction du bruit par filtrage morphologique

Afin d'améliorer significativement la voix œsophagienne, il est indispensable d'assurer une bonne réduction du bruit. Actuellement, les opérations morphologiques fonctionnent bien dans le domaine du filtrage de bruit. Dans cette partie, on étudie la méthode de dilatation du cepstre de Fourier (Ben Othmane et al., 2018a) pour améliorer la voix œsophagienne et éliminer ce bruit spécifique.

Le taux de dilatation α d'une fonction s $\in L^2$ (R) est appelé le taux d'homothétie α. L'opérateur D associé avec cette transformation est donc défini par :

$$D_\alpha[s](x) = s(\frac{x}{\alpha}), \alpha \in R^{+*} \tag{3.1}$$

En conséquence, l'action de l'opérateur D_α dans le domaine fréquentiel est une dilatation de l'axe des fréquences de rapport $1/\alpha$. Il en résulte un signal dilaté en fréquence.

$$D_\alpha^f = \alpha D_{\frac{1}{\alpha}} \tag{3.2}$$

D_α^f agit sur le support en fréquence d'un signal dans le sens d'une transposition vers les hautes fréquences pour $\alpha > 1$, et dans le sens d'une transposition vers les basses fréquences pour $\alpha < 1$. Pour cette raison, une simple dilatation/compression transformera également les composantes fréquentielles de la parole d'origine sans distordre la voix. Une représentation en pseudo-code d'un algorithme de dilatation cepstrale est donnée en annexe B

Synthèse

Dans la phase test nous avons proposé quatre systèmes différents *GMM*, *DNN*, DNN_{Src} et $DNN_{Srcdilated}$.

— *GMM* : nous avons utilisé le modèle GMM décrit dans la section 2.5.3.

— *DNN* : nous avons utilisé le modèle DNN décrit dans la section 2.5.3 et nous avons optimisé l'architecture de DNN pour trouver la meilleure architecture pour attaquer le problème de la correction vocale.

— DNN_{Src} : les paquets convertis avec DNN ont été seulement utilisés pour estimer les coefficients cepstraux relatifs au signal d'excitation glottique et la phase. Pour préserver les caractéristiques du conduit vocal du locuteur source les premiers paquets cepstraux source ont été utilisés au niveau de la resynthèse et n'ont pas été remplacés par les paquets convertis (Ben Othmane et al., 2017b).

— $DNN_{Srcdilated}$: les premiers paquets source dilatés ainsi que l'excitation et la phase extraites à partir de la base d'apprentissage ont été utilisés au niveau de la resynthèse.

3.6 Résultats expérimentaux

Pour évaluer notre système de correction de voix, nous avons utilisé trois corpus parallèles (PC, MH, AL). 100 phrases prononcées par les locuteurs source (alaryngés) et cible (laryngé) ont été utilisées pour l'apprentissage et un nouvel ensemble de 20 paires d'énoncés a été utilisé comme ensemble de test. Nous avons utilisé les 32 premiers coefficients cepstraux extraits par analyse cepstrale en tant que vecteur cepstral de conduit vocal. Les 224 autres coefficients cepstraux ont été utilisés comme caractéristiques d'excitation de la voix source. Le temps qui sépare deux trames analysées a été fixé à 4ms et la durée de la fenêtre de Hamming a été fixée à 32 ms.

Nous avons optimisé plusieurs paramètres, tels que le nombre de classes et de sous-classes de chaque GMM. En conséquence, nous avons utilisé 8 classes et 16 sous-classes. Pour l'extraction des caractéristiques de chaque trame, nous avons utilisé une FFT de taille 512. Nous avons optimisé la structure DNN afin d'obtenir la meilleure précision de conversion dans les données d'évaluation du test de validation croisée. En conséquence, nous avons utilisé 4 couches cachées.

Le facteur de dilatation a été fixé à 0.8 pour toutes les expériences. Nous avons réalisé des évaluations objectives et subjectives. Dans les évaluations objectives, nous avons évalué l'impact de la fonction de conversion vocale utilisée sur l'estimation des vecteurs excitatifs. Nous avons également évalué chaque méthode de conversion (DNN et GMM). Pour les évaluations subjectives, nous avons effectué deux tests d'opinion : l'un pour le naturel et l'autre pour l'intelligibilité.

Quatorze auditeurs ont évalué les 5 types de voix suivants :

- ES : La voix œsophagienne enregistrée.
- GMM : synthèse de la voix à l'aide des vecteurs cepstraux de conduit vocal convertis via de la méthode GMM. L'excitation et la phase ont été extraites à partir de la base de données d'apprentissage de la voix laryngée.
- DNN : la voix resynthétisée a été obtenue grâce aux vecteurs cepstraux du conduit vocal convertis et grâce à l'excitation et à la phase qui ont été extraites à partir de la base de données d'apprentissage cible.
- DNN_{Src} : la voix resynthétisée a été obtenue grâce aux vecteurs cepstraux du conduit vocal source et à l'excitation et à la phase extraites à partir de la base de données d'apprentissage cible (en utilisant le vecteur cepstral du conduit vocal converti par DNN_{vt}).
- $DNN_{Srcdilated}$: la voix resynthétisée a été obtenue grâce aux vecteurs cepstraux du conduit vocal source dilatés ainsi que l'excitation et la phase extraites à partir de la base d'apprentissage à l'aide des vecteurs cepstraux du conduit vocal convertis DNN_{vt}.

La figure 3.8 montre des exemples de formes d'onde et de spectrogrammes de la voix œsophagienne et celle de la voix convertie. Nous pouvons remarquer que les caractéristiques acoustiques de la voix convertie varient de manière plus stable que celles de la voix œsophagienne. Le spectrogramme de la voix convertie montre également

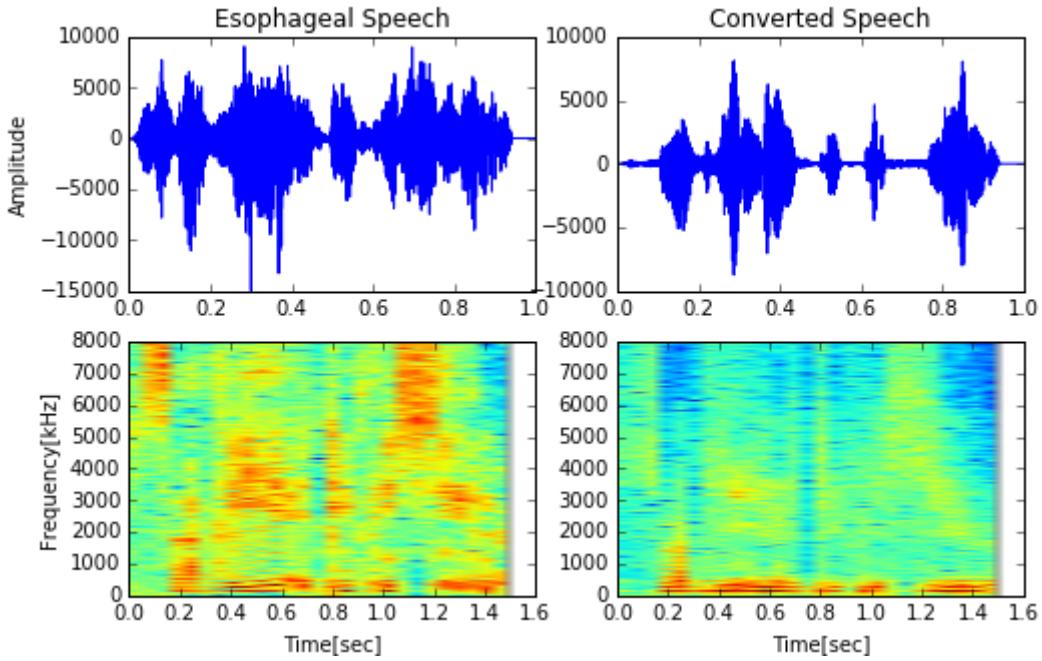


FIGURE 3.8 – *Exemples de formes d'onde et de spectrogrammes des voix convertie et œsophagienne.*

que le système de correction réduit considérablement les bruits spécifiques de la voix œsophagienne.

