Transformation de l'identité d'une voix

Guillaume DENIS

Rapport de stage du DEA ATIAM

Année universitaire 2002/2003
Université de Paris VI - Université Aix Marseille II - École Nationale Supérieure des Télécommunications - Institut National Polytechnique de Grenoble

Responsable de stage : Xavier RODET
Laboratoire d'accueil : IRCAM

23 février 2006

Remerciements

Je tiens à remercier :

Xavier Rodet, pour m'avoir accueilli en stage, pour les discussions et les conseils sur les orientations à suivre.

Geoffroy Peeters, pour avoir partagé ses connaissances en traitement du signal et en synthèse vocale.

Axel Röbel, dont l'aide précieuse m'a permis d'avancer plus vite face aux problèmes informatiques rencontrés.

Éric Daubresse et Damien Tardieu, pour les enregistrements de voix.

Gérard Assayag et Cyrille Defaye, pour leur disponibilité durant cette année de DEA.

Les élèves du DEA et les stagiaires de l'IRCAM pour les moments passés ensemble.

Résumé

La transformation de voix est une opération qui consiste à modifier les enregistrements audio d'une voix afin d'en changer l'identité perçue. On peut par exemple vouloir créer un enregistrement de voix d'homme à partir d'une voix de femme, en changer le timbre, l'intonation, l'accent ou la prononciation sur certaines régions, etc.

La démarche suivie pendant mon stage consiste à effectuer des analyses comparatives sur des couples de voix différentes (la source et la cible) prononçant le même texte, afin d'en dégager les différences les plus importantes d'un point de vue perceptif : le timbre et la prosodie (la hauteur, l'énergie et la durée des phones et des silences). Lors de cette étape d'apprentissage on compare donc localement des extraits audio correspondant aux mêmes phonèmes prononcés mais présents dans le signal sous la forme d'évènements acoustiques différents.

Une des problématiques principales est la difficulté d'aligner parfaitement les voix source et cible sans connaître a priori leurs comportements respectifs dans un contexte de prononciation précis (ce qu'on cherche à faire). La solution retenue est alors d'effectuer les comparaisons en certains points à la fois caractéristiques des signaux et en correspondance sur les deux enregistrements d'après l'alignement trouvé par programmation dynamique.

Les résultats de l'apprentissage mettent en avant les transformations timbrales et prosodiques à effectuer pour imiter au mieux la voix cible à partir d'enregistrements de la voix source. Ces manipulations du signal audio sont alors effectuées grâce à l'analyse/synthèse PSOLA et au Super Vocodeur de Phase.

Mots-clé : transformation de voix, conversion de voix, morphing, synthèse vocale, apprentissage, DTW, PSOLA.

Table des matières

Table des figures

Abbréviations

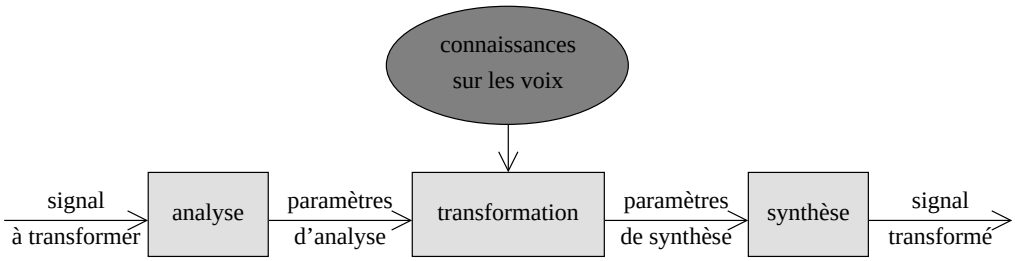
AR	auto-régressif
DFW	dynamic frequency warping *(programmation dynamique fréquencielle)*
DTW	dynamic time warping *(programmation dynamique temporelle)*
LPC	linear predictive coding *(codage par prédiction linéaire)*
MFCC	mel-frequency cepstrum coefficients *(coefficients cepstraux sur l'échelle mel)*
PSOLA	pitch synchronous overlap-add *(addition-recouvrement synchrone au pitch)*
RI	réponse impulsionnelle
TF	transformée de Fourier
TFD	transformée de Fourier discrète

Chapitre 1

Introduction

La transformation de voix est une opération qui consiste à modifier les enregistrements audio d'une voix afin d'en changer l'identité perçue. On peut par exemple vouloir créer un enregistrement de voix d'homme à partir d'une voix de femme, en changer le timbre, l'intonation, l'accent ou la prononciation sur certaines régions, etc.

La conversion de voix s'inscrit dans ce contexte en se fixant plus particulièrement comme objectif d'imiter une voix précise (dite cible) en transformant les enregistrements d'une voix source. On veut donc créer l'illusion perceptive que la cible a prononcé ce que la source a enregistré. Cette démarche peut être incluse dans un schéma général d'analyse/synthèse :



S'il est important d'utiliser des algorithmes de synthèse de grande qualité, afin que les signaux générés paraissent naturels et qu'aucun artéfact ne vienne les dégrader, le point central de la procédure est la fonction de transformation des paramètres d'analyse en fonction de connaissances sur les voix source et cible. Ces connaissances proviennent d'une étape préalable d'apprentissage des différences des voix selon les descripteurs les plus caractéristiques de leur identité.

A cette fin, nous commençons au chapitre 2 par établir quels sont ces descripteurs et quelles sont leurs correspondances avec les paramètres bas niveau manipulés lors d'une analyse/synthèse. Au chapitre 3, nous traitons de l'alignement des signaux de voix, préalable à l'apprentissage lui-même (qui doit être fait sur des portions de signal correspondant aux mêmes phonèmes prononcés, donc alignés), développé au

chapitre 4 suivant diverses solutions et améliorations possibles. La synthèse de ces premiers éléments se fait au chapitre 5 en présentant dans son ensemble la procédure de transformation retenue et les implémentations Matlab associées. Au dernier chapitre, nous discuterons des résultats obtenus, des améliorations et des pistes à suivre pour un travail futur.

Mais commençons par présenter l'état de l'art en transformation de voix et l'équipe Analyse/Synthèse de l'IRCAM pour situer le travail que j'ai effectué en stage.

1.1 État de l'art

Après l'apparition et le développement qualitatif de nombreux modèles de signaux bien adaptés à la voix, la question du contrôle des paramètres bas niveau de ces modèles est centrale dans une perspective d'extension des possibilités des technologies de synthèse vocale. Manipuler de manière cohérente ces paramètres nécessite une gestion de leurs comportements haut niveau, mieux adaptés à l'homme. Ceux-ci pourraient être idéalement le locuteur, son état émotionnel, sa manière de parler (cri, chuchotement, exposé oral, discussion normale), etc. Dans ce contexte, de nombreux chercheurs ont participé depuis les années 1980 à faire évoluer l'état de l'art dans le domaine de la transformation de voix.

En 1985, Childers et al. [10] étudient les facteurs responsables de la qualité d'une voix de synthèse par modèle source-filtre : la forme de l'excitation glottique, l'enveloppe spectrale, la prosodie et la reproduction des consonnes plosives ([p], [b], [t], [d], [k], [g]), très courtes, pour une bonne intelligibilité. Ils effectuent une première conversion de voix femme/homme en transformant en moyenne certains paramètres : facteur de modification du pitch et de la largeur de l'impulsion glottique, compression/dilatation de l'enveloppe spectrale pour les segments voisés et correction de gain.

En 1988, Abe et al. [11] définissent une procédure de transformation qui met l'accent sur l'étape d'apprentissage : enregistrement d'un même jeu de mots par plusieurs locuteurs, marquage et alignement des segments audio des différentes voix par programmation dynamique (ou *dynamic time warping*, DTW), extraction et comparaison des descripteurs de la voix retenus (enveloppe spectrale, pitch et énergie). L'apprentissage se fait par quantification vectorielle : un dictionnaire de traduction *(mapping codebook)* de ces descripteurs est créé à partir des segments alignés. Lors de la conversion, chaque segment de la voix source est transformé d'après les règles de correspondance établies dans le dictionnaire qui est défini sur les segments centroïdes les plus représentatifs. Cette démarche introduit une fonction de transformation dépendante du contexte (phonème prononcé), cependant le signal converti est généré à partir d'un nombre fixe d'enveloppes spectrales et le résultat n'est pas de très bonne qualité (à cause des discontinuités dans le signal) et même pour un grand nombre de classes (512, cf. [16]). Par ailleurs, les auteurs procèdent à une évaluation de leur procédure par des tests perceptifs qui encouragent les travaux

effecués.

En 1991, Abe [12] propose d'effectuer les transformations sur des unités acoustiques (les phonèmes) plutôt que sur des fenêtres de longueur fixe afin de prendre en compte les caractéristiques dynamiques de prononciation (trajectoire des formants).

En 1992, Valbret et al. [13] préfèrent utiliser la synthèse TD-PSOLA [8] plutôt que le modèle source-filtre car des transpositions de pitch et des modifications de l'échelle temporelle importantes sont permises tout en conservant une bonne qualité à l'écoute (avec certaines limitations comme nous pourrons le voir pas la suite). De plus, ils proposent deux nouvelles approches pour l'apprentissage :

– Le dictionnaire de traduction est optimisé par régression linéaire multiple afin de choisir la meilleure base des paires d'entrées-sorties sur l'ensemble des paires alignées.
– Les paires d'entrées-sorties sont considérées comme des classes (qui définissent le contexte de la transformation). À chaque classe est associée une fonction non-linéaire de compression/dilatation par morceaux des enveloppes spectrales (ou *dynamic frequency warping*, DFW) afin de projeter au mieux le relief (les formants) du spectre de la source sur celui de la cible.

Après évaluation perceptive des résultats, la première méthode donne des résultats plus proches de la cible, mais la seconde, si elle conduit à une voix de synthèse intermédiaire, génère moins d'artéfacts et paraît plus naturelle.

Au début des années 1990, l'équipe analyse/synthèse de l'IRCAM [14] [15] est conduite à proposer et à mettre en oeuvre un système de création de voix de castrat de grande qualité pour la bande son du film *Farinelli* de Gérard Corbiau (1994). Il s'agit de faire renaître une voix aux caractéristiques exceptionnelles : tessiture étendue, tenue des sons, caractère juvénile, puissant, agile et homogène sur l'ensemble de sa tessiture. La solution retenue est de créer une voix hybride à partir d'enregistrements d'un contre ténor et d'une soprano colorature. D'une part la voix du contre ténor est rendue plus juvénile (en retirant notamment les évènements bruités : souffle, rugosité, etc.), d'autre part la voix de soprano est transformée pour imiter le timbre du contre ténor (et étendre artificiellement son registre). Pour cette seconde étape, la fonction de transformation est effectuée uniquement sur les voyelles (les plus présentes dans le registre visé) et dépend du contexte (cinq filtrages différents suivant la voyelle chantée).

En 1995, Stylianou et al. [16] tirent profit des progrès récents en reconnaissance de locuteur par l'utilisation de modèles statistiques pour la classification des enveloppes spectrales. Ils inversent le problème et utilisent cette classification en noyaux de gaussiennes pour créer une fonction de transformation continue : les discontinuités dues à la quantification vectorielle disparaissent et améliorent le naturel des voix synthétisées. Pour valider cette nouvelle méthode, ils mesurent une nette diminution de la distortion spectrale entre voix convertie et voix cible par rapport aux approches précédentes. Il faut cependant entraîner les algorithmes d'apprentissage sur de longs enregistrements alignés des différentes voix, ce qui se révèle coûteux en temps de calcul et en organisation.

Proposé par Childers en 1995 [17], l'excitation glottique fait partie des descripteurs appris et transformés dans la démarche présentée par Arslan et Talkin en 1997 [18]. Une procédure robuste est proposée en 1999 par Arslan [25], qui fait la synthèse des recherches précédentes en utilisant comme modèle spectral les paires de lignes spectrales aux bonnes propriétés d'interpolation, et propose une modification contextuelle de la prosodie par l'application d'un rythme local de prononciation suivant le phonème.

En 2001, Kain et Macon [26] mettent l'accent sur l'évaluation des transformations par la constitution de corpus (répertoires des enregistrements sur lesquels sont entraînés les algorithmes d'apprentissage). Ils proposent également de prédire (au lieu de transformer) l'excitation glottique (dont l'estimée est l'erreur résiduelle d'un modèle LPC) en s'appuyant sur l'assomption que dans un contexte donné (la classe du phonème prononcé) les résidus sont similiares.

Si cet historique n'est pas exhaustif, il est représentatif des directions prises par les chercheurs, notamment en fonction des avancées dans d'autres recherches connexes : l'amélioration des techniques de synthèse pour la création de voix plus naturelles et pour un meilleur contrôle des transformations, la reconnaissance automatique de locuteur et de texte [34] [35] et leurs outils statistiques, l'alignement, etc. L'essor des performances de l'outil informatique n'est pas non plus à négliger lorsqu'il s'agit d'analyser des corpus rassemblant plusieurs heures d'enregistrement.

