Transformation de l'intonation : application à la synthèse de la parole et à la transformation de voix

Damien Lolive

N° d'ordre: 03801

THÈSE

présentée

devant l'Université de Rennes 1

pour obtenir

le grade de : Docteur de l'Université de Rennes 1
Mention Informatique

par

Damien Lolive

Équipe d'accueil : Cordial - IRISA
École Doctorale : Matisse
Composante universitaire : IFSIC

Titre de la thèse :

Transformation de l'intonation
Application à la synthèse de la parole et à la transformation de voix

soutenue le 27 novembre 2008 devant la commission d'examen

M. :	Christophe	D'Alessandro	Président
Mme :	Véronique	Aubergé	Rapporteurs
M. :	Thierry	Dutoit	
MM. :	Thierry	Moudenc	Examinateur
M. :	Olivier	Boëffard	Directeur
Mlle :	Nelly	Barbot	Co-directrice

La théorie, c'est quand on sait tout et que rien ne fonctionne. La pratique, c'est quand tout fonctionne et que personne ne sait pourquoi. Ici, nous avons réuni théorie et pratique : Rien ne fonctionne... et personne ne sait pourquoi !

Albert Einstein

Remerciements

Je tiens tout d'abord à remercier les membres du jury d'avoir bien voulu juger mes travaux. En particulier, je remercie chaleureusement Véronique Aubergé, Chargé de recherche CNRS et Thierry Dutoit, Professeur à la faculté polytechnique de Mons, d'avoir accepté la charge de rapporter mon travail de thèse. Je remercie également Christophe d'Alessandro, Directeur de recherche CNRS, de m'avoir fait l'honneur de présider mon jury de thèse. Merci également à Thierry Moudenc, responsable de l'équipe synthèse de parole à Orange Labs, d'avoir participé à mon jury.

Je tiens à remercier vivement Olivier Boëffard et Nelly Barbot qui ont encadré mes travaux de thèse. Cette aventure a débuté il y a maintenant quatre années par un stage de Master Recherche encadré par Olivier et Nelly. Vous avez su me montrer la richesse du monde de la recherche et je vous en remercie. Un grand merci à Olivier qui m'a incité à me montrer créatif et également à Nelly pour sa rigueur nécessaire.

Ce travail de recherche a été réalisé au sein de l'équipe Cordial dirigée par Laurent Miclet et je le remercie tout naturellement pour m'avoir accueilli dans son équipe mais également pour ses conseils avisés. Je souhaite également remercier Jean-Christophe Pettier et Daniel Rocacher pour m'avoir donné la chance d'enseigner et de m'avoir fait confiance. Beaucoup d'autres personnes font partie de ma liste de remerciements, je pense en particulier à Laure, Gaëlle, Joëlle, Nelly, Sabri, Ali, Vincent, Arnaud, Hélène et également à tous les autres membres de l'équipe. Merci à vous tous qui avez partagé avec moi tous ces moments de travail et également de détente ainsi que toutes ces discussions plus ou moins sérieuses.

Je souhaite également remercier ma compagne, Lydie, et mon fils, Elouan. Vous avez été à mes côtés dans les moments les plus difficiles et supporté tout au long de mon travail.

Table des matières

Remerciements **v**

Table des matières **vii**

Introduction **1**

I Etat de l'art **5**

Introduction à la première partie **7**

1 Parole et prosodie **9**

1.1 Description acoustique du signal de parole 9

1.2 Description physiologique de la parole 11

1.3 La prosodie 17

1.3.1 Définition de la prosodie 17

1.3.2 Intonation 19

1.3.3 Accent 19

1.3.4 Intensité et débit 21

1.4 Fonctions de la prosodie 22

1.4.1 Structuration de l'énoncé 22

1.4.2 Focalisation 22

1.4.3 Modalité 23

1.4.4 Fonctions non linguistiques 23

1.5 Paramètres de la prosodie 23

1.5.1 Fréquence fondamentale 23

1.5.2 Intensité 25

1.5.3 Durée 25

1.6 Caractérisation de la voix 25
1.6.1 Enjeux et applications 25
1.6.2 Études perceptives 26
1.6.3 Principaux paramètres acoustiques 27
1.7 Conclusion 28