3.6.1 Évaluation objective

Les évaluations objectives permettent d'évaluer les performances de conversion et pourraient être pratiques pour comparer différents algorithmes. Nous avons calculé le rapport signal sur erreur (SER) de la voix rehaussée. Ensuite, le signal converti a été évalué en utilisant la distance cepstrale. Nous avons comparé les résultats concernant les vecteurs source (ES) et les quatre méthodes de conversion.

Pour construire un système de conversion vocale basé sur DNN, nous avons préalablement optimisé plusieurs paramètres. La tâche la plus importante est d'atteindre une architecture DNN optimale. Nous avons évalué les DNNs, avec différents nombres de couches cachées en changeant le nombre de neurones des couches cachées et en calculant les SERs.

Les SERs obtenus pour différentes architectures DNN sont présentés dans le tableau 3.1 : le DNN avec 4 couches cachées et 512 neurones par couches fournit les meilleures performances. La figure 3.10 montre les enveloppes spectrales du signal vocal source, cible et converti. On peut voir que l'enveloppe spectrale obtenue par la méthode proposée imite la même forme et présente des pics aux mêmes fréquences de l'enveloppe spectrale cible.

TABLE 3.1 – *Estimation du SER pour certaines architectures DNN.*

DNN	PC	MH
512*2	12.54	11.60
256*4	12.61	11.73
512*3	12.85	11.55
512*4	12.98	11.90

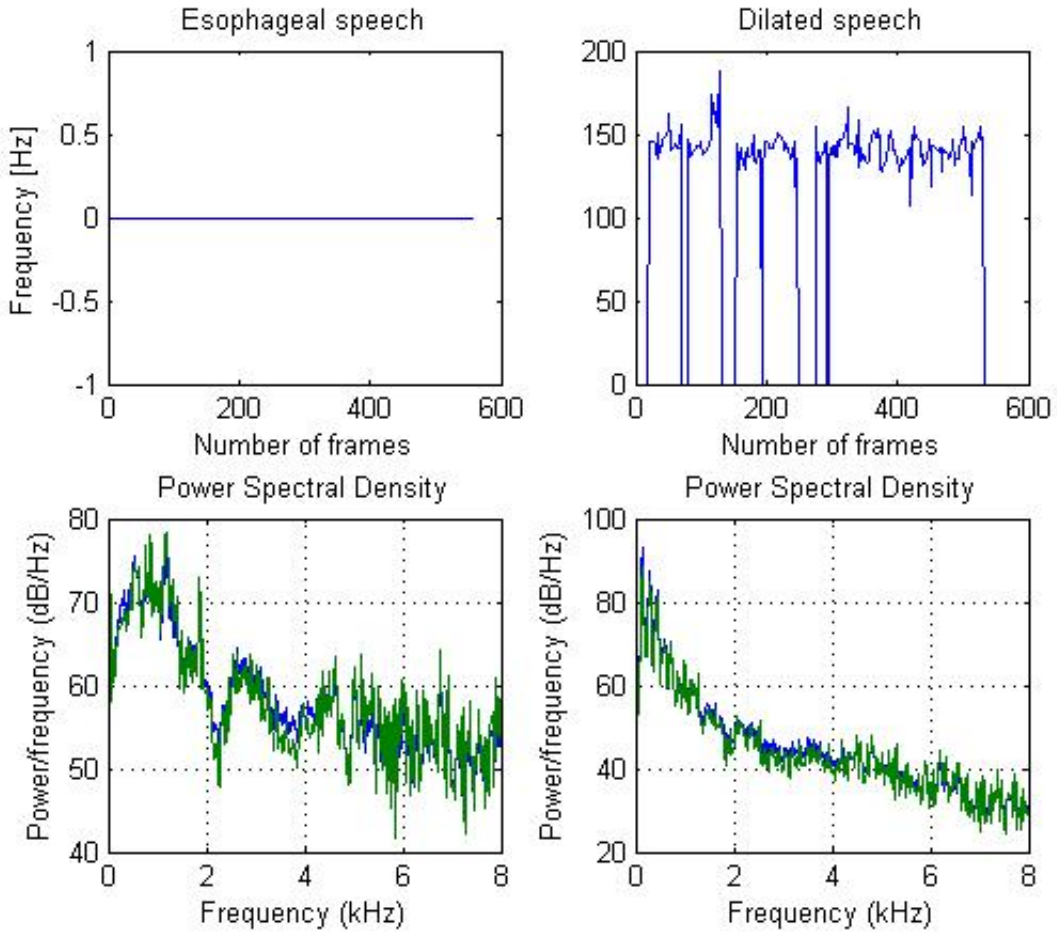


FIGURE 3.9 – *Exemples des contours F0 et de la puissance spectrale de la parole œsophagienne et convertie*

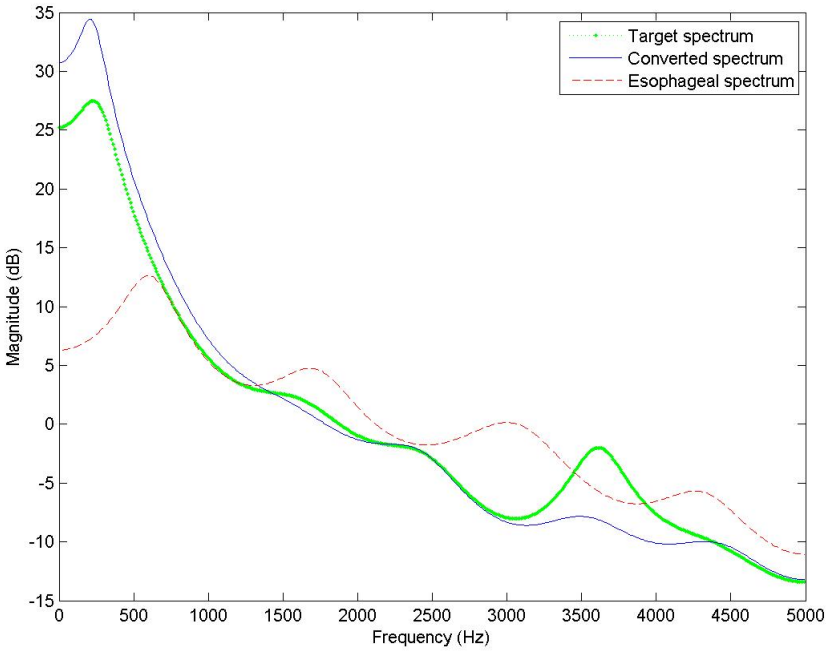


FIGURE 3.10 – *Exemples d'enveloppes spectrales d'un signal vocal source, cible et converti.*

La figure 3.9 montre des exemples des contours F0 et de la puissance spectrale de chacun des échantillons de parole de : a) la voix œsophagienne enregistrée et b) la voix convertie via la méthode $DNN_{Srcdilated}$.

On peut remarquer que les caractéristiques acoustiques de la voix convertie varient de manière plus stable que celles de la voix œsophagienne.

La méthode $DNN_{Srcdilated}$ fournit une voix convertie dont les propriétés sont presque similaires à celles d'une voix laryngée. De plus, il était facile de trouver un contour F0 pour la voix convertie. L'intensité a été déterminée en mesurant la puissance spectrale de la source et du signal de parole convertie. La puissance spectrale de la voix convertie fluctue beaucoup moins que celle de la voix œsophagienne. Par conséquent, la méthode proposée offre une excellente stabilité d'intensité. Le tableau 3.2 montre les différentes valeurs de SER entre les cepstres de Fourier source et cible, puis entre les cepstres de Fourier convertis (obtenus par les quatre méthodes proposées) et les cepstres de Fourier cible. Nous pouvons vérifier que les vecteurs cepstraux de la voix œsophagienne sont très différents de ceux de la voix normale. Nous pouvons également vérifier que ces différences importantes sont considérablement réduites par les différentes méthodes proposées. L'approche proposée basée sur le DNN fonctionne beaucoup mieux que la méthode basée sur le GMM. On observe que le SER de DNN_{Src} est inférieur au SER de DNN. La diminution de SER est due à l'utilisation de vecteurs cepstraux du conduit vocal source, ce qui augmente la distance entre les vecteurs cepstraux cible et les vecteurs cepstraux convertis.