Au final, la transformation de voix s'inscrit dans une volonté générale de manipuler des paramètres de synthèse de haut niveau, adaptés à une manipulation instinctive par l'homme : le sens (synthèse à partie de texte), l'identité du locuteur, son état émotionnel, etc.

1.2 Applications

En s'inscrivant dans une démarche d'amélioration qualitative de la synthèse vocale, de l'expressivité et du contrôle, la transformation de voix a de nombreuses applications, commerciales et artistiques :

- Traitements numériques sur les voix parlées ou chantées, notamment au cinéma : coloration des voix de doublage par le timbre des voix originales, effets spéciaux (transformation femme-homme), voix hybrides, etc.
- Enrichissement du timbre et des comportements prosodiques des synthétiseurs vocaux.
- À l'inverse : adaptation de voix pour la reconnaissance vocale.
- De manière générale, applications aux systèmes qui utilisent des enregistrements audio (monophoniques) : imitation, variations, morphing.

On peut d'ailleurs citer un certain nombre de produits issus de l'industrie musicale et largement distribués qui témoignent d'un marché réel et qui peut tendre à se développer avec l'apparition de nouvelles possibilités et de meilleures performances :

• Le VT-1 Voice Transformer de BOSS Instruments (commercialisé en 1996) est un effet numérique qui permet de manipuler indépendamment le pitch du son monophonique en entrée et un décalage constant de ses formants pour générer un nouveau signal. Son utilisation est destinée aux stations de radio pour la création de nouvelles voix, ainsi qu'aux DJ et aux musiciens (harmonisation vocale, manipulation des sons, etc.).

• Le VoicePrism de TC-Helicon (commercialisé en 2000) reprend le principe du VT-1 et en étend les possibilités dans une optique dédiée à l'harmonisation vocale sur scène : polyphonie (synthèse de choeurs), programmation MIDI des paramètres (formants, transposition suivant un contexte harmonique local), autres effets (reverb, délai, etc.).

• Le Voice Modeler de TC-Helicon (commercialisation attendue) est un plug-in destiné à modifier les caractéristiques timbrales d'enregistrement de voix suivant certains paramètres haut-niveau : raucité, souffle, corps, etc.

Par ailleurs, le Vocaloid de Yamaha (commercialisation prévue pour 2004) est un plug-in de chant virtuel synthétisé à partir de samples obtenus après segmentation des enregsitrements de vrais chanteurs. Pour changer de qualité vocale il est nécessaire d'acheter de nouvelles librairies de samples. On voit donc que cette technologie de synthèse par concaténation d'unités bénéficierait de la souplesse permise par des modules temps-réel de transformation de voix.

1.3 L'équipe Analyse/Synthèse

L'équipe Analyse/Synthèse, dirigée par Xavier Rodet qui m'a encadré durant mon stage, développe des procédés de synthèse et de transformation des sons. Elle jouit d'une longue expérience en modélisation des signaux sonores (sinusoïdale, source-filtre, SINOLA, PSOLA, granulaire, TFD à court-terme, par modèles physiques, etc.).

Ses travaux de recherche ont donné lieu à la conception et au développement d'outils de traitement sonore pour les compositeurs et les musiciens : Additive (analyse/synthèse par modélisation sinusoïdale), Super Vocodeur de Phase (modélisation par TFD à court-terme) et son interface graphique AudioSculpt, Chant (modélisation granulaire par formes d'ondes formantiques), Diphone (environnement graphique de contrôle pour l'hybridation de sons à partir de différentes techniques d'analyse/synthèse), différents portages sur MacIntosh sous la forme de patch Max/MSP, etc.

Parmi les recherches actuellement menées au sein de l'équipe, on peut citer la synthèse par concaténation d'unités, la séparation de sources, la caractérisation et

la classification des sons par des descripteurs de haut niveau (projet CUIDADO), l'aide à l'analyse musicale (reconnaissance d'extraits audio, recherche de structures, résumés sonores, alignement de partition, etc.).

1.4 Stage : objectifs et moyens

Etant donné les différentes techniques mises en jeu (analyse/synthèse, alignement, apprentissage) et les différentes possibilités de mises en œuvre pour chacune d'entre elles, nous avons décidé d'effectuer des expérimentations en Matlab en utilisant au mieux les développements déjà disponibles dans l'équipe Analyse/Synthèse de l'IRCAM.

Les principales ressources logicielles que j'ai pu utiliser sont les suivantes :

XSpect pour la visualisation, l'écoute et l'analyse (spectre, LPC, f0) de fichiers sons. Cet outil permet la validation des expérimentations, par comparaison des voix converties et des voix cibles.

Super Vocodeur de Phase je l'ai utilisé pour des opérations élémentaires bien adaptées à la transformation de voix (pitch-shift, time-stretch, rééchantillonnage et filtrage).

f0 pour l'estimation de la fréquence fondamentale.

Analyse PSOLA positionnement de marqueurs synchrones au picth, calcul du voisement, etc. Les paramètres issus de cet analyse sont alors transformés avant resynthèse.

Divers codes Matlab développés pour d'autres projets (calcul des MFCC, du DTW, synthèse PSOLA).

En utilisant cet environnement, j'ai eu à proposer une procédure permettant de comparer différentes transformations suivant les paramètres d'analyse transformés, et ce pour différents cas (transformation homme-femme, femme-enfant). J'ai effectué les étapes d'apprentissage sur des extraits audio brefs (une dizaine de secondes) plus souples pour les expérimentations, et je n'ai donc pas exploré les solutions relatives à l'apprentissage sur des corpus alignés rassemblant plusieurs heures d'enregistrement. Ce point sera d'ailleurs discuté dans la partie 3 (Méthodes de synthèse vocale) du chapitre suivant.

Chapitre 2

Descripteurs et modification de la voix

2.1 La production de parole

La parole peut être décrite comme le résultat de l'action volontaire et coordonnée d'un certain nombre de muscles. L'appareil respiratoire fournit l'énergie nécessaire à la production des sons, en poussant l'air à travers la trachée. Au sommet de celle-ci se trouve le larynx où la pression de l'air est modulée avant d'être appliquée au conduit vocal (Fig. 2.1).

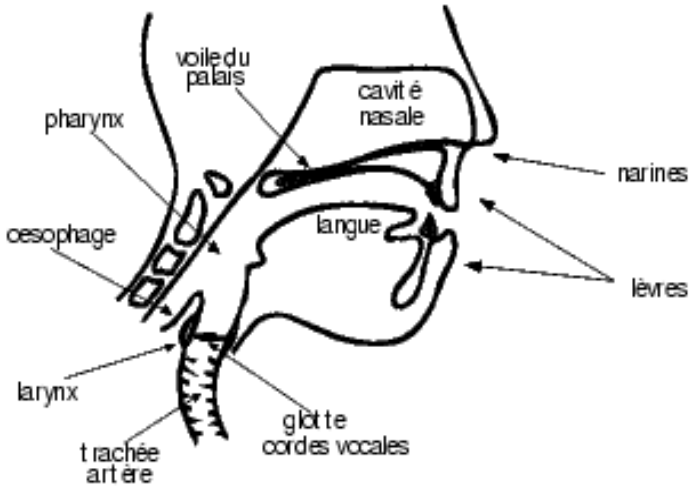


FIG. *2.1:* L'appareil phonatoire

Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée. Les cordes vocales sont deux lèvres symétriques placées en travers du larynx (Fig. 2.2). Elles peuvent fermer

complètement le larynx et, en s'écartant progressivement, déterminer une ouverture triangulaire (la glotte). L'air y passe librement pendant la respiration ou la voix chuchotée, ainsi que pendant la phonation des sons non voisés. Les sons voisés résultent au contraire d'une vibration périodique des cordes vocales. Le larynx est d'abord complètement fermé, ce qui accroît la pression en amont des cordes vocales et les force à s'ouvir, ce qui fait tomber la pression et permet aux cordes vocales de se refermer. Des impulsions périodiques de pression sont ainsi appliquées au conduit vocal, composé des cavités pharyngienne, buccale et nasale. On remarque donc l'adéquation du modèle source-filtre fréquemment utilisé en synthèse vocale.

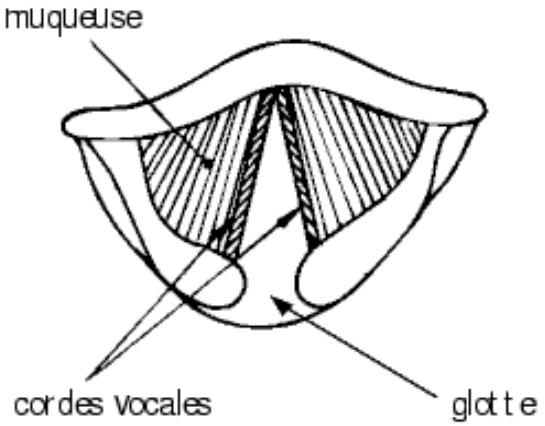


FIG. *2.2:* Le larynx, vu du dessus

Les sons émis sont caractérisés par un grand nombre de paramètres [38] [39] :
- le type de phonation (le mode de vibration des cordes vocales) : son voisé ou non voisé, chuchoté, crié, soufflé, etc.
- le lieu d'articulation (la région de rétrécissement maximal du canal buccal) suivant le position du larynx, du voile du palais, de la langue, des mâchoires, des dents et des lèvres. Plus généralement, une configuration géométrique donnée de ces éléments fixe certaines résonances (les formants) qui permettent de distinguer les sons voisés entre eux.
- le mode d'articulation (pour les sons non voisés) : occlusif ou plosif (le passage de l'air est fermé et le son résulte de son ouverture subite, par exemple pour les consonnes [p], [b], [t]), fricatif (le passage se rétrécit mais n'est pas interrompu, par exemple [f], [s], [v]) et d'autres sous-catégories.
- le caractère nasal suivant la position du voile du palais qui enclenche ou non la résonance de la cavité buccale.

De nombreux facteurs de variabilité dans le contrôle articulatoire du conduit vocal sont donc responsables du caractère unique et reconnaissable d'un locuteur particulier. Cependant d'autres composantes que la physiologie d'un individu interviennent

dans la caractérisation d'une voix comme nous le décrivons dans la partie suivante.

2.2 L'identité du locuteur

Le signal acoustique d'une voix parlée contient différents types d'information : le message (ce qui est dit), des informations propres au locuteur (qui l'a dit) et à l'environnement (où, quand, comment cela a été dit, enregistré). Pour une transformation de voix, nous désirons modifier les attributs relatifs au locuteur en préservant ceux porteurs du message et ceux témoins de l'environnement. Ces attributs spécifiques du locuteur peuvent être groupés en plusieurs niveaux :

Phonématique *(segmental)* regroupe l'ensemble des facteurs définissant la qualité d'une voix : son timbre. Celui-ci dépend des propriétés physiologiques et physiques de l'organe vocal du locuteur, ainsi que de son état émotionnel. On pourra par exemple décrire le timbre d'une voix selon [37] son caractère nasal, bruité, rauque, etc. D'un point de vue signal, on observe le timbre à partir des représentations spectrales des enregistrements (cf. section 2.3) à court terme (position, amplitude et largeur de bande des formants) et à long terme (couleur globale).

Prosodique *(suprasegmental)* correspond aux composantes de l'expression et du style, c'est-à-dire l'intonation et l'accent, qui dépendent des conditions sociales et psychologiques du locuteur. Au niveau du signal, la prosodie correspond à la hauteur du son, à l'énergie et à la durée des phones et des silences, observés à la fois en moyenne (hauteur et volume moyens, taux de prononciation) et dans leur évolution dynamique (contours).

Linguistique inclue la volonté et l'identité sociale du locuteur dans sa manière de s'exprimer : le choix des mots au niveau sémantique, lexical, syntaxique ou pragmatique [38] et selon une prononciation personnelle (liaisons).

Il existe bien sûr des dépendances entre ces différents étages. Par exemple, la prosodie est en moyenne caractéristique d'un individu mais dépend localement d'un contexte linguistique. Nous serons pourtant amenés à les considérer indépendamment lors de la phase d'apprentissage, et nous ne prendrons pas en compte le niveau de description linguistique : les transformations envisagées se font en conservant un texte identique entre la voix source et la voix transformée.
Les descripteurs retenus pour modéliser, apprendre et transformer une voix sont donc ses caractéristiques spectrales et prosodiques. Nous présentons comment les manipuler dans les sections suivantes de ce chapitre. Nous commençons par présenter les différentes méthodes - ou stratégies - usuelles en synthèse vocale suivant un champ d'application bien précis. Une fois ce domaine spécifié, nous ferons un panorama des différentes techniques (un savoir-faire s'appliquant à la description d'un modèle de signal théorique) de synthèse vocale et mettrons en avant la solution retenue : analyse/synthèse PSOLA (addition-recouvrement synchrone au pitch).

2.3 Méthodes de synthèse vocale

Différents niveaux de synthèse vocale peuvent être envisagés [19] suivant qu'elle s'effectue à partir de :

Concepts ce niveau intervient dans des problématiques de dialogue homme-machine. Le texte à transmettre est lui-même synthétisé avant le signal sonore le réalisant.