2 La transformation de voix et ses applications 31
2.1 La transformation de voix 31
2.1.1 Principe 31
2.1.2 Apprentissage de la transformation 32
2.1.3 Transformation d'une phrase 34
2.1.4 Fonctions de transformation 34
2.2 Synthèse de la parole à partir du texte 36
2.2.1 Analyse morphosyntaxique 37
2.2.1.1 Pré-traitement 37
2.2.1.2 Analyse morphologique 38
2.2.1.3 Analyse syntaxique 39
2.2.2 Phonétisation 40
2.2.3 Génération de la prosodie 41
2.2.3.1 Prédiction de la durée 41
2.2.3.2 Prédiction de la fréquence fondamentale 42
2.2.4 Synthèse du signal de parole 42
2.2.4.1 Modélisation du signal de parole 42
2.2.4.2 Synthèse par règles - Synthèse par concaténation d'unités 44
2.2.5 Synthèse de la parole et transformation de la voix 44
2.3 Reconnaissance du locuteur 45
2.3.1 Définition 45
2.3.2 Sources d'erreurs de reconnaissance 46
2.3.3 Architecture 47
2.3.4 Choix des caractéristiques de la voix 48
2.3.5 Reconnaissance dépendante/indépendante du texte 49
2.3.6 Reconnaissance du locuteur et transformation de la voix 50
2.4 Conclusion 51

3 Modélisation de la prosodie : un état de l'art 53
3.1 Stylisation de la fréquence fondamentale 54
3.1.1 Tones and Break Indices 54

3.1.2 INTSINT (INternational Transcription System for INTonation) . 55
3.1.3 PaInTE (Parametric INTonation Event) 57
3.1.4 Modèle RFC : Rise/Fall/Connection 59
3.1.5 Tilt 61
3.1.6 Stylisation MoMel 62
3.1.7 Modèle de Fujisaki 63
3.1.7.1 Présentation 63
3.1.7.2 Interprétation physiologique 65
3.1.7.3 Estimation des paramètres 67
3.1.8 Systèmes dynamiques 68
3.2 Classification de contours de F_0 69
3.2.1 Normalisation et stylisation 70
3.2.2 Classification 70
3.3 Conclusion 72

4 Transformation de la prosodie : un état de l'art 73
4.1 Méthodes de transformation 74
4.1.1 Gaussian Normalization 74
4.1.2 Scatterplot ou Nth Order Conversion Function 74
4.1.3 Mean-variance (Ceyssens) 76
4.1.4 Méthode de Gillett et King 78
4.1.5 Table de correspondance 79
4.1.6 Fonction de transformation GMM 81
4.1.7 Transformation par arbre de régression et de classification 82
4.2 Évaluation de la transformation 85
4.2.1 Évaluation subjective 86
4.2.2 Appliquer la prosodie au signal de parole 87
4.2.3 Corpus parallèles ou non parallèles 87
4.3 Conclusion 88