TABLE 3.2 – *Estimation du SER entre les vecteurs cepstraux cible et source (de la voix œsophagienne) et les différents vecteurs cepstraux convertis.*

Corpus	Extrait	GMM	DNN	DNN_{Src}	$DNN_{Srcdilated}$
PC	2.7	12.33	12.98	9.38	10.56
MH	2.97	11.39	11.90	11.01	11.53

TABLE 3.3 – *Estimation de la distance cepstrale CD entre les vecteurs cepstraux cible et convertis.*

Corpus	Extrait	DNN_{Src}	$DNN_{Srcdilated}$
PC	9.28	8.64	7.72
MH	9.03	8.45	8.09

L'utilisation de la dilatation temporelle dans $DNN_{Srcdilated}$ permet de réduire cette distance et de minimiser la distance spectrale (voir le tableau 3.3).

Les résultats obtenus dans le tableau 3.3 montrent que la méthode de conversion proposée est très efficace pour améliorer toutes les caractéristiques acoustiques, à savoir le cepstre du conduit vocal, l'excitation et la phase.

3.6.2 Évaluation subjective

Nous avons effectué deux tests subjectifs du type MOS : l'un pour l'intelligibilité et l'autre pour la qualité. Quatorze auditeurs ont évalué les 5 types de voix. Les auditeurs ont évalué l'intelligibilité et la qualité globale de la voix source et de la voix convertie en donnant à chaque fois une note de 1 à 5 points. Chaque auditeur a jugé 32 échantillons vocaux pour chacun des deux tests.

Résultats expérimentaux sur l'intelligibilité

La figure 3.11 montre les résultats des tests d'intelligibilité. Les résultats établissent que la méthode DNN fonctionne mieux que la méthode GMM.

Les résultats de test MOS confirment ceux obtenus par les tests objectifs. Les résultats des tests MOS d'intelligibilité moyens confirment que les sujets préfèrent la méthode $DNN_{Srcdilated}$: les stimuli obtenus par cette méthode leur semblent plus naturels que ceux obtenus par DNN_{Src} et DNN. En effet, comme le montre la figure 3.11 la conversion $DNN_{Srcdilated}$ a obtenu un MOS égal à 3.49 contre 3.10 pour la conversion DNN_{Src}. Nous pouvons conclure que la méthode proposée $DNN_{Srcdilated}$, préserve, voire améliore l'intelligibilité de la parole œsophagienne et permet de réduire considérablement les bruits spécifiques de la voix œsophagienne grâce à l'algorithme de dilatation proposé.

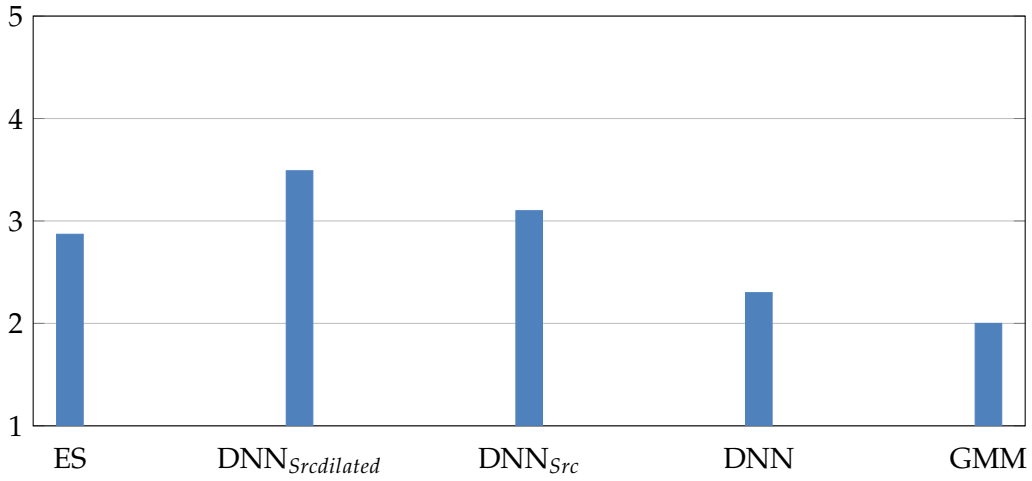


FIGURE 3.11 – *Résultats des tests MOS d'intelligibilité*

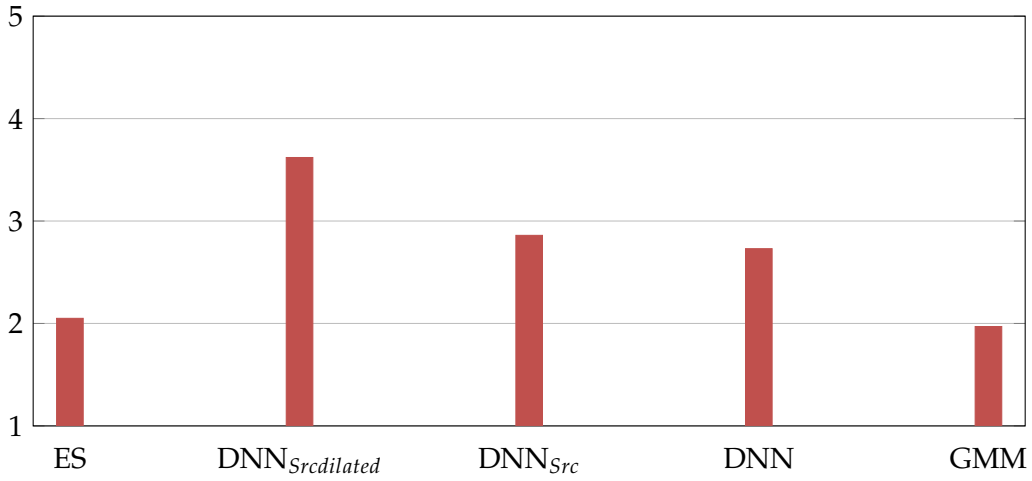


FIGURE 3.12 – *Résultats des tests MOS de qualité*

Résultats expérimentaux sur la qualité

La figure 3.12 montre les résultats du test MOS pour la qualité. La qualité est améliorée par l'excitation réaliste et l'estimation de phase que nous proposons. En considérant les MOS moyens, il s'avère que les auditeurs préfèrent la méthode $DNN_{Srcdilated}$. Comme le montre la figure 3.12 la méthode $DNN_{Srcdilated}$ a obtenu un MOS égal à 3.62 contre 2.86 pour la méthode DNN_{Src}

Les résultats montrent également qu'une amélioration significative est obtenue avec la méthode $DNN_{Srcdilated}$. De plus nos résultats montrent que la méthode $DNN_{Srcdilated}$ produit une voix plus naturelle par rapport aux autres voix synthétisées par les autres méthodes.

Ces résultats démontrent que la méthode $DNN_{Srcdilated}$ permet effectivement une

amélioration de l'intelligibilité et de la qualité de la voix œsophagienne.

3.7 Conclusion

Cette section a présenté une nouvelle méthode d'amélioration de la voix œsophagienne à l'aide d'une méthode de conversion vocale. L'appareil vocal, l'excitation et la phase sont estimés séparément. Deux méthodes ont été utilisées pour transformer les vecteurs cepstraux du conduit vocal, l'une basée sur le modèle GMM et l'autre sur le modèle DNN. Les vecteurs cepstraux du conduit vocal sont convertis afin d'extraire l'excitation et la phase de l'espace d'apprentissage cible (laryngé) par un algorithme de sélection de trame utilisant un arbre binaire KD-tree.

Dans la phase de synthèse (dans les expériences $DNN_{Srcdilated}$ et DNN_{Src}), les vecteurs cepstraux convertis ne sont pas utilisés afin de préserver les caractéristiques du conduit vocal du locuteur source. Les vecteurs convertis ne sont utilisés que pour prédire l'excitation et la phase. Dans ce travail, les cepstres de Fourier dilatés sont utilisés afin de réduire le bruit œsophagien et d'améliorer les vecteurs cepstraux du conduit vocal source. Nous avons effectué des évaluations objectives et subjectives : les résultats expérimentaux ont prouvé que la méthode proposée $DNN_{Srcdilated}$ offre des améliorations importantes du caractère naturel de la voix œsophagienne convertie tout en préservant, voire améliorant, son intelligibilité.