Texte ce niveau regroupe les systèmes *text-to-speech*.

Texte phonétique il n'y a plus à ce niveau d'analyse linguistique : les entrées du système sont une séquence phonétique et la prosodie associée.

Paramètres les paramètres de commandes sont envoyés directement au synthétiseur, ils sont donc directement dépendants de la technique de synthèse choisie (cf. partie suivante) et ne sont plus lisibles ou intelligibles.

On peut passer d'un niveau à l'autre en employant différentes analyses (linguistique, phonétique, signal sonore), et on peut donc envisager un synthétiseur à partir des paramètres comme un sous-module d'un synthétiseur à partir de texte phonétique, lui-même sous module d'un synthétiseur à partir de texte, etc.

En se situant en-dessous du niveau conceptuel (qui a cependant une grande importance pour la génération de prosodie) les technologies développées en synthèse vocale peuvent être classées suivant deux grandes catégories de méthodes :

Synthèse par règles cette approche est fondée sur un modèle paramétrique du signal vocal et sur un ensemble de règles gouvernant l'évolution temporelle de ces paramètres. Par exemple, dans le modèle source-filtre, on manipule l'excitation (la forme et la fréquence de l'impulsion, le bruit) et les formants (la réponse du filtre). Cette méthode est particulièrement bien adaptée au cas où une analyse a déjà été effectuée et que l'on dispose d'une évolution temporelle cohérente des paramètres. On peut alors les modifier et effectuer une resynthèse. Cependant, il est important de noter que pour des applications de synthèse à partir de texte (sans analyse du son préalable), il est difficile de définir des règles d'interpolation des trajectoires réalistes des divers paramètres entre des zones stables de prononciation (milieu d'un phonème). S'il est possible de générer un signal compréhensible, la coarticulation n'est pas de bonne qualité et il est très difficile d'obtenir une voix cohérente et naturelle.

Synthèse par concaténation d'unités acoustiques ici, l'effort d'élaboration des règles est remplacé par le stockage et la classification d'un répertoire de segments de parole élémentaires[1] extraits d'enregistrements d'un locuteur réel. L'opération de reconstruction consiste à concaténer au mieux la séquence appropriée de ces unités. On distingue deux approches qui peuvent coexister :

[1] Un signal de voix peut être découpé suivant différentes unités acoustiques élémentaires : phones (une réalisation d'un phonème), diphones (du milieu d'un phone au milieu du phone suivant), triphones (groupement de deux diphones), etc.

- Chaque unité n'est disponible qu'une fois dans la base de données et on modifie ses caractéristiques prosodiques grâce à une technique d'analyse/synthèse (additive par exemple pour le projet MBROLA [40]), avant d'effectuer un lissage acoustique des discontinuités. Ce système souple requiert une segmentation et un stockage d'environ 1500 diphones pour la langue française. Cette méthode permet une bonne intelligibilité et un naturel acceptable mais perfectible : la parole est hyper-articulée et souffre d'une certaine pauvreté expressive (le signal généré est caricatural).
- La sélection s'effectue dans une grande base de données où on trouve plusieurs instances de chaque unité dans des contextes prosodiques différents. Il s'agit ici de choisir au mieux chaque unité en minimisant un coût de synthèse global qui tient compte d'un coût de représentation (dans quelle mesure les segments choisis correspondent-ils au contexte phonétique et prosodique dans lequel on les insère ?) et d'un coût de concaténation (dans quelle mesure la juxtaposition des segments choisis amène-t-elle des discontinuités ?). Les résultats obtenus peuvent alors être de grande qualité mais nécessitent des enregistrements longs et soignés, une segmentation automatique très précise ainsi qu'un accès très rapide à plusieurs gigaoctets de données, tout cela pour un seul modèle de voix.

À ce point, on peut également envisager deux stratégies de transformation de voix : l'apprentissage se fait sur de grands corpus pour bénéficier d'un grand nombre de contextes de prononciation (la transformation tendrait à devenir un synthétiseur par concaténation d'unités adapté à la voix cible) ou sur des corpus plus petits sur lesquels on désire extraire une quantité d'information réduite mais pertinente. Cette deuxième solution, bien adaptée à la synthèse par règles, est celle que j'ai suivie pour des raisons de souplesse et de cohérence avec la durée du stage de DEA. Elle s'inscrit par ailleurs dans la volonté de ne pas constituer de nouvelles bases de données quand on cherche à imiter une nouvelle voix.

2.4 Techniques d'analyse/synthèse

Afin d'estimer et de manipuler les caractéristiques spectrales et prosodiques d'un signal de voix nous disposons de différentes techniques d'analyse/synthèse. Sans essayer d'en dresser une classification, il est important de noter que chacune s'applique à un modèle de signal cohérent avec les paramètres qu'elle manipule [9] :

modèle	**technique**
TFD à court terme	vocodeur de phase
mélanges de sinusoïdes et bruit	additive
harmonicité/bruit et enveloppe spectrale	source-filtre
formes d'onde élémetaires	PSOLA, formes d'onde formantiques
mécano-acoustique du conduit vocal	modèle physique

On peut ainsi manipuler des signaux de voix en utilisant des modélisations différentes. Il est possible d'en estimer les enveloppes spectrales (Fig. 2.3) :

• en joignant les sommets de chaque raie spectrale, obtenues après une analyse additive ou une TFD effectuée sur quelques périodes du signal.

• en l'associant à la TFD de deux périodes de signal (marquées par une analyse PSOLA) pondérées par une fenêtre de hanning (cf. [5] et [8] pp.10-11). Le fenêtrage de cette forme d'onde entraîne un léger lissage des résonances étroites, mais reste acceptable pour les formants d'une voix.

• en l'associant à la réponse en fréquence du filtre auto-régressif correspondant à une analyse LPC.

Prenons un exemple sur un signal de voix :

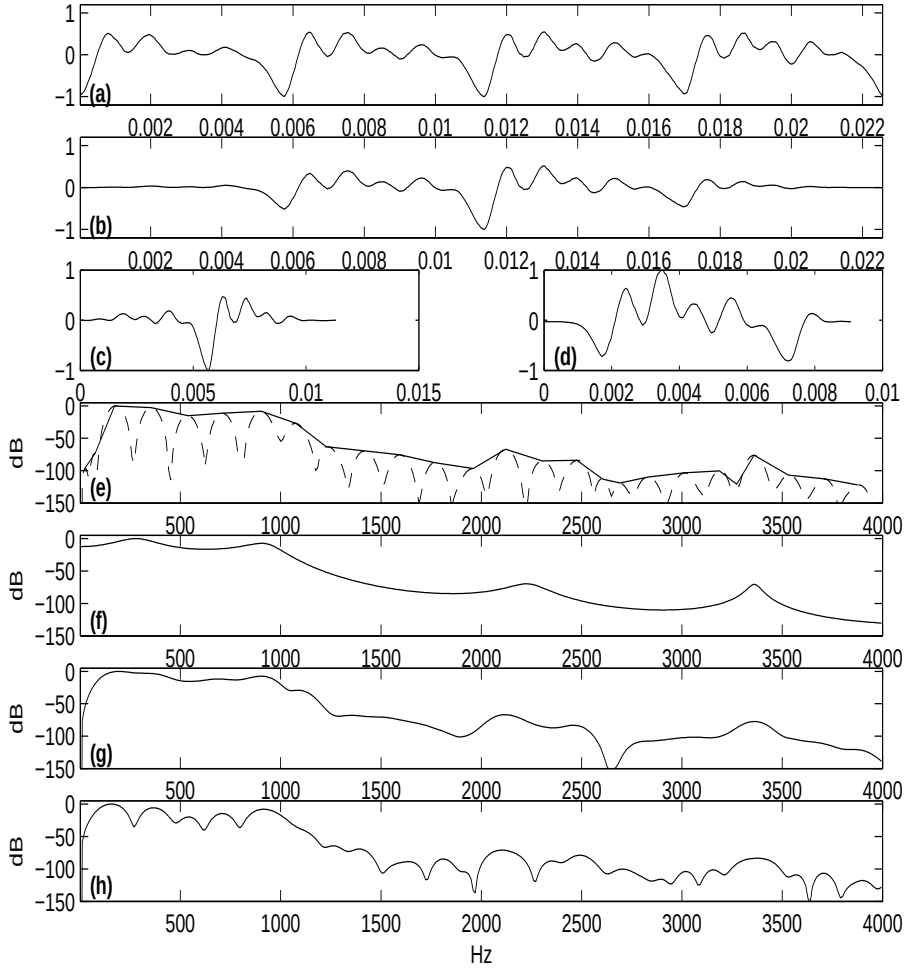


FIG. *2.3:* Estimations d'enveloppes spectrales

Détaillons les différentes partie de la figure 2.3 :

– les signaux temporels (en secondes) :
  (a) signal original de 4 périodes.
  (b) même signal pondéré par une fenêtre de hanning.
  (c) 2 périodes pondérées par une fenêtre de hanning centrée sur un maximum d'énergie.
  (d) 2 périodes pondérées par une fenêtre de hanning non centrée sur un maximum d'énergie.

– les estimées de l'enveloppe spectrale :
  (e) sommets des raies après TFD sur le signal (b).
  (f) réponse du filtre auto-régressif d'ordre 50 calculé sur le signal (b).
  (g) TFD sur le signal (c).
  (h) TFD sur le signal (d).

Tout d'abord remarquons que le signal voisé de la figure 2.3 est marqué par de fortes concentrations locales d'énergie (comme décrit dans la partie 1 de ce chapitre) correspondant aux instants de fermeture de la glotte (l'impulsion dans un modèle source-filtre).

Quand on désire estimer l'enveloppe spectrale en analysant une petite portion de signal (deux périodes), il est donc important de considérer la position de la fenêtre de hanning (Fig. 2.4) par rapport à ces pics d'énergie :

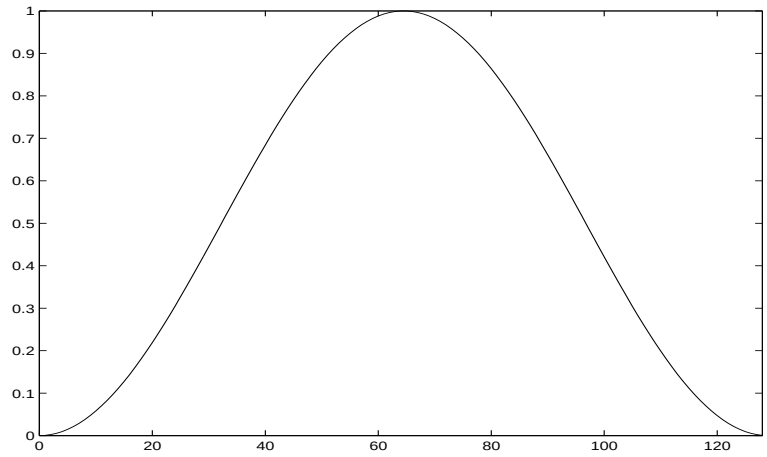


FIG. *2.4:* Fenêtre de hanning sur 128 points

1. lorsqu'elle est centrée sur un pic (c), celui-ci n'est guère détérioré (zone stable de la fenêtre). Les pics précédent et suivant sont pondérés par des valeurs proches de zéro (extrémités de la fenêtre). Ainsi on analyse la contribution d'une seule réponse impulsionnelle faiblement distordue.

2. losqu'elle contient deux pics (d), ceux-ci sont sont détériorés par des régions à forte pente de la fenêtre de hanning. De plus la périodisation du signal entraîne l'apparition des pics spectraux après TFD, on est donc loin d'une bonne estimation de l'enveloppe spectrale.

Ainsi, un marquage du signal temporel centré sur les maxima d'énergie permet de passer aisément dans le domaine des enveloppes spectrales et donc d'effectuer des opérations sur celles-ci. Ce marquage est effectué par le logiciel d'analyse PSOLA développé par Geoffroy Peeters et Joseph Escribe, et c'est donc la solution que nous avons choisie pour réaliser les transformations.

Il était aussi possible d'utiliser une technique source-filtre basée sur un modèle auto-régressif (dont les anti-formants sont cependant absents), ou les techniques additive ou vocodeur de phase (pas spécialement adaptées à la voix et maniant un grand nombre de paramètres).

Nous présentons donc dans la partie suivante la technique PSOLA qui, comme nous allons le voir, est également efficace pour certaines modifications de prosodie.

2.5 PSOLA

La technique PSOLA est bien adapté au signaux comme celui de la voix où l'interprétation en tant que répétition d'une forme d'onde est possible [8]. Le but de l'analyse est d'effectuer un fenêtrage exactement centré sur les périodes fondamentales du signal. Le signal de synthèse est alors reconstitué par superposition addition *(overlap-add)* de ces formes d'onde élémentaires.

Les opérations d'analyse à effectuer pour le marquage du signal sont les suivantes :

Détection des singularités on recherche les concentrations temporelles importantes de l'énergie du signal local : les instants de ferneture de la glotte ou les transitoires indépendantes d'une excitation périodique (dues aux plosives par exemple).