Conclusion de la première partie 89

II Contributions 91

Introduction à la deuxième partie 93

5 Stylisation du F_0 par un modèle B-Spline **95**
5.1 Introduction . . . 95
5.2 Modélisation B-spline . . . 98
5.2.1 Description du modèle . . . 98
5.2.2 Discussion sur les splines . . . 100
5.2.3 Estimation des points de contrôle . . . 103
5.3 Estimation des nœuds . . . 104
5.3.1 Maximum de vraisemblance . . . 105
5.3.2 Positionnement libre des nœuds . . . 105
5.3.3 Optimisation par recuit-simulé . . . 106
5.4 B-splines et MDL . . . 108
5.4.1 Principe de solution . . . 108
5.4.2 Bornes théoriques sur les points de contrôle . . . 109
5.4.3 Influence de la précision des points de contrôle . . . 110
5.4.4 Critères MDL pour les B-splines . . . 112
5.5 Protocole expérimental . . . 112
5.6 Comparaison entre B-splines et splines : résultats et discussion . . . 113
5.7 Optimisation du nombre de nœuds : résultats et discussion . . . 117
5.7.1 Présentation d'un exemple . . . 117
5.7.2 Relation entre la RMS et les degrés de liberté du modèle . . . 118
5.7.3 Sensibilité des critères MDL par rapport à ε . . . 119
5.7.3.1 Critère (a) . . . 120
5.7.3.2 Critère (b) . . . 120
5.7.4 Analyse des critères MDL proposés . . . 121
5.8 Conclusion . . . 123

6 Apprentissage non supervisé de classes de contours mélodiques **125**
6.1 Introduction . . . 125
6.2 Modélisation des classes par des HMM . . . 126
6.2.1 Le modèle . . . 126
6.2.2 Gaussian Splitting . . . 127
6.3 Apprentissage non supervisé des classes . . . 128
6.4 Méthodologie . . . 130
6.4.1 Corpus de F_0 . . . 130
6.4.2 Préparation des données . . . 131
6.4.3 Évaluation . . . 131

6.5 Résultats et discussion 132
6.5.1 Exemple de contour mélodique 132
6.5.2 Résultats pour le critère CMSE 133
6.5.3 Exemple de partitionnement avec 16 classes 135
6.5.4 Comportement du critère de sélection de classe 138
6.6 Conclusion 141

7 Transformation de la prosodie par adaptation de GMM 143
7.1 Introduction 143
7.2 Pré-traitement des données 145
7.2.1 Interpolation et lissage 145
7.2.2 Représentation de la durée 145
7.2.3 Stylisation du F_0 145
7.2.4 Prosodie d'une syllabe 146
7.3 Transformation de la prosodie 146
7.3.1 Modélisation par mélange de lois gaussiennes, GMM 147
7.3.2 Adaptation du GMM source 148
7.3.3 Transformation de la durée et du pitch 149
7.4 Protocole expérimental 151
7.4.1 Données 151
7.4.2 Expériences 152
7.5 Résultats et discussion 152
7.5.1 Séries d'expériences 1 : adaptation de la moyenne 152
7.5.2 Séries d'expériences 2 : adaptation moyenne/variance 156
7.6 Conclusion 161

Conclusion de la deuxième partie 163

Conclusion 165

Bibliographie 182

Table des figures 183

Introduction

Depuis de nombreuses années, la production artificielle de la parole a été une préoccupation constante de la communauté scientifique. Les premiers efforts pour créer des machines parlantes sont apparus dans la deuxième moitié du 18^{e} siècle. Ainsi, en 1791, le baron von Kempelen a présenté une « machine parlante » construite à partir des connaissances de l'époque sur les mécanismes physiologiques de la parole. Cette machine pouvait émettre une vingtaine de sons différents.

Avec l'apparition de l'électricité et de l'électronique au début du 20^{e} siècle, des machines plus perfectionnées ont vu le jour. Notamment, en 1939, Homer Dudley présentait le VODER, *Voice Operation Demonstrator*, à l'exposition universelle de New York. Le fonctionnement du VODER repose sur l'excitation d'un ensemble fixe de filtres qui jouent le rôle de résonateurs. L'utilisation de cette machine nécessitait un entraînement assez long.

Depuis les années 50, époque à laquelle sont apparus les premiers synthétiseurs à formants, le domaine de la synthèse de la parole a grandement évolué. Notamment, l'évolution considérable des connaissances en traitement du signal et de la puissance de calcul des ordinateurs dans les années 70 a permis de révolutionner le monde de la synthèse vocale. De nos jours, les connaissances et les moyens techniques mis en œuvre permettent de créer une parole synthétique de bonne qualité.