Chapitre 4

Conclusions et perspectives

L'étude réalisée au cours de cette thèse a un double objectif : le développement d'un système de conversion vocale performant et robuste que nous pouvons appliquer à la correction de la voix œsophagienne L'objectif de la conversion vocale est de modifier les caractéristiques vocales d'un signal émis par un locuteur source de manière à ce qu'il semble à l'écoute, être prononcé par un autre locuteur désiré.

Pour développer un système de conversion de voix, il est nécessaire de bien comprendre les mécanismes de production de la voix et il faut avoir une bonne connaissance des paramètres acoustiques caractérisant l'identité du locuteur. C'était l'objectif du premier chapitre.

Dans le deuxième chapitre nous avons étudié les principes d'un système de conversion de voix et nous avons dressé un état de l'art des méthodes de conversion présentées dans la littérature.

Nous avons examiné dans cette thèse trois problèmes principaux et nous avons cherché une solution pour chacun d'eux.

— Trouver une fonction de transformation plus performante et plus efficace pour transformer les coefficients cepstraux du conduit vocal.
— Améliorer l'alignement temporel.
— Trouver une méthode efficace et intelligente pour prédire l'excitation glottique afin d'éviter l'extraction de la fréquence fondamentale F0. En effet cette technique classique n'est pas efficace car l'onde excitative générée est artificielle et son spectre ne correspond pas à un spectre excitatif réaliste.

Dans un premier temps, nous avons pris ces questions en considération et nous proposons d'adapter la stratégie d'apprentissage des fonctions de conversion. À cet effet, nous nous sommes inspirés des méthodes proposées par les approches qui tentent de réduire l'effet du sur-apprentissage dans un contexte d'optimisation des paramètres de transformation linéaires par GMM.
Nous avons réussi à concevoir une fonction de transformation locale GMM-LMF. Cette fonction a été déterminée à l'aide d'une seconde classification. L'avantage de cette mé-

thode est qu'elle permet de trouver une fonction de transformation plus précise (ce qui minimise l'effet de sur-lissage) tout en réduisant le temps de calcul.

Par une étude comparative entre notre modèle et le modèle proposé dans l'état de l'art JD-GMM, nous avons remarqué que les fonctions de conversion locales permettent de réduire le temps de calcul pour la phase d'apprentissage et de conversion tout en améliorant les performances du système.

Dans ce chapitre nous avons proposé deux autres modèles basés sur la mise en cascade de la méthode proposé GMM-LMF et un réseau de neurones profond DNN, DNN-GMM et GMM-DNN-GMM, afin de profiter des avantages des méthodes GMM et DNN.

Nous avons effectué une étude comparative des méthodes proposées DNN-GMM et GMM-DNN-GMM et celle-ci a mis en évidence leur très bonne performance par rapport aux techniques classiques fondées sur GMM ou DNN

Nous avons proposé de répéter le processus d'alignement DTW et de conversion, dans le but d'affiner le chemin d'alignement temporel et par conséquence amender la liste de correspondances. L'indépendance entre l'enveloppe spectrale et la fréquence fondamentale est une hypothèse courante dans le domaine du traitement du signal vocal.

Ainsi, tous les codeurs de la voix étudient ces deux paramètres séparément. Les transformations prosodiques et spectrales sont généralement traitées indépendamment l'une de l'autre.

Des études antérieures en synthèse vocale ont prouvé une dépendance entre l'enveloppe spectrale et le pitch. Cette dépendance est un élément important qui ne peut être négligé dans notre système de conversion de voix.

À cet effet, nous avons supposé qu'il existe une forte corrélation entre le vecteur cepstral du conduit vocal et l'excitation. Notre étude a abouti à une méthode intelligente pour la prédiction de l'excitation à partir du vecteur cepstral du conduit vocal via un arbre binaire KD-tree. Par la suite nous avons proposé d'extraire l'excitation et la phase à partir de l'espace d'apprentissage cible, préalablement codé sous la forme d'un arbre binaire, afin d'estimer un signal excitatif réaliste.

Cette méthode intelligente d'extraction de l'excitation et de la phase nous a permis de préserver l'identité du locuteur cible. En se basant sur le système de conversion vocale proposé, nous avons développé des algorithmes de correction vocale. Les résultats expérimentaux obtenus par ces algorithmes nous ont encouragé à creuser dans le domaine de la correction vocale afin d'améliorer la qualité de la voix œsophagienne.

Dans le troisième chapitre nous avons étudié la problématique de la parole oesophagienne (alaryngée). L'étude de ce type de voix pose plusieurs contraintes : 1) les corpus de la parole oesophagienne existants ne sont pas dédiés à la conversion, à cause d'un manque de données (uniquement un seul corpus de voix oesophagienne) ; 2) contrairement à la voix laryngée, la voix œsophagienne est caractérisée par une faible intelligibilité, un bruit spécifique élevé, et une fréquence fondamentale F0 instable.

Ainsi la voix produite est une voix faible en intensité, rauque, non naturelle et par conséquent difficile à comprendre. Les systèmes de conversion de la voix laryngée peuvent être adaptés à ce type de voix mais avec des pertes en performance.

En effet, il est difficile de compenser les distorsions cepstrales entre ces deux types de voix (laryngée et alaryngée) ; en outre l'estimation de la fréquence fondamentale est difficile voire même impossible pour la voix œsophagienne.

Pour apporter une solution à tous ces défis, nous avons essayé de trouver une fonction de transformation efficace et performante pour convertir les vecteurs cepstraux de conduit vocal et nous avons utilisé avantageusement notre méthode intelligente d'extraction de l'excitation et de la phase à partir de l'espace d'apprentissage cible (la voix laryngée).

Deux techniques de correction vocale ont été proposées et exploitées dans ce mémoire : la méthode DNN_{Src} et la méthode $DNN_{Srcdilated}$. Nous avons procédé de la même manière que pour la conversion de la voix laryngée à ceci près que les premiers paquets cepstraux n'ont pas été modifiés au cours de la resynthèse afin de préserver les caractéristiques du conduit vocal du locuteur source. Après resynthèse par le synthétiseur OLA-FFT, une voix «laryngée» plus naturelle que l'originale a été obtenue, avec une reconstruction effective des informations prosodiques. Et ce, tout en préservant, et c'est le point fort de notre étude, les caractéristiques du conduit vocal inhérentes au locuteur source.

Dans le but de réduire la distorsion spectrale et le bruit spécifique de la voix œsophagienne nous avons proposé un algorithme de dilatation cepstrale qui a été utilisé comme un filtre pour la voix œsophagienne $DNN_{Srcdilated}$. Au niveau de la resynthèse les premiers paquets cepstraux source dilatés sont utilisés. Les résultats expérimentaux montrent que ce système de correction a permis de réduire considérablement les bruits spécifiques œsophagiens et a amélioré la qualité de la voix œsophagienne.

4.1 Perspectives

Le travail présenté dans cette thèse est une démarche pour répondre à la problématique que nous nous sommes fixée. Néanmoins, ces solutions sont certainement incomplètes et laissent entrevoir des perspectives prometteuses.

Dans cette étude, seules des modifications à l'échelle segmentale ont été effectuées.

Or le style d'élocution, comme la prosodie est également un élément essentiel de son identité vocale. Il est par conséquent nécessaire de mettre en œuvre une méthode qui vise à imiter le style d'élocution du locuteur cible. À cette fin, il conviendrait de trouver une méthode qui permet une modélisation et transformation de la prosodie d'un locuteur.

Nous envisageons d'étendre notre corpus de la voix œsophagienne. Nous envisageons aussi de mettre en œuvre d'autres types de réseaux de neurones comme les ré-

seaux modernes LSTM. Afin de réduire le bruit œsophagien, l'algorithme de dilation cepstrale peut être remplacé par d'autres techniques de débruitage plus sophistiquées pour traiter les premiers paquets cepstraux source.

Bibliographie

(Abdallah, 2012) M. B. Abdallah, 2012. Registre des cancers. nord-tunisie. données 2004–2006.

(Abe et al., 1990) M. Abe, S. Nakamura, K. Shikano, & H. Kuwabara, 1990. Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan (E) 11*(2), 71–76.