Calcul de la fréquence fondamentale l'estimation de la fréquence fondamentale du signal est utile aux étapes suivantes et peut être effectué par un module extérieur (logiciel f0 dans mon cas).

Voisement on détecte ici le caractère voisé/non voisé des fenêtres d'analyse suivant un degré de confiance sur l'harmonicité des signaux par exemple.

Détection des transitoires on sépare ici les impulsions périodiques des transitoires, par exemple en se servant de la cohérence avec la période fondamentale trouvée.

Placement des marques PSOLA dans les régions non voisées et transitoires le marquage se fait suivant un pas constant puisque la fréquence fondamentale n'est pas définie (il n'y a donc pas de synchronie possible). Pour les régions voisées le marquage s'effectue suivant le respect de deux contraintes : le centrage sur les maxima d'énergie (pour ne pas détériorer le signal après fenêtrage

de hanning par exemple), le respect d'une distance entre deux fenêtres proche de la période fondamentale. Ces deux contraintes sont spécifiées car elles peuvent être localement contradictoires dans le cas de légères irrégularités des instants de fermeture de glotte.

A la synthèse, il suffit alors pour modifier la hauteur d'un son de changer la distance entre deux formes d'onde élémentaires. Pour changer le déroulement de l'axe temporel on peut également supprimer ou répéter les formes d'onde.

Ces deux opérations peuvent être effectuées indépendamment et il est donc possible de jouer sur deux composantes prosodiques (hauteur et durées à la resynthèse). Cependant des modifications trop grandes peuvent altérer le naturel du signal synthétisé, par exemple :

– dans le cas d'une dilatation des durées la répétition des formes d'onde peut conduire (malgré une interpolation) à un signal trop stable et peu naturel.
– réduire la hauteur en grande proportions peut conduire à disjoindre les formes d'ondes élémentaires et à créer des trous dans le signal.

En résumé, PSOLA nous permet d'effectuer aisément des modifications de prosodie, des opérations sur les enveloppes spectrales (en passant par la TFD des formes d'onde élémentaires). L'étape critique est ici l'analyse où la précision du marquage dépend notamment de l'estimation de la fréquence fondamentale.

Chapitre 3

L'alignement

La fonction de transformation que nous désirons définir se fait par un apprentissage des comportements de deux voix sur les mêmes phonèmes prononcés. Nous présentons dans ce chapitre les solutions proposées pour effectuer la mise en correspondance de ces phonèmes, présents dans le signal sous la formes d'évènements acoustiques différents (les phones).

3.1 La programmation dynamique

La programmation dynamique est fréquemment utilisée dans les opérations de recherche d'un chemin où les décisions consécutives des directions à suivre (ou des états à choisir) dépendent les unes des autres et dont la séquence globale (le trajet parmi tous les états) doit conduire à un résultat optimal suivant certains critères à minimiser [32].

Dans le cas de l'alignement de séquences temporelles suivant leurs similitudes, l'approche de la programmation dynamique est non temps-réel puisqu'elle prend en compte le passé et le futur de chaque état intermédiaire. On peut illustrer cette opération par la mise en correspondance des sommets et des vallées de deux reliefs similaires (Fig. 3.1).

La programmation dynamique est notamment utilisée en reconnaissance vocale (sous la dénomination *dynamic time warping*) afin de s'affranchir du rythme de prononciation particulier d'un locuteur. Dans notre cas, cet algorithme est utilisé à des fins similaires : avant l'apprentissage on cherche à établir des correspondances entre les phonèmes identiques prononcés par la source et la cible mais présents dans le signal acoustique enregistré sous la forme de phones différents à des temps différents.

Nous présentons ici le principe de l'algorithme, en désignant par S et C les séquences temporelles à aligner (avec pour illustration la figure 3.2) :

$$S = s_1, s_2, ..., s_n$$
$$C = c_1, c_2, ..., c_m$$

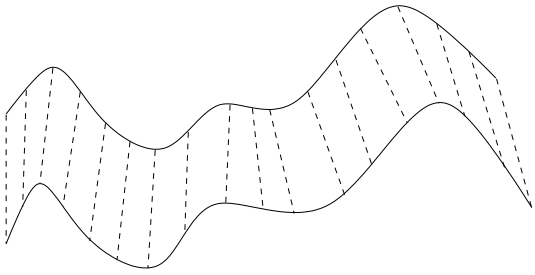


FIG. *3.1:* Une utilisation possible du DTW : les séquences mesurent la position d'un geste de la main capté suivant une direction de l'espace. On désire reconnaître ces gestes en comparant leur forme générale avec un dictionnaire de gestes de référence qui ne sont pas forcément à la même échelle.

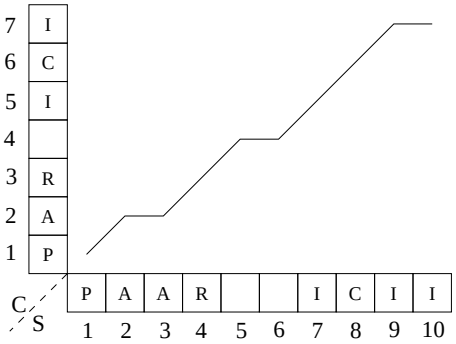


FIG. *3.2:* Le bon chemin : $(1,1)(2,2)(3,2)(4,3)(5,4)(6,4)(7,5)(8,6)(9,7)(10,7)$

Tout d'abord il convient de projeter les points s_i et c_j dans un espace où on peut définir une distance bien adaptée aux modèles de ces séquences (des portions de signal audio dans notre cas). On construit alors une matrice de taille $n * m$ dans laquelle l'élément (i, j) contient la distance $d(s_i, c_j)$ entre les deux points s_i et c_j.
Un chemin W se définit par un ensemble de couples $w_k = (i_k, j_k)$ où $(1 \leq k \leq K)$ qui vérifie les contraintes suivantes :

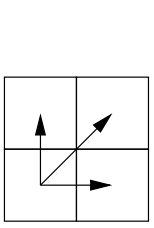
Conditions limites $w_1 = (1, 1)$ et $w_K = (n, m)$.

Monotonie pour $w_k = (a, b)$ et $w_{k+1} = (\alpha, \beta)$ on a $a \leq \alpha$ et $b \leq \beta$.

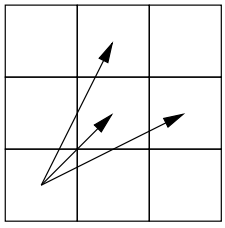
Voisinage autorisé seulement certains déplacements entre w_k et w_{k+1} sont autorisés.

Sur la figure 3.3, nous voyons qu'un voisinage fixe un type d'incrément possible qui doit être cohérent avec les connaissances a priori sur les signaux à aligner : le cas (a) est utile quand des sauts sont à prévoir (une séquence avance sans l'autre, e.g. silences décalés dans le signal audio), le (b) empêche ces sauts de mener à un chemin trivial, le (c) sert dans un contexte particulier (une séquence avance toujours plus vite que l'autre), etc.

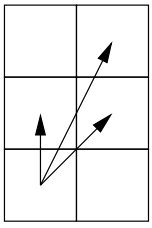
Parmi l'ensemble des chemins W, on veut alors trouver celui qui minimise le poids



(a) les directions horizontales, verticales et diagonales sont autorisées



(b) seules les diagonales sont autorisées



(c) parcours asymétrique

FIG. *3.3:* Exemples de voisinages

du trajet :

$$DTW = min_W \sqrt{\frac{1}{K} \sum_{(i,j) \in W} d(s_i, c_j)}$$

Le facteur $\frac{1}{K}$ sert ici à compenser des longueurs de chemin différentes.

L'algorithme de programmation dynamique effectue cette optimisation par récurrence en utilisant la distance cumulative $\delta(i,j) = d(s_i, c_j) + min(\delta(\hat{i}, \hat{j}))$ où $(\hat{i}, \hat{j})$ représente l'ensemble des prédecesseurs possibles de (i,j). Par exemple, pour le voisinage (b) on a :

$$\delta(i,j) = d(s_i, c_j) + min(\delta(i-1, j-1), \delta(i-1, j-2), \delta(i-2, j-1))$$

Ainsi, en partant de (s_1, c_1) pour aller jusqu'à (s_n, c_m) et en passant par tous les couples (i,j), on construit une matrice des distances cumulatives.

On obtient alors le DTW en faisant un trajet retour - l'algorithme est fondamentalement non temps-réel - de (s_n, c_m) à (s_1, c_1) en choisissant à chaque fois le couple du voisinage précédent autorisé qui est associé à la distance cumulative la plus petite.

L'implémentation que j'ai utilisée prévoit quelques modifications :

– afin d'éviter un temps de calcul trop long, on préfère calculer la matrice des distances cumulatives sur une région limitée dans un voisinage de la diagonale (la diagonale représentant le cas particulier où les deux séquences sont reconnues comme étant parfaitement synchronisées, à un facteur d'échelle près) ou d'une autre région à définir. Par ailleurs, le fait de préciser cette région des chemins acceptables en fonction de connaissances a priori sur les signaux empêche d'obtenir un DTW grossièrement faux.
– une pondération des directions peut être incluse dans le calcul des distances cumulatives. Par exemple, dans le cas (a) de la figure 3.3, la distance

$$\delta(i,j) = d(s_i, c_j) + min(\delta(i-1, j-1), \delta(i, j-1), \delta(i-1, j))$$

est remplacée par

$$\delta(i,j) = d(s_i, c_j) + min(p_{11} * \delta(i-1, j-1), p_{01} * \delta(i, j-1), p_{10} * \delta(i-1, j))$$

Cette pondération permet de favoriser certains trajets attendus.

En résumé, il m'a donc été possible d'adapter le calcul du DTW aux signaux audio en définissant un certain nombre de paramètres :

1. l'espace de projection des signaux et la distance associée.
2. les voisinages d'incrément autorisés.
3. la région de calcul des distances cumulatives.
4. la pondération des directions d'incrément.

En ce qui concerne les trois derniers points, différentes possibilités ont été testées pour essayer d'évaluer la meilleure combinaison (pour un exemple donné). Nous présentons en détail dans la partie suivante le choix d'une distance adaptée aux signaux sonores numérisés.

3.2 La distance

Afin de comparer de manière automatique des signaux de voix, il est habituel (notamment en reconnaissance de texte) d'utiliser une distance euclidienne sur les coefficients cepstraux sur l'échelle mel ou MFCC *(mel-frequency cepstrum coefficients)* qui tient compte des particularités de l'oreille humaine.

L'échelle mel prend en compte une contribution prépondérante des basses fréquences (échelle quasi-linéaire en basses fréquences, logarithmique en hautes fréquences) :

$$mel(f_{hz}) = 2595 * log_{10}(1 + \frac{f_{hz}}{700})$$

La calcul des MFCC s'effectue ainsi [33] :

1. on choisit le nombre M de coefficients MFCC, typiquement de l'ordre de 10 à 20 pour la voix.
2. on définit $N > M$ filtres triangulaires d'importance énergétique équivalente (et donc de largeur de bande variable) sur l'échelle mel.
3. on pondère la portion de signal par une fenêtre d'analyse (hanning) pour minimiser la distortion spectrale avant d'en calculer la TFD.
4. on calcule les énergies du signal analysé X_k $(k = 1..N)$ en sortie des filtres mel.
5. on considère la séquence X_k comme un signal temporel discret dont les ondulations (les formants) sont décrites par les coefficients cosinusoïdaux de la série de Fourier de ce signal :

$$MFCC_i = \sum_{k=1}^{N} X_k * cos[i(k - \frac{1}{2})\frac{\pi}{N}]; i = 1..M$$

6. on récupère les M premiers coefficients $MFCC_i$. En effet, les premiers coefficients prennent en compte les reliefs basse fréquence du spectre (l'enveloppe spectrale), quand les coefficients d'ordre supérieur décrivent des variations plus rapides dépendantes du pitch et des raies spectrales.

Par ailleurs, l'utilisation des dérivées du premier et du second ordre des MFCC (les DMFCC et DDMFCC) peut permettre d'améliorer une caractérisation objective d'une portion de signal vocal par la prise en compte de paramètres dynamiques : l'évolution temporelle de l'enveloppe spectrale.

On dispose alors d'une projection du signal sur la base de ses MFCC, DMFCC, DDMFCC. On peut effecuer des pondérations de ces coefficients (notamment augmenter l'importance du premier coefficient MFCC mesurant l'énergie du signal) pour trouver le meilleur alignement (cf. partie Résultats).

3.3 Problématique

Malgré la bonne caracérisation permise par les MFCC et l'efficacité du DTW, il est important de noter que si l'on pouvait aligner parfaitement deux voix, cela signifierait qu'on saurait les comparer et les mettre en correspondance, c'est-à-dire de passer de l'une à l'autre. Or c'est justement le point de départ de notre étude : l'apprentissage des différences entre ces voix.