La synthèse de la parole s'inscrit dans le domaine de la communication personne-machine. Elle apporte un certain confort pour l'utilisateur puisque l'oral est une interface naturelle, quasi-universelle. Dans certains cas, il s'agit du seul moyen de communication possible, par exemple pour des mal-voyants. Dans d'autres cas, un message oral peut être plus efficace qu'un message écrit. Par exemple, les limites spatiales du champ de vision ne s'appliquent pas à l'ouïe, ce qui permet à un message oral d'attirer plus facilement l'attention de l'usager.

Pour obtenir une communication efficace, la parole synthétique doit être de bonne qualité, elle doit être intelligible, la plus naturelle possible et également cohérente avec

la tâche à réaliser. Des progrès doivent encore être fait principalement au niveau de la qualité de la voix et notamment celui de la prosodie.

Un champ d'usage des méthodologies de traitement de la parole, complémentaire à celui de la synthèse, concerne la transformation de la voix. On peut définir la transformation de voix comme étant une modification des caractéristiques acoustiques d'une phrase issue d'un locuteur *source* de telle sorte qu'elle soit perçue comme si elle était prononcée par un locuteur *cible*. L'objectif de cette thèse peut être vu comme un sous-ensemble de la transformation de voix puisqu'il s'agit de modifier les caractéristiques prosodiques d'un locuteur source vers celles d'un locuteur cible.

Pour y parvenir, il est nécessaire de s'intéresser tout d'abord à la modélisation de la prosodie, et dans notre cas, nous nous focaliserons sur l'intonation qui est principalement portée par l'évolution des contours mélodiques. Ensuite, lorsque l'on dispose d'un modèle de prosodie pour chaque locuteur, on peut se poser la question de la mise en correspondance de ces deux modèles afin de construire une fonction de transformation. Cette fonction permettrait de passer d'un modèle à l'autre et donc de la prosodie d'un locuteur à l'autre. Ce modèle doit donc être en mesure de capter les spécificités prosodiques liées au locuteur.

Une application possible de cette méthodologie est la synthèse de la parole afin de réduire le temps et le coût de développement d'une voix de synthèse. Il serait en effet intéressant de disposer d'une voix de référence que l'on puisse transformer pour atteindre une cible particulière désirée ou bien pour créer de nouvelles voix de synthèse. Un autre champ d'application est la biométrie : connaître la fonction de passage d'une voix à une autre apporte des connaissances sur ce qui fait l'identité de la voix. Une telle méthodologie serait donc applicable dans des systèmes d'authentification ou de reconnaissance du locuteur.

Ce document est structuré en deux parties. L'une, Partie I, *État de l'art*, présente des notions élémentaires sur la parole et sa synthèse ainsi qu'une description bibliographique des méthodes de stylisation et de transformation de la prosodie. L'autre, Partie II, *Contributions*, décrit les travaux de recherche menés sur la stylisation et la transformation de la prosodie ainsi que leurs résultats.

La première partie, Partie I, *État de l'art*, présente le contexte scientifique général de cette thèse ainsi que les travaux existants sur la transformation de la prosodie. Ainsi, le chapitre 1 page 9 décrit tout d'abord la parole sur les plans acoustique et physiologique, puis définit de manière plus précise la prosodie. Enfin, ce chapitre introduit le problème de la caractérisation de la voix. Le chapitre suivant, chapitre 2 page 31, aborde le domaine de la transformation de voix et décrit deux de ses applications possibles : la

synthèse de la parole et la reconnaissance du locuteur. Le chapitre 3 page 53 aborde le problème de la modélisation de la prosodie, et en particulier de l'intonation, en donnant un aperçu des principaux modèles proposés par la communauté scientifique. Le dernier chapitre de cette partie, chapitre 4 page 73 propose un tour d'horizon des principales approches de transformation de la prosodie. L'évaluation des systèmes de transformation de la prosodie y est également discutée.