(Ali et Jebara, 2006) R. H. Ali & S. B. Jebara, 2006. Esophageal speech enhancement using excitation source synthesis and formant structure modification. *Signal Processing for Image Enhancement and Multimedia Processing (SITIS) 31*, 615–624.

(Arslan, 1999) L. M. Arslan, 1999. Speaker transformation algorithm using segmental codebooks (stasc) 1. *Speech Communication 28*(3), 211–226.

(Arya, 1996) S. Arya, 1996. *Nearest neighbor searching and applications*. Thèse de Doctorat, University of Maryland, College Park.

(Baer et al., 1983) T. Baer, A. Löfqvist, & N. S. McGarr, 1983. Laryngeal vibrations : A comparison between high-speed filming and glottographic techniques. *The Journal of the Acoustical Society of America 73*(4), 1304–1308.

(Bagshaw et al., 1993) P. C. Bagshaw, S. Hiller, & M. A. Jack, 1993. Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching. Dans les actes de *Third European Conference on Speech Communication and Technology*.

(Bahja et al., 2015) F. Bahja, J. Di Martino, E. I. Elhaj, & D. Aboutajdine, 2015. An overview of the cate algorithms for real-time pitch determination. *Signal, Image and Video Processing 9*(3), 589–599.

(Bahja et al., 2016) F. Bahja, J. Di Martino, E. I. Elhaj, & D. Aboutajdine, 2016. A corroborative study on improving pitch determination by time–frequency cepstrum decomposition using wavelets. *SpringerPlus 5*(1), 564.

(Baken, 1992) R. J. Baken, 1992. Electroglottography. *Journal of Voice 6*(2), 98–110.

(Bandoin et Stylianou, 1996) G. Bandoin & Y. Stylianou, 1996. On the transformation of the speech spectrum for voice conversion. Dans les actes de *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, Volume 3, 1405–1408. IEEE.

(Beauregard et al., 2005) G. T. Beauregard, X. Zhu, & L. Wyse, 2005. An efficient algorithm for real-time spectrogram inversion. Dans les actes de *Proceedings of the 8th international conference on digital audio effects*, 116–118.

(Bellandese et al., 2001) M. H. Bellandese, J. W. Lerman, & H. R. Gilbert, 2001. An acoustic analysis of excellent female esophageal, tracheoesophageal, and laryngeal speakers. *Journal of speech, language, and hearing research* *44*(6), 1315–1320.

(Ben Othmane et al., 2017a) I. Ben Othmane, J. Di Martino, & K. Ouni, 2017a. Enhancement of esophageal speech using voice conversion techniques. Dans les actes de *International Conference on Natural Language, Signal and Speech Processing-ICNLSSP 2017*.

(Ben Othmane et al., 2017b) I. Ben Othmane, J. Di Martino, & K. Ouni, 2017b. Vers la transformation de la parole oesophagienne en voix laryngée à l'aide de techniques de conversion vocale. Dans les actes de *7ème Journées de Phonétique Clinique-JPC 7*.

(Ben Othmane et al., 2018a) I. Ben Othmane, J. Di Martino, & K. Ouni, 2018a. Enhancement of esophageal speech obtained by a voice conversion technique using time dilated fourier cepstra. *International Journal of Speech Technology*, 1–12.

(Ben Othmane et al., 2018b) I. Ben Othmane, J. Di Martino, & K. Ouni, 2018b. Enhancement of esophageal speech using statistical and neuromimetic voice conversion techniques. *Journal of International Science and General Applications* *1*(1), 10.

(Ben Othmane et al., 2018c) I. Ben Othmane, J. Di Martino, & K. Ouni, 2018c. Improving the computational performance of standard gmm-based voice conversion systems used in real-time applications. Dans les actes de *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, 1–5. IEEE.

(Bengio et al., 2007) Y. Bengio, P. Lamblin, D. Popovici, & H. Larochelle, 2007. Greedy layer-wise training of deep networks. Dans les actes de *Advances in neural information processing systems*, 153–160.

(Bernoulli, 1738) D. Bernoulli, 1738. Hydrodynamica. *Dulsecker. Consultable en ligne http ://imgbase-scd-ulp. u-strasbg. fr/displayimage. php*.

(Bi et Qi, 1997) N. Bi & Y. Qi, 1997. Application of speech conversion to alaryngeal speech enhancement. *IEEE transactions on speech and audio processing* *5*(2), 97–105.

(Bishop, 2006) C. M. Bishop, 2006. Pattern recognition and machine learning (information science and statistics) springer-verlag new york. *Inc. Secaucus, NJ, USA*.

(Blagnys et Montgomery, 2008) H. Blagnys & P. Montgomery, 2008. Without a larynx. *BMJ* *336*(Suppl S3), 0803124.

(Boll, 1979) S. Boll, 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing* *27*(2), 113–120.

(Chalstrey et al., 1994) S. Chalstrey, N. Bleach, D. Cheung, & C. Van Hasselt, 1994. A pneumatic artificial larynx popularized in hong kong. *The Journal of Laryngology & Otology 108*(10), 852–854.

(Chen et al., 2013) L.-H. Chen, Z.-H. Ling, Y. Song, & L.-R. Dai, 2013. Joint spectral distribution modeling using restricted boltzmann machines for voice conversion. Dans les actes de *Interspeech*, 3052–3056.

(Cho et al., 2011) K. Cho, A. Ilin, & T. Raiko, 2011. Improved learning of gaussian-bernoulli restricted boltzmann machines. Dans les actes de *International conference on artificial neural networks*, 10–17. Springer.

(Chollet et al., 2017) F. Chollet et al., 2017. Keras https ://github. com/fchollet/keras.

(Del Pozo et Young, 2006) A. Del Pozo & S. Young, 2006. Continuous tracheoesophageal speech repair. Dans les actes de *Signal Processing Conference, 2006 14th European*, 1–5. Citeseer.

(Del Pozo et Young, 2008) A. Del Pozo & S. Young, 2008. Repairing tracheoesophageal speech duration. Dans les actes de *Proc Speech Prosody*, 187–190. Citeseer.

(Dempster et al., 1977) A. P. Dempster, N. M. Laird, & D. B. Rubin, 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

(Deng et al., 2014) L. Deng, D. Yu, et al., 2014. Deep learning : methods and applications. *Foundations and Trends® in Signal Processing 7*(3–4), 197–387.

(Desai et al., 2010) S. Desai, A. W. Black, B. Yegnanarayana, & K. Prahallad, 2010. Spectral mapping using artificial neural networks for voice conversion. *IEEE Transactions on Audio, Speech, and Language Processing 18*(5), 954–964.

(Desai et al., 2009) S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, & K. Prahallad, 2009. Voice conversion using artificial neural networks. Dans les actes de *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 3893–3896. IEEE.

(Di Martino, 1984) J. Di Martino, 1984. *Contribution à la reconnaissance globale de la parole : mots isolés et mots enchaînés*. Thèse de Doctorat.

(Diamond, 2011) L. Diamond, 2011. Laryngectomy : The silent unknowns and challenges of surgical treatment. *Journal of the American Academy of PAs 24*(8), 38–42.

(Dibazar et al., 2006) A. A. Dibazar, T. W. Berger, & S. S. Narayanan, 2006. Pathological voice assessment. Dans les actes de *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, 1669–1673. IEEE.

(Doi et al., 2010a) H. Doi, K. Nakamura, T. Toda, H. Saruwatari, & K. Shikano, 2010a. Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models. *IEICE TRANSACTIONS on Information and Systems 93*(9), 2472–2482.

(Doi et al., 2010b) H. Doi, K. Nakamura, T. Toda, H. Saruwatari, & K. Shikano, 2010b. Statistical approach to enhancing esophageal speech based on gaussian mixture models. Dans les actes de *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 4250–4253. IEEE.

(Doi et al., 2014) H. Doi, T. Toda, K. Nakamura, H. Saruwatari, & K. Shikano, 2014. Alaryngeal speech enhancement based on one-to-many eigenvoice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* *22*(1), 172–183.

(DriX, 2004) DriX, 2004. La respiration pour le chant. https://fr.audiofanzine.com/techniques-de-chant/editorial/dossiers/la-respiration-pour-le-chant,p.5.html.