On peut alors se douter que l'alignement va souffrir de certains évèmenents : la prononciation d'un même texte peut être interprétée différemment suivant le locuteur (respect des liaisons) et la suite d'évènements acoustiques n'est donc plus la même, les réalisations acoustiques d'un même phonème peuvent être spectralement trop éloignées pour être synchronisées, le chemin peut prendre localement des raccourcis aberrants en sautant une portion de signal gênante, etc. Avant d'effectuer l'apprentissage il sera donc important de sélectionner des zones de confiance où l'on est sûr que l'alignement est adéquat.

3.4 Résultats

Comme nous l'avons détaillé, différents paramétrages du calcul du DTW sont possibles. Après expérimentations, nous avons choisi d'effectuer le calcul des MFCC sur des fenêtres de 30 ms avec recouvrement (pas d'incrément de 5ms). Pour une voix d'homme descendant jusqu'à 100 hertz on est alors sûrs d'avoir des fenêtres d'analyse contenant à chaque fois quelques périodes (au moins 3 de 10 ms chacune) ce qui garantit une certaine stabilité.

En effet, puisqu'on avance à pas constant (5 ms) non synchronisé sur le pitch, il est nécessaire d'avoir au moins plusieurs périodes de signal sur chaque région d'analyse pour que le fenêtrage de hanning n'ait pas trop d'influence suivant sa position par rapport aux pics d'énergie.

Les signaux vocaux sont donc projetés sur une base de MFCC (20 coefficients) toutes les 5ms et les correspondances de l'alignement se font selon ce pas d'échantillonnage.

Nous discutons maintenant des autres possibilités de paramétrage suivant le voisinage d'incrément. Les figures 3.4 et 3.5 illustrent l'alignement de deux voix (une femme et un homme) prononçant 'paf' (issu de 'pas fait') pour deux voisinages différents mais avec une même distance euclidienne sur les MFCC.

Dans le cas d'un voisinage autorisant les directions horizontales et verticales (sauts) et la diagonale (cas (a), Fig. 3.3) on observe les particularités suivantes (Fig. 3.4) :
– la mise en correspondance de certaines zones est très précise (zones transitoires : plosives, début d'un mot après un silence) : placement des marques 266, 269.
– d'autres zones ne sont pas marquées (entre les points 267 et 268).

Étant donné la possibilité pour l'algorithme d'effectuer des déplacements horizontaux et verticaux, il effectue des sauts dans le marquage. Ceux-ci peuvent avoir plusieurs conséquences : l'algorithme est robuste face aux longs silences et aux rythmes de prononciation très différents, mais en contre-partie, il a tendance à suivre localement des chemins aberrants. En effet, on observe sur la figure 3.4 que les zones voisées du [a] des deux voix sont très différentes, il saute donc cette région avant de se recaler à l'apparition du [f]. On peut alors interpoler entre les marqueurs présents, mais quand ceux-ci sont trop peu nombreux la précision est faible.

Dans le cas d'un voisinage autorisant trois directions diagonales aux pentes différentes (cas (b), Fig. 3.3), l'algorithme est obligé de poser des marqueurs régulièrement (Fig. 3.5) malgré des distances entre les deux voix qui peuvent être localement importantes. Le chemin est donc plus homogène et il n'est pas nécessaire d'interpoler entre les marqueurs. Cependant, dans le cas (que nous avons évité) où les deux signaux comprennent de longs silences décalés entre les mots prononcés, le chemin ne peut pas prendre une pente assez raide et une erreur se propage.

Ainsi on a préféré utiliser un algorithme n'effectuant pas de sauts de marquage. Les DMFCC et DDMFCC ont également été pris en compte dans le calcul de la distance, mais cela n'a pas eu beaucoup d'influence sur les résultats observés (99,7% des marques identiques).

3.5 Alignement de texte

Dans le cadre d'un projet de l'équipe Analyse/Synthèse visant à recréer artificiellement la voix de Jean Cocteau, j'ai été amené à tester le découpage automatique de longs enregistrements en diphones. Le but de l'opération est de créer un corpus contenant un grand nombre de réalisations de chaque diphone pour ensuite procéder à une synthèse par concaténations d'unités acoustiques (cf. partie 2.3).

Certaines instances des diphones ont été repérées à la main par écoute et observation des signaux, mais ce travail fastidieux n'a pas été mené jusqu'à obtenir au moins

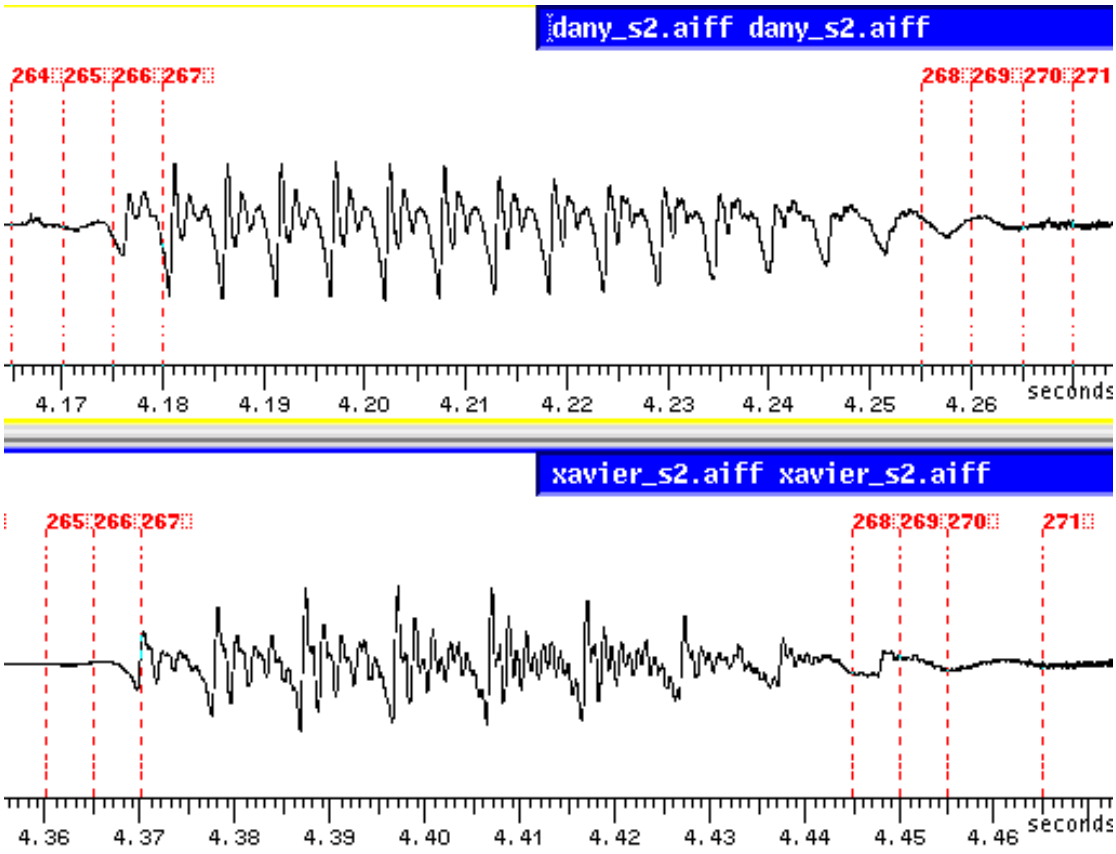


FIG. *3.4:* Alignement avec sauts sur deux voix prononçant 'paf'

une instance de chaque diphone de la langue française. On a alors complété artificiellement le corpus à développer en concaténant des moitiés de phones (35 pour 35 phonèmes) pour imiter les diphones restants.

Dans la solution envisagée on dispose d'un extrait à découper (la source) et de sa transcription en diphones. À partir de cette transcription on créé un second signal (la cible) par concaténation des diphones et faux diphones (deux moitiés de phones) du corpus. On aligne alors la source et la cible et, connaissant le découpage en diphones de la cible on peut reporter les temps de début, milieu et fin de chaque diphone sur la source d'après le DTW trouvé. On peut observer un exemple de découpage sur la figure 3.6. La plupart des découpages sont précis mais il existe des erreurs notamment parce que le premier corpus contient des fausses réalisations de diphones et que la qualité de l'alignement n'est pas facile à évaluer. L'algorithme doit donc être amélioré pour détecter les découpages erronés et affiner localement les autres.

Le découpage automatique n'est donc pas parfaitement réalisé par le prototype présenté, mais c'est une bonne piste à suivre pour éviter le découpage à la main de nombreuses heures d'enregistrement de voix. Par ailleurs, il est important de noter que le procédé s'améliore naturellement à chaque itération : le corpus s'agrandit et les faux diphones (deux moitiés de phones) sont enlevés petit à petit.

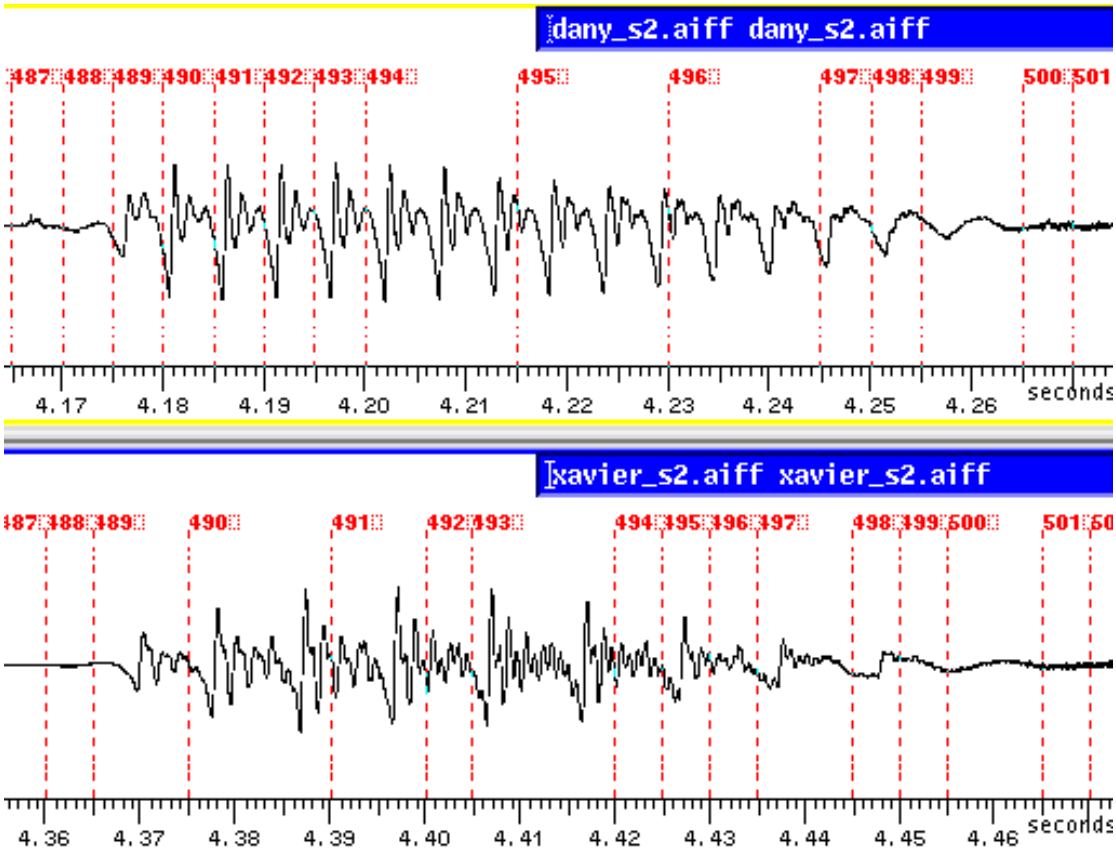


FIG. *3.5:* Alignement sans saut sur deux voix prononçant 'paf'

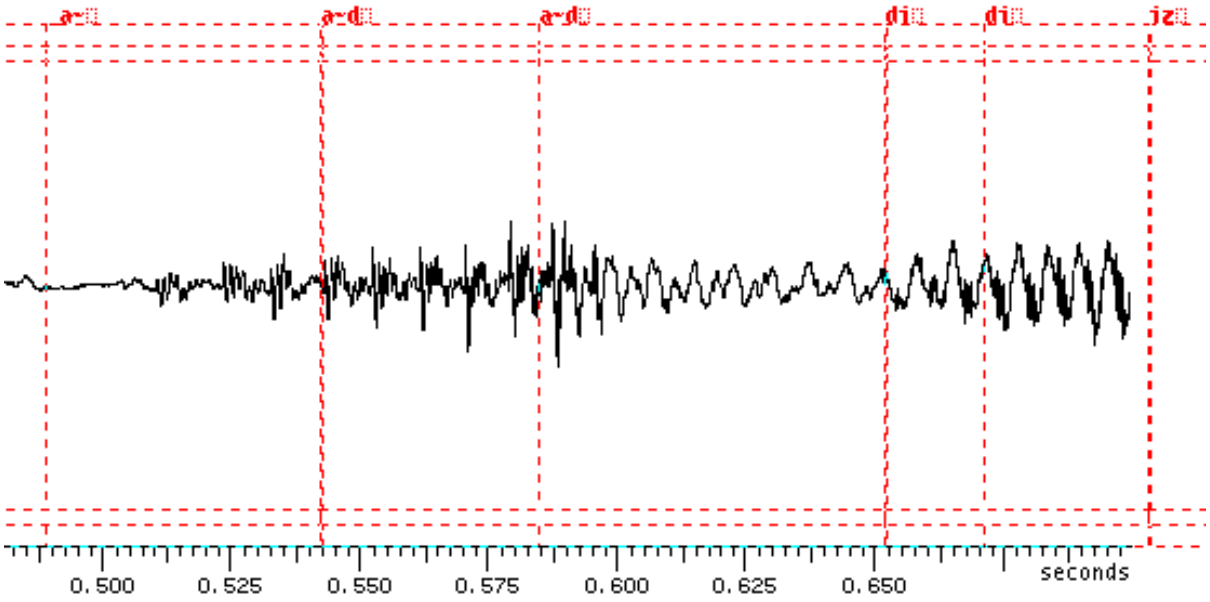


FIG. *3.6:* Alignement d'un extrait audio avec sa succession de diphones

Chapitre 4

L'apprentissage

Notre objectif est ici de définir des fonctions de transformation génériques, c'est-à-dire qui s'appliquent au mieux à toutes les instances présentes dans un signal vocal.