Les chapitres de la première partie apportent les informations nécessaires pour disposer d'une vision générale du domaine de traitement de la parole et notamment de la transformation de la prosodie. La deuxième partie, Partie II, *Contributions*, introduit les trois axes de recherche chronologiques qui ont animé ce travail de thèse. Le premier, décrit dans le chapitre 5 page 95, porte sur la stylisation de contours mélodiques par un modèle B-spline pour lequel nous proposons un critère permettant de choisir de manière automatique le nombre de paramètres du modèle. Ce modèle permet ici d'obtenir une précision de description des contours mélodiques suffisante avec un nombre de paramètres réduit. Cependant, ce modèle ne permet pas de représenter directement l'espace des contours mélodiques d'un locuteur par un ensemble restreint de modèles. Nous proposons donc, dans le chapitre 6 page 125, un modèle HMM qui établit un partitionnement de cet espace de manière non supervisée. Il permet en outre d'intégrer les variations de longueur des contours mélodiques afin de rendre comparables deux contours de longueurs différentes. Dans un contexte de transformation de la prosodie, il est également nécessaire de pouvoir mettre en correspondance les contours mélodiques réalisés par le locuteur source et ceux réalisés par le locuteur cible. Cette mise en correspondance peut être effectuée de manière individuelle (un contour source est associé à un contour cible), ou bien globale. Dans le chapitre 7 page 143, une méthodologie permettant d'adapter un modèle des contours mélodiques du locuteur source aux contours mélodiques du locuteur cible est proposée.

Première partie

Etat de l'art

Introduction à la première partie

Cette première partie permet de présenter les notions fondamentales concernant notre objet d'étude qu'est la parole. Le chapitre 1 page 9 est consacré à la description de la parole sur les plans acoustique et physiologique ainsi qu'à la définition de la prosodie qui représente le point central de cette étude. Modéliser la prosodie est, en effet, une étape indispensable pour tout système de synthèse de la parole à partir du texte afin d'obtenir une parole proche du naturel.

Le chapitre 2 page 31 aborde le domaine de la transformation de voix et ses applications. Tout d'abord, le problème de transformation de voix est défini et l'architecture d'un tel système est détaillée. La suite de ce chapitre est consacrée à deux applications possibles qui sont la synthèse de la parole et la reconnaissance du locuteur.

Dans le chapitre 3 page 53, les principaux modèles de prosodie sont décrits. Un panel assez large de modèles est présenté en partant des représentations symboliques de la prosodie, dont le système d'annotation ToBI est un représentant incontournable, du moins pour l'anglais, jusqu'à des modèles mathématiques issus du traitement du signal ou encore de l'apprentissage artificiel.

Le dernier chapitre de cette partie, chapitre 4 page 73, traite de la transformation de la prosodie d'un locuteur à un autre. L'objectif de ce chapitre est de présenter les différentes méthodes existantes pour modifier la prosodie d'un locuteur source de sorte qu'elle soit perçue comme celle d'un locuteur cible. Nous verrons que pour atteindre cet objectif de transformation, il est tout d'abord nécessaire de modéliser correctement la prosodie des locuteurs source et cible.

Chapitre 1

Parole et prosodie

La parole résulte de phénomènes physiques, physiologiques et cognitifs propres à l'homme. Sur le plan acoustique, la parole se traduit par une mise en vibration de l'air. La mesure de cette vibration par l'intermédiaire d'un microphone permet de représenter la parole sous la forme d'un signal qui possède une forme caractéristique que nous allons étudier dans une première partie. Dans une deuxième partie, nous allons présenter les mécanismes complexes qui mettent en jeu un ensemble d'organes de phonation pour produire la parole. La troisième partie est, quant à elle, consacrée à une description succincte de la prosodie, qui comme nous le verrons plus loin, peut être définie comme un vecteur d'information accompagnant le message. Enfin, la dernière partie de ce chapitre traite de la caractérisation de l'identité de la voix.