(Erhan et al., 2010) D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, & S. Bengio, 2010. Why does unsupervised pre-training help deep learning ? *Journal of Machine Learning Research* *11*(Feb), 625–660.

(Erro et al., 2010a) D. Erro, A. Moreno, & A. Bonafonte, 2010a. Inca algorithm for training voice conversion systems from nonparallel corpora. *IEEE Transactions on Audio, Speech, and Language Processing* *18*(5), 944–953.

(Erro et al., 2010b) D. Erro, A. Moreno, & A. Bonafonte, 2010b. Voice conversion based on weighted frequency warping. *IEEE Transactions on Audio, Speech, and Language Processing* *18*(5), 922–931.

(Eslava et Bilbao, 2008) D. E. Eslava & A. M. Bilbao, 2008. Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models. *Barcelona, Spain : PhD Thesis, Universitat Politechnica de Catalunya*.

(Espy-Wilson et al., 1998) C. Y. Espy-Wilson, V. R. Chari, J. M. MacAuslan, C. B. Huang, & M. J. Walsh, 1998. Enhancement of electrolaryngeal speech by adaptive filtering. *Journal of Speech, Language, and Hearing Research* *41*(6), 1253–1264.

(Flanagan et Golden, 1966) J. L. Flanagan & R. Golden, 1966. Phase vocoder. *Bell System Technical Journal* *45*(9), 1493–1509.

(García et al., 2002) B. García, J. Vicente, & E. Aramendi, 2002. Time-spectral technique for esophageal speech regeneration. Dans les actes de *11th EUSIPCO (European Signal Processing Conference). IEEE, Toulouse, France*, 113–116.

(García et al., 2005) B. García, J. Vicente, I. Ruiz, A. Alonso, & E. Loyo, 2005. Esophageal voices : Glottal flow restoration. Dans les actes de *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, Volume 4, iv–141. IEEE.

(George, 1991) E. B. George, 1991. *An analysis-by-synthesis approach to sinusoidal modeling applied to speech and music signal processing*. Thèse de Doctorat, Georgia Institute of Technology.

(George et Smith, 1997) E. B. George & M. J. Smith, 1997. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. *IEEE Transactions on Speech and Audio Processing 5*(5), 389–406.

(Gnann, 2014) V. Gnann, 2014. *Signal Reconstruction from Multiresolution Magnitude Spectrograms for Audio*. Shaker Verlag Gmbh, Germa.

(Godoy et al., 2009) E. Godoy, O. Rosec, & T. Chonavel, 2009. Alleviating the one-to-many mapping problem in voice conversion with context-dependent modelling. Dans les actes de *InterSpeech 09 : 10th Annual Conference of the International Speech Communication Association*.

(Gold et Rabiner, 1969) B. Gold & L. Rabiner, 1969. Parallel processing techniques for estimating pitch periods of speech in the time domain. *The Journal of the Acoustical Society of America 46*(2B), 442–448.

(Gopi, 2014) E. Gopi, 2014. *Digital speech processing using Matlab*. Springer.

(Griffin et Lim, 1984) D. Griffin & J. Lim, 1984. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing 32*(2), 236–243.

(Hamon et al., 1989) C. Hamon, E. Mouline, & F. Charpentier, 1989. A diphone synthesis system based on time-domain prosodic modifications of speech. Dans les actes de *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, 238–241. IEEE.

(Helander et al., 2008a) E. Helander, J. Nurminen, & M. Gabbouj, 2008a. Lsf mapping for voice conversion with very small training sets. Dans les actes de *ICASSP*, 4669–4672.

(Helander et al., 2008b) E. Helander, J. Schwarz, J. Nurminen, H. Silen, & M. Gabbouj, 2008b. On the impact of alignment on voice conversion performance. Dans les actes de *Ninth Annual Conference of the International Speech Communication Association*.

(Helander et al., 2012) E. Helander, H. Silén, T. Virtanen, & M. Gabbouj, 2012. Voice conversion using dynamic kernel partial least squares regression. *IEEE transactions on audio, speech, and language processing 20*(3), 806–817.

(Helander et al., 2010) E. Helander, T. Virtanen, J. Nurminen, & M. Gabbouj, 2010. Voice conversion using partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing 18*(5), 912–921.

(Hinton et al., 2012) G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., 2012. Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups. *IEEE Signal processing magazine 29*(6), 82–97.

(Hinton et al., 2006) G. E. Hinton, S. Osindero, & Y.-W. Teh, 2006. A fast learning algorithm for deep belief nets. *Neural computation 18*(7), 1527–1554.

(Hinton et Salakhutdinov, 2006) G. E. Hinton & R. R. Salakhutdinov, 2006. Reducing the dimensionality of data with neural networks. *science 313*(5786), 504–507.

(Hisada et Sawada, 2002) A. Hisada & H. Sawada, 2002. Real-time clarification of esophageal speech using a comb filter. Dans les actes de *Proc. ICDVRAT*, 39–46.

(Ishaq et Zapirain, 2013) R. Ishaq & B. G. Zapirain, 2013. Esophageal speech enhancement using modified voicing source. Dans les actes de *Signal Processing and Information Technology (ISSPIT), 2013 IEEE International Symposium on*, 000210–000214. IEEE.

(Itakura, 1975) F. Itakura, 1975. Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America 57*(S1), S35–S35.

(Kahane, 1978) J. C. Kahane, 1978. A morphological study of the human prepubertal and pubertal larynx. *American Journal of Anatomy 151*(1), 11–19.

(Kain et Macon, 1998) A. Kain & M. W. Macon, 1998. Spectral voice conversion for text-to-speech synthesis. Dans les actes de *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, Volume 1, 285–288. IEEE.

(Kain et Macon, 2001) A. Kain & M. W. Macon, 2001. Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. Dans les actes de *icassp*, 813–816. IEEE.

(Kain, 2001) A. B. Kain, 2001. High resolution voice transformation.

(Kawahara et al., 2001) H. Kawahara, J. Estill, & O. Fujimura, 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. Dans les actes de *Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*.

(Kawahara et al., 1999) H. Kawahara, I. Masuda-Katsuse, & A. De Cheveigne, 1999. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction : Possible role of a repetitive structure in sounds1. *Speech communication 27*(3-4), 187–207.

(Kawanami et al., 2003) H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, & K. Shikano, 2003. Gmm-based voice conversion applied to emotional speech synthesis. Dans les actes de *Eighth European Conference on Speech Communication and Technology*.

(Kominek et Black, 2004) J. Kominek & A. W. Black, 2004. The cmu arctic speech databases. Dans les actes de *Fifth ISCA workshop on speech synthesis*.

(Ladefoged, 1996) P. Ladefoged, 1996. *Elements of acoustic phonetics*. University of Chicago Press.

(Laures et Bunton, 2003) J. S. Laures & K. Bunton, 2003. Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions. *Journal of communication disorders 36*(6), 449–464.

(Laures et Weismer, 1999) J. S. Laures & G. Weismer, 1999. The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech, Language, and Hearing Research 42*(5), 1148–1156.

(Le Huche et Allali, 2001) F. Le Huche & A. Allali, 2001. *La voix : anatomie et physiologie des organes de la voix et de la parole*, Volume 1. (DEPRECIATED).

(Ling et al., 2013) Z.-H. Ling, L. Deng, & D. Yu, 2013. Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing 21*(10), 2129–2139.

(Liu et Ng, 2007) H. Liu & M. L. Ng, 2007. Electrolarynx in voice rehabilitation. *Auris Nasus Larynx 34*(3), 327–332.

(Liu et al., 2006a) H. Liu, Q. Zhao, M. Wan, & S. Wang, 2006a. Application of spectral subtraction method on enhancement of electrolarynx speech. *The Journal of the Acoustical Society of America 120*(1), 398–406.

(Liu et al., 2006b) H. Liu, Q. Zhao, M. Wan, & S. Wang, 2006b. Enhancement of electrolarynx speech based on auditory masking. *IEEE Transactions on Biomedical Engineering 53*(5), 865–874.

(Lu et al., 2013) H. Lu, S. King, & O. Watts, 2013. Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis. Dans les actes de *Eighth ISCA Workshop on Speech Synthesis*.