4.1 Choix des phonèmes caractéristiques

Une fois l'alignement effectué, nous disposons d'un grand nombre d'information pour effectuer des comparaisons. Cependant celle-ci doit être triée pour ne pas entraîner l'apprentissage sur des régions où l'alignement peut se révéler imprécis (notamment dans le cas où les prononciations sont très éloignées entre les locuteurs). On sélectionne donc des portions de signal à la fois sûres et caractéristiques des deux voix, c'est-à-dire une zone de stabilité au centre des phones. On cherche donc (Fig. 4.1) à repérer des portions de signal qui correspondent conjointement :

– à un minimum du flux spectral du signal pour repérer des portions proches d'un régime établi, en mesurant les variations de l'amplitude de la TFD à court terme d'une fenêtre d'analyse à la suivante.
– à un maximum d'énergie du signal au dessus d'un seuil (fixé à -20 dB après expérimentations) pour sélectionner des portions représentatives.

Les deux conditions n'étant que rarement remplies en même temps, on autorise un décalage possible (fixé à 100 ms).

Une fois cette première sélection faite, il convient de ne retenir que les fenêtres d'analyse caractéristiques (appelées désormais phones caractéristiques) des deux voix qui sont également en correspondance d'après l'alignement trouvé (avec un autre décalage permis de 100 ms). Pour un extrait de 5 secondes nous en repérons ainsi une dizaine sur lesquelles nous entraînons l'algorithme d'apprentissage.

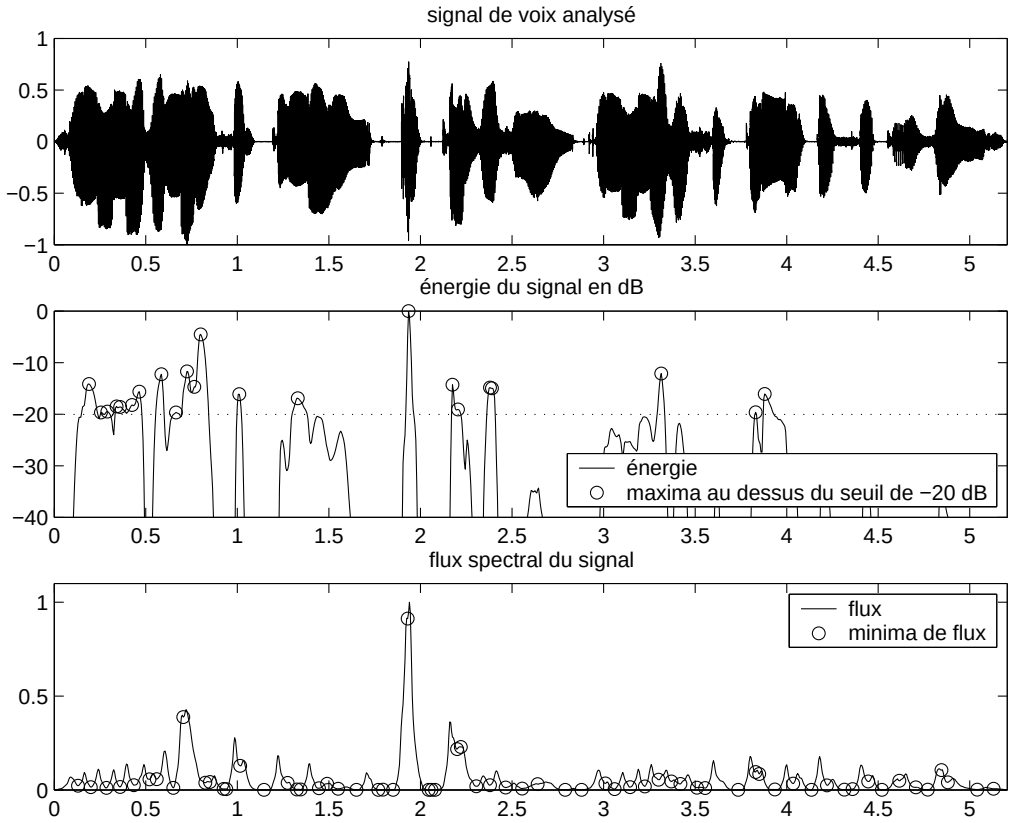


FIG. *4.1:* Calcul des zones stables et caractéristiques d'un signal vocal

4.2 Dilatation/compression constante des enveloppes spectrales

Les transformations que j'ai testées se font principalement dans le cas d'une transformation de genre (dans les deux sens femme-homme et homme-femme). Les facteurs discriminants les plus importants sont alors [28] la fréquence fondamentale, la position et la largeur de bande des formants. Les formants dans un signal de voix féminine sont plus hauts en fréquence en raison d'un conduit vocal plus court (la position des cordes vocales est moins profonde dans la gorge). Il est donc naturel d'envisager la transformation spectrale comme une compression (femme vers homme) ou une dilatation (homme vers femme) de l'axe fréquentiel.

Pour estimer ce facteur de dilatation/compression on effectue une minimisation sur les fenêtres d'analyse sig_{source} et sig_{cible} :

$$\alpha = min_a \sum_{f=0Hz}^{f=3000Hz} \|Env(sig_{source}, f) - a * Env(sig_{cible}, f)\|$$

où $Env(sig, f) = 20 * log_{10}(|TFD_{sig}(f)|)$ pondéré par la réponse moyenne de l'oreille. On tronque la somme des distances à 3000 Hz pour favoriser les régions formantiques.

On utilisera à la synthèse (cf. partie 4 du chapitre suivant) la moyenne des coefficients α sur l'ensemble des phones caractéristiques comme facteur d'homothétie sur les enveloppes spectrales.

On trouve sur les exemples étudiés (femme vers homme) $\alpha = 0.76$.

4.3 Dilatation/compression par morceaux des enveloppes spectrales

On veut ici tester une modification plus précise des enveloppes spectrales par dilatation/compression par morceaux de l'axe fréquentiel (DFW), afin de s'approcher mieux de la voix cible spécifique.

On utilise alors l'algorithme de programmation dynamique pour la mise en correspondance des reliefs des enveloppes, en appliquant une distance sur les pentes des enveloppes. On peut observer le résultat sur un phone caractéristique sur la figure 4.2. Durant la transformation on applique donc la fonction de dilatation/compression par morceaux de l'axe fréquentiel sur les enveloppes spectrales avant resynthèse. N'ayant pas eu le temps de définir une fonction dépendante du contexte, j'ai appliqué un DFW moyen en repérant à la main les correspondances les plus souvent répétées sur l'ensemble des phones caractéristiques. On peut alors observer le DFW retenu (spécifique aux voix analysées) sur la figure 4.3.

4.4 Filtrage à long terme

Les deux types de transformation envisagées ont pour effet de déplacer les résonances des enveloppes spectrales, mais elles modifient également la répartition énergétique des voix sur l'axe fréquentiel. Par exemple, une compression constante des enveloppes (dans le sens femme-homme) a pour conséquence d'appauvrir les hautes fréquences et de conférer à la voix une couleur inhabituelle.

En comparant les TFD à long terme de la voix transformée et de la voix cible (Fig. 4.4) on définit le gabarit d'un filtre qui va permettre de corriger le timbre de la voix transformée afin d'en améliorer le naturel.

Dans le cas de la figure 4.4 on applique un gain de 3 dB à partir de 1000 Hz, 10 dB à partir de 15000 Hz, 20 dB au-delà de 18000 Hz.

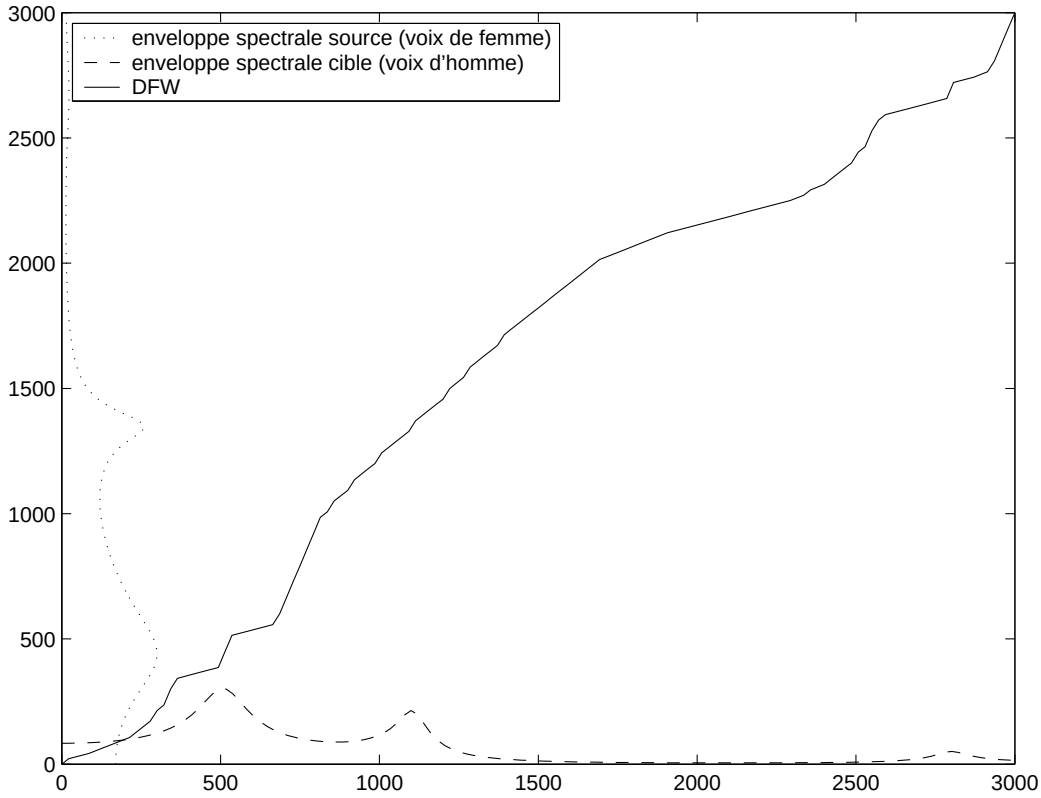


FIG. *4.2:* Chemin d'alignement du relief des enveloppes (DFW) entre 0 et 3000 Hz

4.5 Modification de hauteur

Pour les transformations de genre envisagées, on évalue un facteur de multiplication moyen des fréquences fondamentales en comparant les rapports de fréquence sur les phones caractéristiques. Sur les mêmes exemples de voix illustrés dans les parties précédentes on trouve un rapport moyen de 0.65 dans le sens femme-homme.

On ne modifie donc pas le contour général de la hauteur (la mélodie) mais seulement son échelle. Une extension des transformations envisagées devrait prendre en compte un modèle d'apprentissage des différentes composantes prosodiques. Il est clair que cette piste devrait permettre des améliorations perceptives très importantes.