1.1 Description acoustique du signal de parole

Le signal de parole possède une structure complexe et peut être analysé en termes d'événements de nature temporelle et fréquentielle. Cette structure temps-fréquence découle des caractéristiques statiques et dynamiques de l'appareil de phonation. Un exemple de signal est donné sur la figure 1.1 page suivante, il correspond à la phrase « un beau tir groupé ». On constate que le signal représenté est constitué d'une alternance de zones « périodiques » et de zones bruitées, appelées *zones voisées* et *non voisées*. En bas de cette figure, le spectrogramme du signal est représenté. Il s'agit d'une représentation à trois dimensions sur laquelle on trouve le temps en abscisses et les fréquences en ordonnées. Sur ce spectrogramme, une troisième dimension, représentée par le degré de noirceur de la figure, est l'énergie du signal. Sur cette figure, la zone allant de 0.87s à environ 1.15s est une zone voisée et correspond à une réalisation du son [i]. La forme

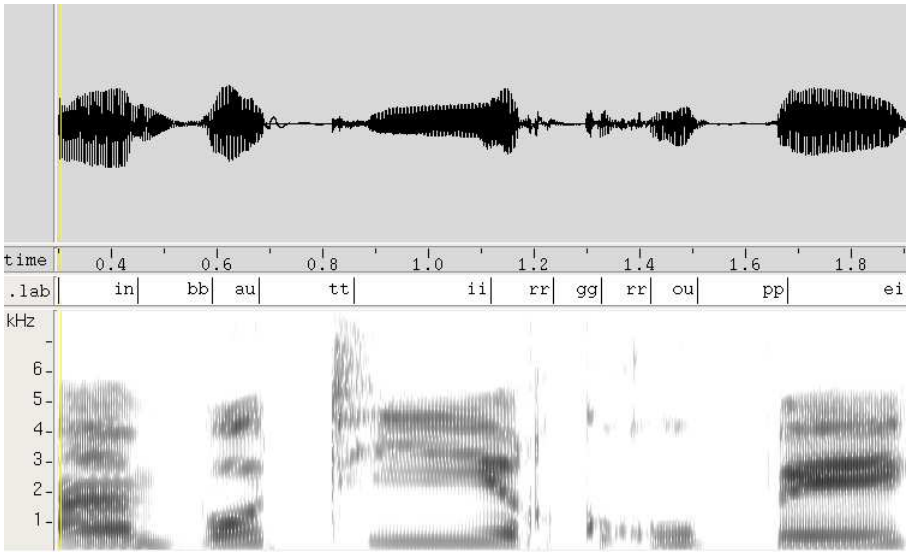


FIG. 1.1 – Signal et spectrogramme de « un beau tir groupé ». De haut en bas, on trouve (a) le signal de parole correspondant à « un beau tir groupé », (b) la transcription en phones et (c) le spectrogramme du signal (analyse en bande étroite).

d'onde du signal de parole pendant ce son peut être considérée comme périodique. Au contraire, le son [p] est non voisé et la forme d'onde du signal n'est pas périodique.

D'après la théorie de Fourier :

« Tout signal périodique, voire apériodique, d'énergie finie peut se décomposer sur la base d'une série de composantes sinusoïdales. »

Cette théorie met en avant le fait qu'un signal complexe peut s'écrire sous la forme d'une somme de signaux élémentaires. Un signal élémentaire, une sinusoïde, est décrit par trois paramètres :

- *l'amplitude* correspond, pour un signal électrique, à la valeur maximale du signal et est directement liée à la quantité d'énergie portée par un son. L'énergie d'un signal sonore s'exprime le plus souvent sur une échelle logarithmique en *décibels* (dB).
- *la fréquence* est le nombre d'oscillations de la valeur du signal autour de ces valeurs extrêmes par unité de temps. Elle s'exprime en cycle par seconde ou Hertz (Hz).
- *la phase* mesure la position de la sinusoïde à l'instant initial de l'onde. Cette quantité représente un décalage angulaire et s'exprime en radian modulo 2π.