(Machado et Queiroz, 2010) A. F. Machado & M. Queiroz, 2010. Voice conversion : A critical survey. *Proc. Sound and Music Computing (SMC)*, 1–8.

(Mantilla-Caeiros et al., 2010) A. Mantilla-Caeiros, M. Nakano-Miyatake, & H. Perez-Meana, 2010. A pattern recognition based esophageal speech enhancement system. *Journal of applied research and technology 8*(1), 56–70.

(Matsui, 1997) K. Matsui, 1997. Enhancement of esophageal speech using formant synthesis method. Dans les actes de *Proc. Spring Meet. Acoust. Soc. Jpn.*, Volume 311.

(Matsui et al., 2002) K. Matsui, N. Hara, N. Kobayashi, & H. Hirose, 2002. Enhancement of esophageal speech using formant synthesis. *Acoustical Science and Technology 23*(2), 69–76.

(Mattice, 2015) S. Mattice, 2015. Why alaryngeal speech has a reduced level of intelligibility and how it can be maximized.

(McAulay et Quatieri, 1986) R. McAulay & T. Quatieri, 1986. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing 34*(4), 744–754.

(Mizuno et Abe, 1995) H. Mizuno & M. Abe, 1995. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt. *Speech communication 16*(2), 153–164.

(Mohammadi et Kain, 2017) S. H. Mohammadi & A. Kain, 2017. An overview of voice conversion systems. *Speech Communication 88*, 65–82.

(Moorer, 1978) J. A. Moorer, 1978. The use of the phase vocoder in computer music applications. *Journal of the Audio Engineering Society 26*(1/2), 42–45.

(Moulines et Charpentier, 1990) E. Moulines & F. Charpentier, 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication 9*(5-6), 453–467.

(Moulines et Laroche, 1995) E. Moulines & J. Laroche, 1995. Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech communication 16*(2), 175–205.

(Moulines et Verhelst, 1995) E. Moulines & W. Verhelst, 1995. Time-domain and frequency-domain techniques for prosodic modification of speech. *Speech coding and synthesis*, 519–555.

(Nair et Hinton, 2010) V. Nair & G. E. Hinton, 2010. Rectified linear units improve restricted boltzmann machines. Dans les actes de *Proceedings of the 27th international conference on machine learning (ICML-10)*, 807–814.

(Nakamura et al., 2006) K. Nakamura, T. Toda, H. Saruwatari, & K. Shikano, 2006. Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech. Dans les actes de *Ninth International Conference on Spoken Language Processing*.

(Nakamura et al., 2009) K. Nakamura, T. Toda, H. Saruwatari, & K. Shikano, 2009. Electrolaryngeal speech enhancement based on statistical voice conversion.

(Nakashika et al., 2013) T. Nakashika, R. Takashima, T. Takiguchi, & Y. Ariki, 2013. Voice conversion in high-order eigen space using deep belief nets. Dans les actes de *Interspeech*, 369–372.

(Narendranath et al., 1995) M. Narendranath, H. A. Murthy, S. Rajendran, & B. Yegnanarayana, 1995. Transformation of formants for voice conversion using artificial neural networks. *Speech communication 16*(2), 207–216.

(Ney et al., 2004) H. Ney, D. Suendermann, A. Bonafonte, & H. Höge, 2004. A first step towards text-independent voice conversion. Dans les actes de *Eighth International Conference on Spoken Language Processing*.

(Nguyen, 2009) B. P. Nguyen, 2009. Studies on spectral modification in voice transformation.

(Noll, 1970) A. M. Noll, 1970. Pitch determination of human speech by the harmonic product spectrum, the harmonic surn spectrum, and a maximum likelihood estimate. Dans les actes de *Symposium on Computer Processing in Communication, ed.*, Volume 19, 779–797. University of Broodlyn Press, New York.

(Oppenheim et Schafer, 1968) A. Oppenheim & R. Schafer, 1968. Homomorphic analysis of speech. *IEEE Transactions on Audio and Electroacoustics 16*(2), 221–226.

(Pépiot, 2013) E. Pépiot, 2013. *Voix de femmes, voix d'hommes : différences acoustiques, identification du genre par la voix et implications psycholinguistiques chez les locuteurs anglophones et francophones*. Thèse de Doctorat, Université Paris VIII Vincennes-Saint Denis.

(Pernkopf, 1952) E. Pernkopf, 1952. *Topographische Anatomie des Menschen : Der Hals/von Eduard Pernkopf*. Urban & Schwarzenberg.

(Pravena et al., 2012) D. Pravena, S. Dhivya, et al., 2012. Pathological voice recognition for vocal fold disease. *International Journal of Computer Applications 47*(13).

(Qi, 1990) Y. Qi, 1990. Replacing tracheoesophageal voicing sources using lpc synthesis. *The Journal of the Acoustical Society of America 88*(3), 1228–1235.

(Qi et Weinberg, 1995) Y. Qi & B. Weinberg, 1995. Characteristics of voicing source waveforms produced by esophageal and tracheoesophageal speakers. *Journal of Speech, Language, and Hearing Research 38*(3), 536–548.

(Qi et al., 1995) Y. Qi, B. Weinberg, & N. Bi, 1995. Enhancement of female esophageal and tracheoesophageal speech. *The Journal of the Acoustical Society of America 98*(5), 2461–2465.

(Quatieri et McAulay, 1986) T. Quatieri & R. McAulay, 1986. Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing 34*(6), 1449–1464.

(Rabiner et Schafer, 1978) L. R. Rabiner & R. W. Schafer, 1978. *Digital processing of speech signals*, Volume 100. Prentice-hall Englewood Cliffs, NJ.

(Rao et al., 2007) K. S. Rao, R. H. Laskar, & S. G. Koolagudi, 2007. Voice transformation by mapping the features at syllable level. Dans les actes de *International Conference on Pattern Recognition and Machine Intelligence*, 479–486. Springer.

(Rumelhart et al., 1986) D. E. Rumelhart, G. E. Hinton, & R. J. Williams, 1986. Learning representations by back-propagating errors. *nature 323*(6088), 533.

(Schroeder, 1968) M. R. Schroeder, 1968. Period histogram and product spectrum : New methods for fundamental-frequency measurement. *The Journal of the Acoustical Society of America 43*(4), 829–834.

(Sharifzadeh et al., 2010) H. R. Sharifzadeh, I. V. McLoughlin, & F. Ahmadi, 2010. Reconstruction of normal sounding speech for laryngectomy patients through a modified celp codec. *IEEE Transactions on Biomedical Engineering 57*(10), 2448–2458.

(Srivastava et al., 2014) N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, & R. Salakhutdinov, 2014. Dropout : a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research 15*(1), 1929–1958.

(Stylianou, 1996) Y. Stylianou, 1996. Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification. *Ph. D thesis, Ecole Nationale Superieure des Telecommunications*.

(Stylianou, 2001) Y. Stylianou, 2001. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on speech and audio processing 9*(1), 21–29.

(Stylianou, 2009) Y. Stylianou, 2009. Voice transformation : a survey. Dans les actes de *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 3585–3588. IEEE.

(Stylianou et al., 1998) Y. Stylianou, O. Cappé, & E. Moulines, 1998. Continuous probabilistic transform for voice conversion. *IEEE Transactions on speech and audio processing 6*(2), 131–142.

(Tanaka et al., 2014) K. Tanaka, T. Toda, G. Neubig, S. Sakti, & S. Nakamura, 2014. A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation. *IEICE TRANSACTIONS on Information and Systems 97*(6), 1429–1437.

(Toda et al., 2007) T. Toda, A. W. Black, & K. Tokuda, 2007. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing 15*(8), 2222–2235.

(Toda et al., 2016) T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, & J. Yamagishi, 2016. The voice conversion challenge 2016. Dans les actes de *Interspeech*, 1632–1636.

(Valbret, 1993) H. Valbret, 1993. *Systeme de conversion de voix pour la synthese de parole*. Thèse de Doctorat.

(Valbret et al., 1992) H. Valbret, E. Moulines, & J.-P. Tubach, 1992. Voice transformation using psola technique. *Speech communication 11*(2-3), 175–187.