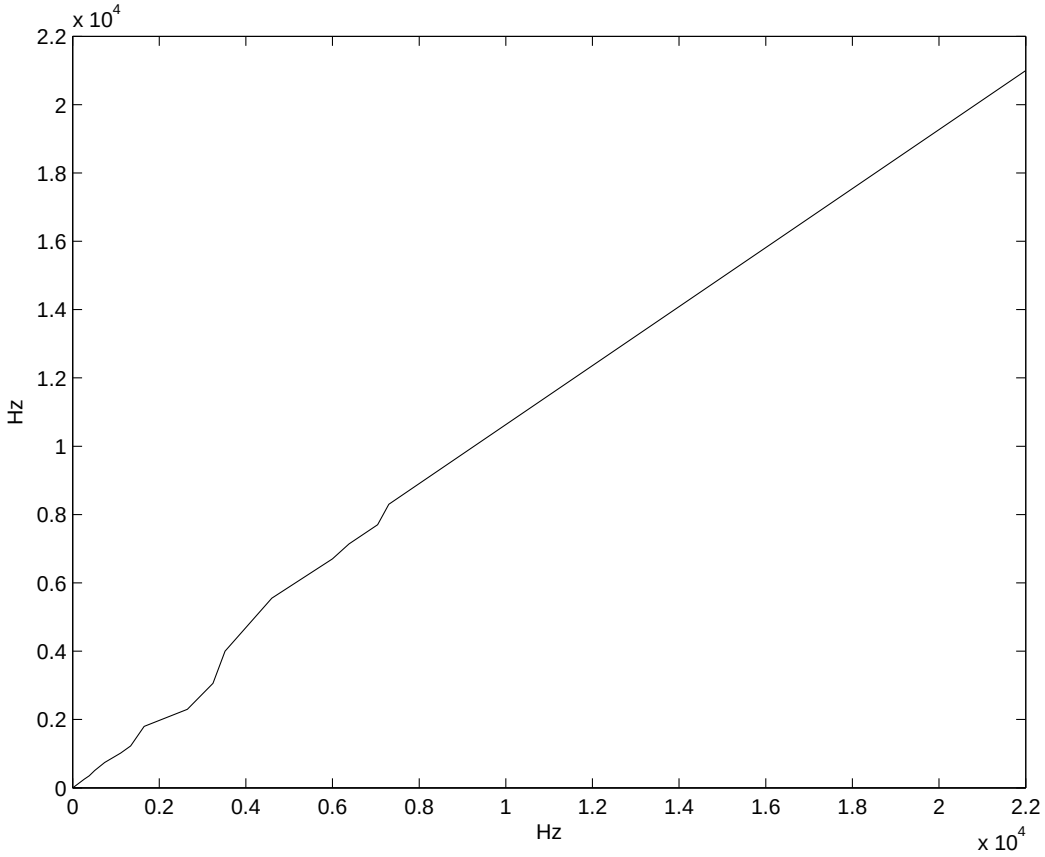


FIG. *4.3:* Correspondances moyennes des enveloppes spectrales (voix féminine en ordonnée, masculine en abscisse)

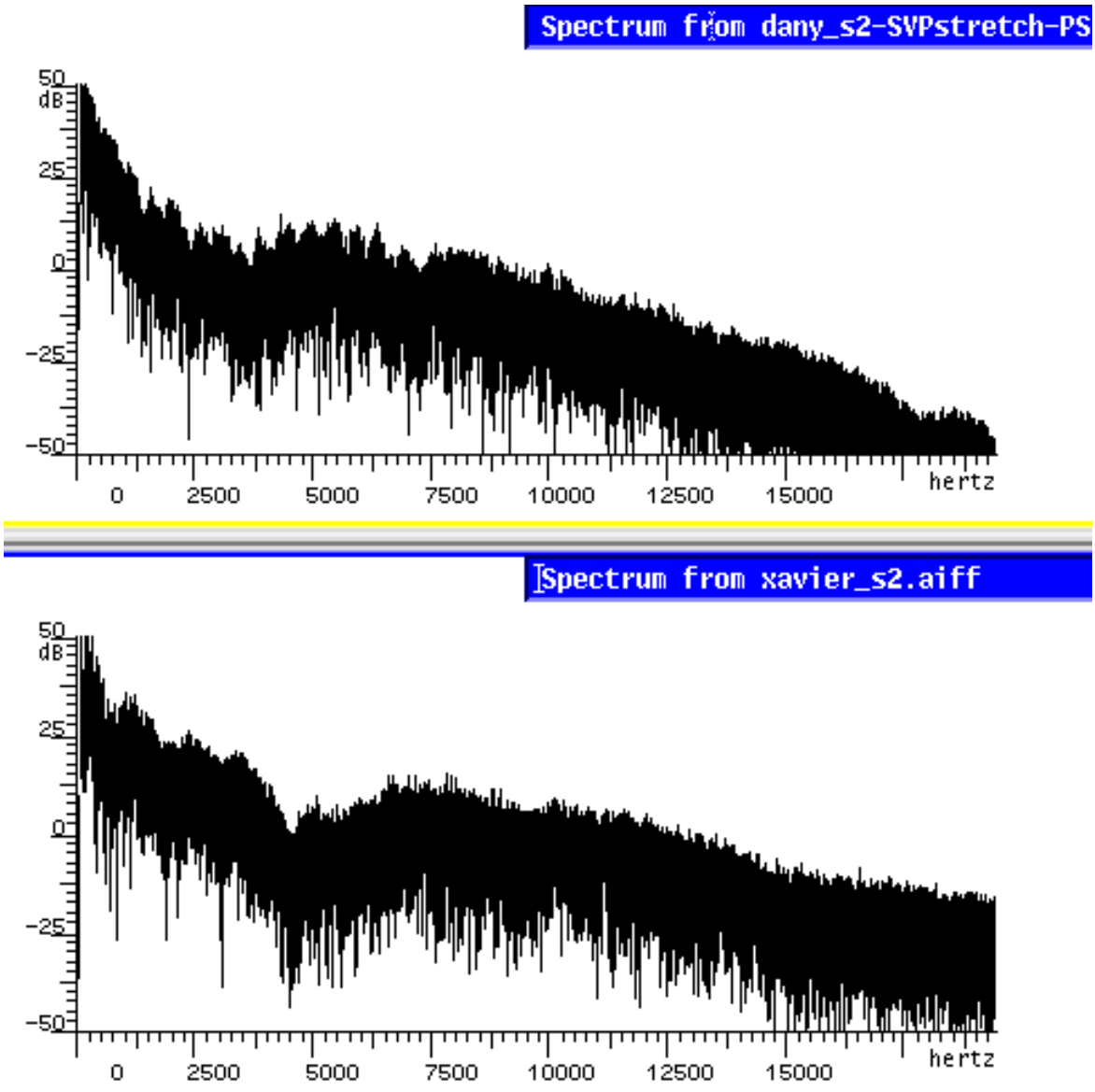


FIG. *4.4:* Les TFD à long terme (sur 5 secondes d'enregistrement) de la voix source transformée (en haut) et de la voix cible (en bas)

Chapitre 5

La procédure de transformation

Dans cette partie nous situons les étapes détaillées dans les chapitres précédents au sein d'une procédure complète de transformation de voix. Nous présentons en fin de chapitre les implémentations réalisées pour la gestion des différents modules, puis les différentes opérations de synthèse effectuées.

5.1 Point de départ : l'enregistrement

Comme nous l'avons détaillé précédemment, les corpus utilisés sont de durées assez courtes (quelques phrases enregistrées pour chaque voix), à la fois pour des raisons pratiques et pour se placer dans des conditions où l'on dispose d'un nombre réduit d'information sur les voix à transformer.

Six voix ont été enregistrées (un enfant, une femme, quatre hommes) en studio d'enregistrement à l'IRCAM sur le texte suivant :

> «Je connais très bien mon temps. Ne jamais travailler demande de grands talents. Il est heureux que je les ai eu.
> Je n'en aurais manifestement eu aucun besoin, et n'en aurais certainement pas fait usage, dans le but d'accumuler des surplus, si j'avais été originellement riche, ou si même j'avais au moins bien voulu m'employer dans un des quelques arts dont j'étais peut-être plus capable que d'autres, en consentant une seule fois à tenir le moindre compte des goûts actuels du public.
> Ma vision personnelle du monde n'excusait de telles pratiques autour de l'argent que pour garder ma complète indépendance ; et donc sans m'engager effectivement à rien en échange.»
>
> Messieurs Halgand, *Nous autres*.

Dans la procéduire présentée, l'apprentissage se fait sur une partie du texte enregistré (la phrase «Je n'en aurais manifestement eu aucun besoin, et n'en aurais certainement pas fait usage »), et la transformation peut être testée sur l'ensemble de l'enregistrement.

5.2 La procédure

Nous résumons ici l'ensemble des opérations à réaliser après l'enregistrement des voix :

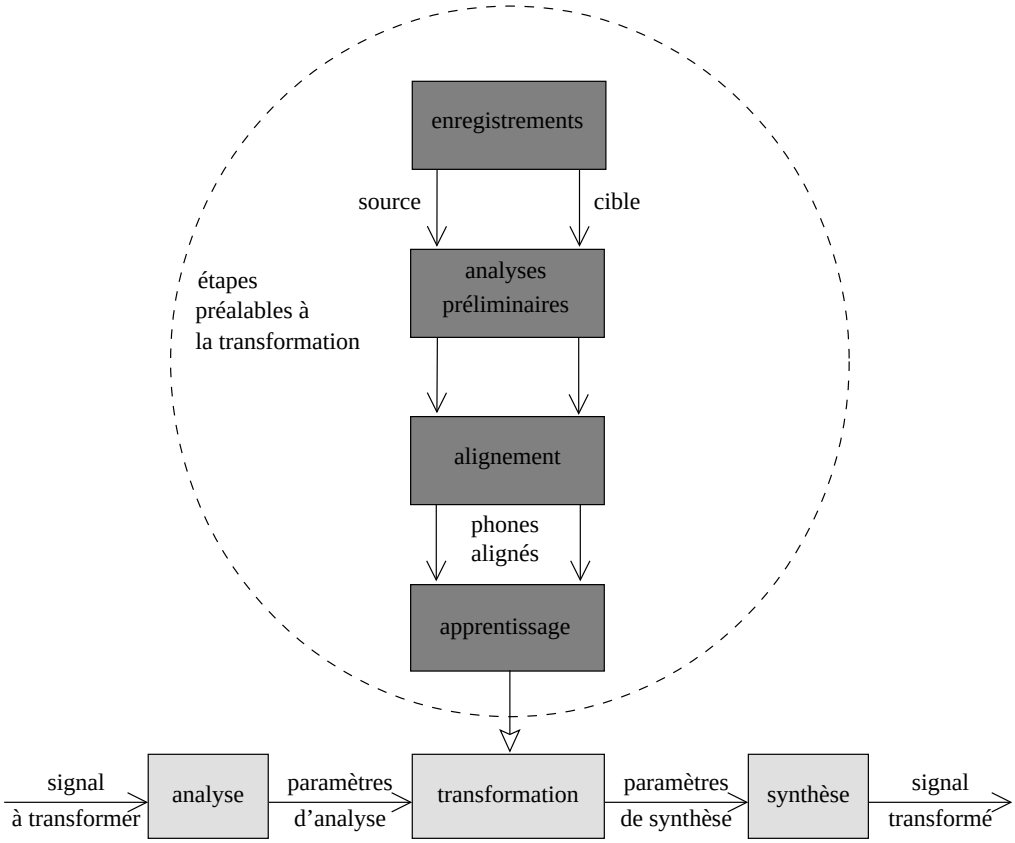
Analyses préliminaires ces analyses servent à l'alignement (calcul des MFCC) et au choix des zones stables des phonèmes caractéristiques pour l'apprentissage (calcul de l'énergie et du flux spectral des signaux).

Alignement renvoie les correspondances temporelles des signaux.

Apprentissage comparaison de certaines propriétés spectrales et prosodiques et définition des fonctions de transformation retenues pour passer le mieux possible d'une voix à l'autre.

Transformation à l'aide du Super Vocodeur de Phase et de l'analyse/synthèse PSOLA, on réalise les transformations définies dans l'étape précédente.

Nous pouvons illustrer la procédure en séparant l'établissement des connaissances et l'analyse/synthèse :



5.3 Implémentation

Certaines étapes de la procédure étant coûteuses en temps de calcul (calcul des MFCC et du DTW), il était important d'en sauvegarder les résultats pour ne pas perdre de temps à chaque fois qu'une nouvelle paramétrisation des autres étapes était prévue.

Ces paramétrisations ont du être testées suivant de nombreuses valeurs (combien de coefficients MFCC, quel type d'incrément autorisé pour le DTW, quel seuil de détection pour les phones caractéristiques, etc.) afin de choisir celles qui donnaient le meilleur résultat.

J'ai donc dès le départ centré mes efforts sur la gestion d'un grand nombre de fichiers d'analyse à lire ou à recalculer si la nouvelle paramétrisation le demande.

Les modules que j'ai moi-même développés sont les analyses de flux et d'énergie, la détection des phonèmes caractéristiques, l'apprentissage, et l'organisation générale du projet. En ce qui concerne les MFCC et le DTW, j'ai eu à faire quelques adaptations pour utiliser des codes déjà développés pour d'autres projets (CUIDADO, alignement de partitions).

5.4 La synthèse

Comme indiqué lors de la description de PSOLA, l'analyse donne un découpage des formes d'onde élémentaires du signal de la voix source. En faisant une TFD sur ces formes d'onde, on passe dans le domaine des enveloppes spectrales que l'on peut modifier suivant la fonction de transformation définie. Après une TFD inverse on créé une nouvelle forme d'onde qui est utilisée pour la synthèse PSOLA (Fig. 5.1).