La théorie de Fourier permet de passer d'une représentation temporelle du signal de parole à une représentation spectrale et réciproquement. Cette dernière peut être la

représentation des amplitudes ou des phases en fonction des fréquences.

La figure 1.2 page suivante présente un son pur (a) et un son complexe issu de la parole (c), ainsi que leurs spectres d'amplitude respectifs en (b) et (d). On peut observer, en (c), le spectre d'amplitude d'un son pur qui est une sinusoïde de fréquence 1000 Hz. Ce spectre présente un pic à la fréquence 1000 Hz qui est entouré de bruit spectral dû aux erreurs de calcul de la transformée de Fourier discrète. En comparaison, un signal de parole et son spectre (resp. en (c) et (d)) sont plus riches que dans le cas d'un son pur.

Sur la partie (c) de la figure 1.2 page suivante, on peut également observer qu'une forme quasi-périodique est présente. Le temps de cycle de cette structure définit la période fondamentale dont l'inverse, le fondamental, est noté F_0. La fréquence fondamentale varie approximativement de 70 à 250 Hz pour les hommes, de 150 à 400 Hz pour les femmes et de 200 à 600 Hz chez les enfants.

On retrouve sur la partie (d), l'enveloppe du spectre d'amplitude calculée à l'aide d'un filtre auto-régressif d'ordre 12 estimé selon la méthode de la covariance modifiée. Le premier pic spectral situé au voisinage de 100 Hz correspond à la fréquence fondamentale. On retrouve cette valeur en calculant l'inverse de la période fondamentale sur le signal (c), voisine de 10 ms. Les pics observables sur le spectre, ou harmoniques, possèdent une fréquence centrale qui s'exprime en multiples de la fréquence fondamentale. La forme générale d'une enveloppe spectrale comporte des pics et des vallées qui correspondent aux résonances et anti-résonances du conduit vocal, lesquels sont appelés formants et anti-formants (Boite *et al.*, 2000). L'évolution temporelle de leur fréquence et leur largeur de bande détermine le timbre du son.

L'enveloppe spectrale d'un son voisé présente une forme caractéristique avec des pics spectraux marqués comme dans l'exemple présenté ici.

1.2 Description physiologique de la parole

La parole, très souvent considérée comme une activité propre à l'homme, met en jeu la coordination de nombreux muscles et organes de phonation. Elle nécessite donc une coordination musculaire qui passe par un apprentissage. La production de parole correspond à un ensemble de commandes du système nerveux central jusqu'aux muscles. Celles-ci permettent de piloter les évolutions physiologiques du conduit vocal au cours du temps.

Tout d'abord, après une inspiration, le diaphragme expulse l'air des poumons et agit comme une force mécanique. Le flux d'air passe alors par la trachée, puis le larynx et