(Vincent et al., 2010) P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, & P.-A. Manzagol, 2010. Stacked denoising autoencoders : Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research 11*(Dec), 3371–3408.

(Viterbi, 1967) A. Viterbi, 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory 13*(2), 260–269.

(Watanabe et al., 2002) T. Watanabe, T. Murakami, M. Namba, T. Hoya, & Y. Ishida, 2002. Transformation of spectral envelope for voice conversion based on radial basis function networks. Dans les actes de *Seventh International Conference on Spoken Language Processing*.

(Wu, 2015) Z. Wu, 2015. *Spectral mapping for voice conversion*. Thèse de Doctorat, PhD Thesis, St. Scho ol of Computer Engineering, Nanyang Technological ....

(Ze et al., 2013) H. Ze, A. Senior, & M. Schuster, 2013. Statistical parametric speech synthesis using deep neural networks. Dans les actes de *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 7962–7966. IEEE.

(Zhu et al., 2006) X. Zhu, G. T. Beauregard, & L. Wyse, 2006. Real-time iterative spectrum inversion with look-ahead. Dans les actes de *Multimedia and Expo, 2006 IEEE International Conference on*, 229–232. IEEE.

Bibliographie personnelle

Revues internationales avec comité de sélection

BEN OTHMANE, I., DI MARTINO, J., & OUNI, K. (2018). ENHANCEMENT OF ESOPHAGEAL SPEECH USING STATISTICAL AND NEUROMIMETIC VOICE CONVERSION TECHNIQUES. JOURNAL OF INTERNATIONAL SCIENCE AND GENERAL APPLICATIONS, 1(1), 10.

BEN OTHMANE, I., DI MARTINO, J., & OUNI, K. (2018). ENHANCEMENT OF ESOPHAGEAL SPEECH OBTAINED BY A VOICE CONVERSION TECHNIQUE USING TIME DILATED FOURIER CEPSTRA. INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY,SPRINGER VERLAG, 1-12.

Conférences d'audience internationale avec comité de sélection

BEN OTHMANE, I., DI MARTINO, J., & OUNI, K. (2017, JUNE). VERS LA TRANSFORMATION DE LA PAROLE OESOPHAGIENNE EN VOIX LARYNGÉE À L'AIDE DE TECHNIQUES DE CONVERSION VOCALE. IN 7ÈME JOURNÉES DE PHONÉTIQUE CLINIQUE-JPC 7.

BEN OTHMANE, I., DI MARTINO, J., & OUNI, K. (2017, DECEMBER). ENHANCEMENT OF ESOPHAGEAL SPEECH USING VOICE CONVERSION TECHNIQUES. IN INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE, SIGNAL AND SPEECH PROCESSING-ICNLSSP 2017.

BEN OTHMANE, I., DI MARTINO, J., & OUNI, K. (2018, DECEMBER). IMPROVING THE COMPUTATIONAL PERFORMANCE OF STANDARD GMM-BASED VOICE CONVERSION SYSTEMS USED IN REAL-TIME APPLICATIONS. IN 2018 INTERNATIONAL CONFERENCE ON ELECTRONICS, CONTROL, OPTIMIZATION AND COMPUTER SCIENCE (ICECOCS) (PP. 1-5). IEEE.

Journées nationales

BEN OTHMANE, I. & OUNI, K. (22-23 AVRIL 2015) TRANSFORMATION DE LA VOIX. APPROCHES ET APPLICATIONS, PRÉSENTATION ORALE ET ARTICLE PUBLIÉ DANS LES PROCEEDINGS DE LA JOURNÉE DOCTORALE DE L'ECOLE POLYTECHNIQUE DE TUNIS (JDEPT), TUNIS, 2015.

BEN OTHMANE, I.& OUNI, K. (24-26 NOVEMBRE 2016) STUDY, IMPLEMENTATION AND APPLICATION OF A VOICE CONVERSION SYSTEM, MYPHD 2016, HAMMAMET, TUNISIA, 2016.

Annexe A

Python

Le travail de programmation réalisé dans cette thèse a été mené en utilisant plusieurs cadres et technologies combinés. Les différents outils utilisés et la raison de leur choix seront présentés ici. Comme mentionné au chapitre 2, les systèmes de conversion de la parole ont été mis en œuvre à l'aide de plusieurs étapes distinctes, à la fois pour la phase d'apprentissage et la phase test. Différentes technologies ont été choisies pour les différentes étapes. L'apprentissage a été réalisée à l'aide d'une bibliothèque « Deep Learning » en Python. Le raisonnement derrière ces choix est présenté dans les sections suivantes.

Apprentissage en profondeur ou "Deep Learning"

Le Deep Learning étant le concept principal utilisé dans cette thèse, pour sa mise en œuvre, il a été nécessaire de trouver la bonne bibliothèque pour faciliter le développement de nos systèmes. Il y a beaucoup de choses à avoir à l'esprit lors du choix d'une bibliothèque, et il faut évaluer quelle est la priorité la plus importante pour les tâches à accomplir.

Est-il primordial d'avoir la bibliothèque la plus récente et la plus riche en fonctionnalités, ou peut-être faut-il choisir le bon langage de programmation ainsi que les bibliothèques adéquates qui sont compatibles avec celui-ci ? Python (étant l'un des langages de programmation les plus utilisés pour ce type de travaux) a été choisi comme langage de programmation pour la mise en œuvre de nos programmes en raison de sa simplicité.

Theano et Tensorflow sont deux bibliothèques très populaires pour mettre en œuvre des réseaux de neurones. Mais Keras (Chollet et al., 2017) est devenue la bibliothèque la plus usitée. Elle possède toutes les propriétés recherchées et nécessaires dans cette thèse. C'était également l'une des trois bibliothèques suggérées par les développeurs.

Numpy

Numpy est une bibliothèque de base pour l'informatique scientifique en Python. Elle gère correctement les données de grandes dimensions et est complètement compatible avec Keras.

h5py

Le format de données hiérarchique (HDF) est un format de fichier pour le stockage et l'organisation de quantité de données, et h5py est parfaitement adapté avec Python. Cette bibliothèque a été spécialement créée pour le partage de données. L'utilisation de ce format de fichier permet d'utiliser les données provenant de n'importe quel programme.

Keras

Keras est une bibliothèque open source permettant de mettre en œuvre des réseaux de neurones de haut niveau, écrite en python et interfaçable avec CNTK, Theano et Tensorflow.

Depuis sa publication initiale en mars 2015, elle a été favorisée pour sa simplicité d'utilisation et sa simplicité syntaxique, facilitant ainsi son développement rapide. Keras offre des blocs de construction de haut niveau permettant d'interagir avec les algorithmes de réseaux de neurones profonds et de " machine learning".

Cette bibliothèque a pour objectif de permettre des expérimentations rapides et permet d'aller de l'idée au résultat avec le plus faible délai possible. Elle ne gère pas ellemême les opérations de bas niveau tel que les produits tensoriels, les convolutions, etc. Pour ce faire, elle s'appuie sur une bibliothèque spécialisée bien optimisée pour la manipulation de ces opérations.

Keras permet de gérer les problèmes de manière modulaire. De plus, plusieurs moteurs peuvent être connectés de manière transparente à Keras.

Theano

Theano est une bibliothèque Python qui facilite l'écriture et l'exploitation de modèles d'apprentissage profonds et offre la possibilité de les exécuter sur des GPUs.

Annexe B

Algorithme de dilatation

Algorithm 1 Cepstral Dilation/Compression algorithm

```
1: procedure modify_cepstrum(X, factor)
2:     Z = np.zeros(len(X), np.double)
3:     if factor ≠ 1 then
4:         for i in range (len(X)) do
5:             ii = i * factor
6:             inew = int(ii)
7:             α = ii − inew
8:             if inew < len(X) − 1 then
9:                 Z[i] = X[inew] * (1 − α) + X[inew + 1] * α
10:            else
11:                if factor > 1 then
12:                    ifinal = i
13:                    for j in range(len(X) − ifinal) do
14:                        Z[ifinal + j] = Z[ifinal − 1 − j]/(j * 0.05 + 1)
15:                    end for
16:                    break
17:                end if
18:            end if
19:        end for
20:    else
21:        Z = X
22:    end if
23:    return Z
24: end procedure
```