La modification moyenne de hauteur est également effectuée lors de la synthèse PSOLA, et le filtrage à long terme destiné à améliorer le rendu naturel de la voix synthétisé est réalisé à l'aide du Super Vocodeur de Phase.

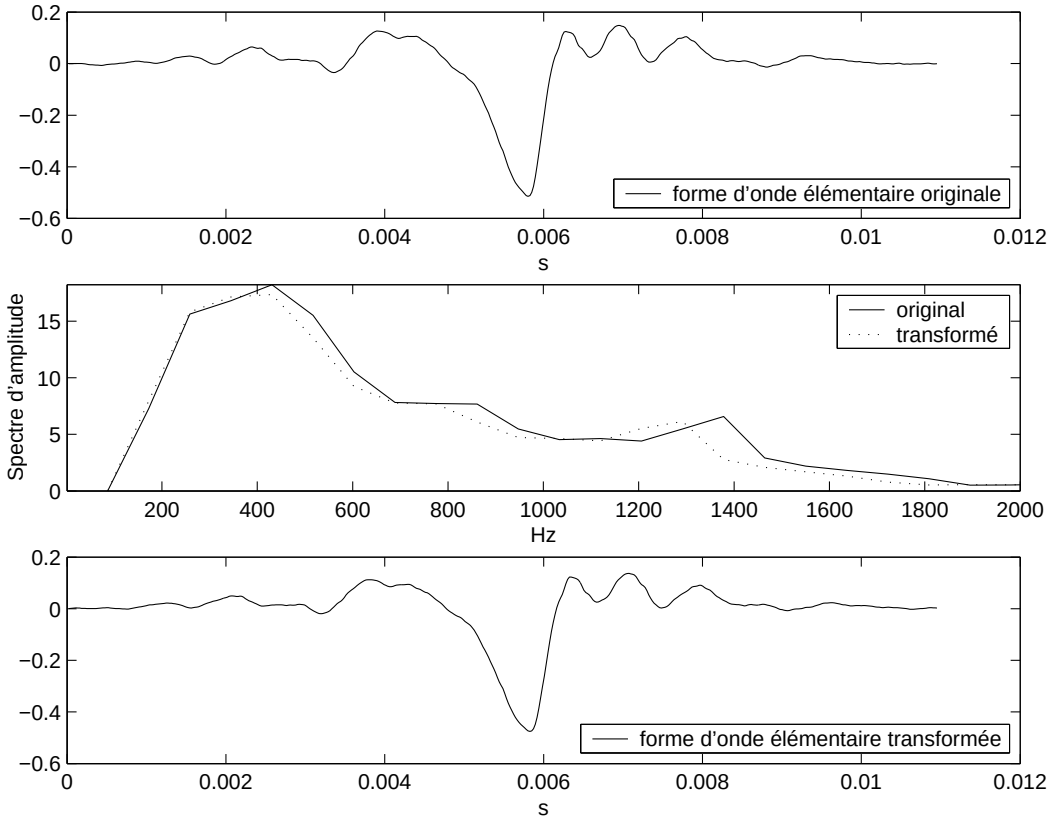


FIG. *5.1:* Modification des formes d'onde élémentaires avec PSOLA

Chapitre 6

Résultats et ouverture

6.1 Résultats de la synthèse

La procédure développée a été principalement testée sur des transformations de genre : dans ce cas les différences timbrales et prosodiques entre la voix source et la voix cible sont très importantes et les résultats de synthèse donnent des voix dont le caractère a été très largement modifié dans le sens désiré.

Certaines transformations auxquelles j'ai participé serviront d'ailleurs pour la bande son du film *Tiresia* de Bertrand Bonnello (sortie le 8 octobre 2003), qui désirait donner un caractère masculin à des enregistrements de voix féminines.

Dans certains cas, les transformations réalisées souffrent cependant d'artéfacts sonores à cause de changements de hauteur trop importants (limite de PSOLA), ou à cause d'analyses délicates. En effet, l'analyse de la fréquence fondamentale peut comporter des erreurs qui influent sur le placement des marqueurs PSOLA et peuvent générer des traitements inappropriés à la synthèse.

Pour les transformations femme-enfant et homme-homme, il est nécessaire d'affiner l'apprentissage suivant différentes pistes que nous discutons dans la partie suivante.

6.2 Améliorations possibles

Ce stage m'a permis d'explorer les nombreux domaines qui touchent à la transformation de voix : l'identité d'un locuteur au travers sa voix, l'alignement de signaux sonores, l'apprentissage, les stratégies et techniques de synthèse vocale, etc. J'ai pu développer une procédure faisant une synthèse des opérations nécessaires pour modifier l'identité perçue à l'écoute d'enregistrements sonores.

J'ai par ailleurs pu mettre en avant les pistes à suivre pour améliorer les résultats obtenus :

– prendre en compte le contexte (le phonème prononcé, les caractéristiques spectrales et prosodiques locales) dans les fonctions de transformation. Des analyses et un découpage automatique des signaux en diphones ou en phones doit alors être réalisé pour définir ce contete.

- travailler avec des modélisations spectrales permettant des interpolations sans artéfact (paires de lignes spectrales, paramètres PLAR [28]).
- améliorer la qualité sonore des voix transformées notamment avec une meilleure analyse de la fréquence fondamentale.
- établir des modèles de prosodie en séparant le contexte de prononciation (phrase affirmative, interrogative, etc.) des caractéristiques d'un locuteur spécifique et travailler sur une transformation des paramètres de ces modèles.
- effectuer des évaluations perceptives des transfomations sur différents critères (la qualité sonore de la voix de synthèse, son naturel, sa conformité à la cible), pour se rendre compte des directions de recherche à creuser.

Les résultats obtenus sont encourageants et les nombreuses pistes évoquées laissent entrevoir que des transformations de grande qualité sont possibles.

Conclusion

Le stage que j'ai effectué au sein de l'équipe Analyse/Synthèse de l'IRCAM m'a permis d'étudier différents domaines de recherche : les méthodes et techniques de synthèse vocale, la reconnaissance de locuteur, la reconnaissance de texte, l'alignement d'extraits audio, etc.

Autour du thème de la transformation de voix, il était important de dégager les dépendances de ces recherches entre elles, notamment pour spécifier un contexte d'application précis et pour en imaginer les moyens de mise en oeuvre : analyse/synthèse PSOLA, Super Vocodeur de Phase, caractérisation (et comparaison) des signaux par les MFCC, alignement par programmation dynamique, etc.

Les résultats très précis de l'alignement effectué conduisent à l'apprentissage d'une fonction de transformation des descripteurs caractéristiques d'une voix. Il devrait également permettre de découper automatiquement les signaux en diphones et donc de définir localement les contextes de prononciation.

La procédure proposée met en jeu de nombreux modules de traitement des sons et des analyses. Elle a été testée dans plusieurs cas, donnant les meilleurs résultats pour les transformations de genre. Elle pourra être réutilisée et améliorée suivant diverses pistes possibles dont la plus prometteuse est sans doute la modélisation et la transformation des comportements prosodiques d'un individu.

Bibliographie

Production de la parole

[1] G. Straka (1965). *Album phonétique*, Presses de l'Université Laval, Québec.

[2] G. Fant (1970). *Acoustic theory of speech production*, Mouton 1970 The Hague.

Traitement du signal et modèles

[3] J. D. Markel, A. H. Gray jr. (1972). *Linear prediction of speech*, Springer-Verlag.

[4] J.L. Flanagan (1972). *Speech analysis synthesis and perception*, Springer-Verlag.

[5] E. Moulines, F. Charpentier (1990). *Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones*, Speech Communication 1990.

[6] D. Schwarz (1998). *Spectral envelopes in sound analysis and synthesis*, mémoire de Master, Université de Stuttgart.

[7] M. Hasegawa-Johnson (2000). *Lecture notes in speech production, speech coding, and speech recognition*, University of Illinois.

[8] G. Peeters (2001). *Modèles et modification du signal sonore adaptés à ses caractéristiques locales*, Thèse de doctorat, Université Paris VI.

[9] P. Depalle (2002). Cours d'analyse/synthèse du DEA ATIAM, IRCAM, Paris.

Transformation de voix

[10] D. G. Childers, B. Yegnanarayana, K. Wu (1985). *Voice conversion : factors responsible for quality*, ICASSP 1985, pp. 748-751.

[11] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara (1988). *Voice conversion through vector quantization*, ICASSP 1988, pp. 655-658.

[12] M. Abe (1991). *A segment-based approach to voice conversion*, ICASSP 1991, pp. 765-768.

[13] H. Valbret, E. Moulines, J.P. Tubach (1992). *Voice transformation using PSOLA technique*, ICASSP 1992, vol. 1, pp. 145-148.

[14] P. Depalle, G. Garcia, X. Rodet (1994). *A virtual castrato (!?)*, ICMC 94, Aarhus, Danemark.

[15] P. Depalle, G. Garcia, X. Rodet (1995). *À la recherche d'une voix perdue*, Résonance n^{o} 8, mars 1995, IRCAM, Paris.

[16] Y. Stylianou, O. Cappé, E. Moulines (1995). *Statistical methods for voice quality transformation*, Proc. Eurospeech 1995 (Madrid, Espagne), pp. 447-450.

[17] D. G. Childers. *Glottal source modelling for voice conversion*, Speech Communication, 16(2), pp. 127-138.

[18] L. M. Arslan, D. Talkin (1997). *Voice conversion by codebook mapping of line spectral frequncies and excitation spectrum*, Proc. Eurospeech 1997 (Rhodes, Grèce), vol.3, pp 1347-1350.

[19] S. Santi (1997). *La synthèse vocale au sein du traitement automatique des langues : place, rôle et perspectives*, BULAG, Actes Du Colloques International FRACTAL 1997, Linguistique et Informatique : Théories et Outils pour le Traitement Automatique des Langues, pp 329-336.

[20] L. M. Arslan (1999). *Speaker transformation algorithm using segmental codebooks (STASC)*, Speech Communication, 28(3), pp 211-228.

[21] A. Kain, M. Macon (2001). *Design and evaluation of a voice conversion algorithm based on spectral enveloppe mapping and residual prediction*, Proc. ICASSP 2001.

[22] A. Kain (2001). *High resolution voice transformation*, Ph. D. Thesis, OGI school of Science and Technology.

[23] A. Kain, M. Macon (2001). *Design and evaluation of a voice conversion algorithm based on spectral enveloppe mapping and residual prediction*, Proc. ICASSP 2001.

[24] A. Kain (2001). *High resolution voice transformation*, Ph. D. Thesis, OGI school of Science and Technology.

[25] L. M. Arslan (1999). *Speaker transformation algorithm using segmental codebooks (STASC)*, Speech Communication, 28(3), pp 211-228.

[26] A. Kain, M. Macon (2001). *Design and evaluation of a voice conversion algorithm based on spectral enveloppe mapping and residual prediction*, Proc. ICASSP 2001.

[27] A. Kain (2001). *High resolution voice transformation*, Ph. D. Thesis, OGI school of Science and Technology.

[28] S. Corveleyn, B.Coose, W.Verhelst (2002). *Voice modification and conversion using PLAR-parameters*, MPCA-2002, Leuven, Belgium.

[29] W. Verhelst, H. Brouckxon (2002). *Voice modification for lip synchronization, voice dubbing and karaoke*, MPCA-2002, Leuven, Belgium.

[30] B. Gillett (2003). *Transforming voice quality and intonation*, MSc. Thesis.

[31] J.M. Gutiérrez-Arriola, J.M. Montero, J.A. Vallejo, R. Córdoba, R. San-Segundo, J.M. Pardo. *A new multi-speaker formant synthetizer that applies voice conversion techniques.*

Alignement

[32] E. J. Keogh, M. J. Pazzani. *Derivative dynamic time warping.*

[33] F. Zheng, G. Zhang., Z. Song (2001). *Comparison of different implementations of MFCC*, J. Computer Science and Technology, 16(6) :582-589, septembre 2001.

Divers

[34] T. Dutoit, L. Couvreur, F. Malfrère, V.Pagel, C. Ris (2002). *Synthèse vocale et reconnaissance de la parole : droites gauches et mondes parallèles*, Actes du 6è Congrès Français d'Acoustique, Lille.

[35] M. Ostendorf, I. Bulyko. *The impact of speech recognition on speech synthesis*, department of electrical engineering, University of Wahington, Seattle.

[36] H. Kopka, P. W. Daly (1999). *A guide to Latex*, Addison-Wesley.

Webographie

[37] *Étude du timbre de la voix*, site du LIMSI : `http ://www.limsi.fr/RS98FF/CHM98FF/TLP98FF/tlp3.html`

[38] Laboratoire de linguistique informatique Talana : `http ://talana.linguist.jussieu.fr`

[39] *Cours de phonétique*, section linguistique de l'université de Lausanne : `http ://www.unil.ch/ling/phon/index.html`

[40] The MBROLA project : `http ://tcts.fpms.ac.be/synthesis/mbrola.html`