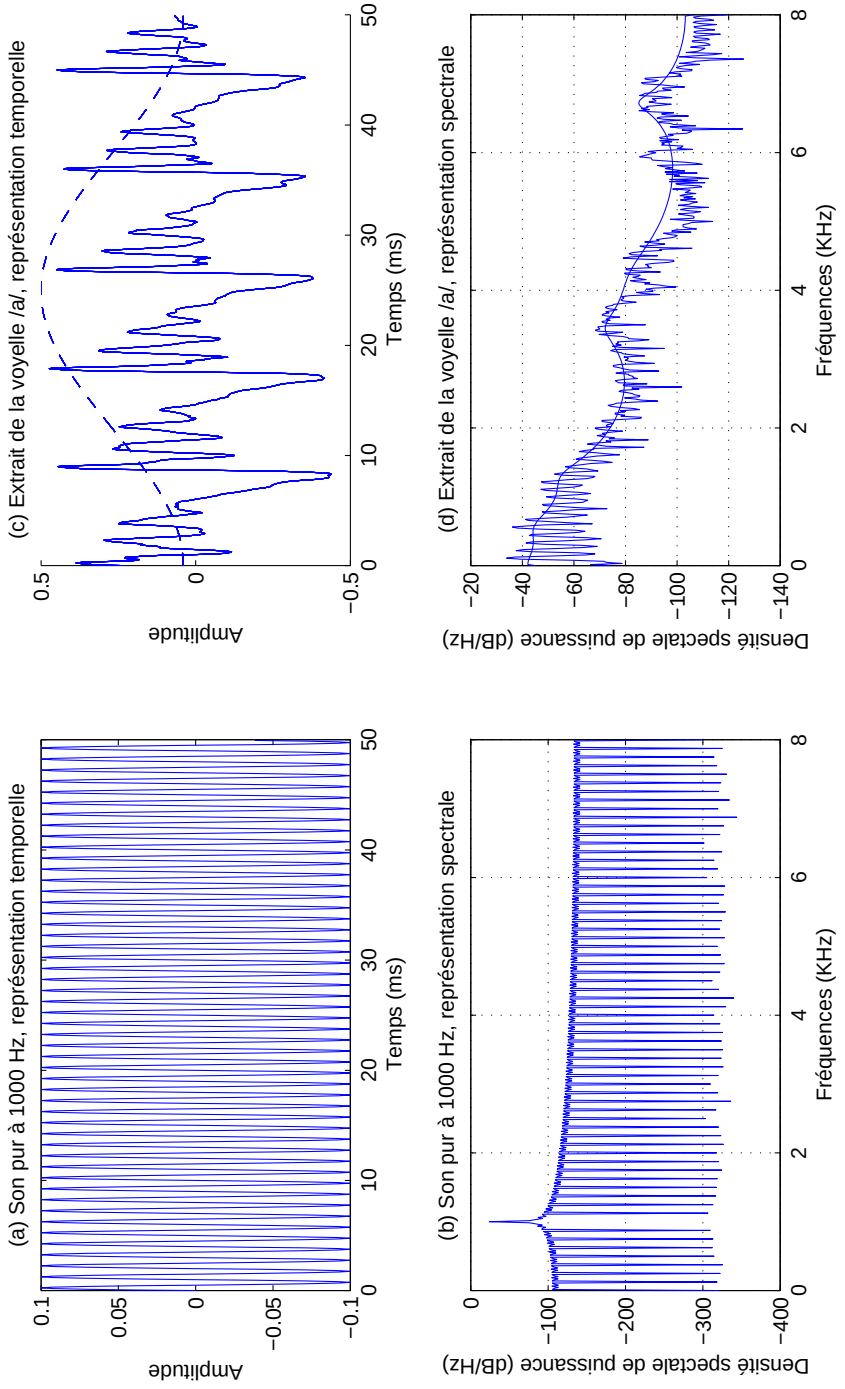


FIG. 1.2 – Représentation spectrale d'un son pur et d'un son complexe. En (a) et (b), on peut observer respectivement les représentations temporelle et spectrale d'un son pur. Le son pur est une sinusoïde de fréquence 1000Hz. En (c) et (d), on peut observer les représentations temporelle et spectrale d'un son complexe, ici un extrait de la voyelle /a/. La courbe discontinue (hors échelle), en (c), représente une fenêtre de Hamming de 50 ms utilisée pour l'estimation du spectre représenté en (d).

enfin par le double conduit vocal (oral et nasal). Au niveau de celui-ci, l'air se divise en deux flux et passe pour partie par le conduit oral puis est ensuite évacué au niveau des lèvres, l'autre partie passant par le conduit nasal pour sortir par les narines. La figure 1.3 page suivante présente les principaux acteurs de la phonation.

L'ensemble poumons-trachée, appelé soufflerie, se comporte comme un générateur d'air permettant « d'exciter » la source vocale. Cette source acoustique peut être de nature périodique et correspond à la vibration des cordes vocales au niveau du larynx, ou de nature apériodique, produisant des bruits d'explosion ou de friction qui peuvent naître à l'intérieur du conduit vocal (de la glotte aux lèvres). Sur un plan physique, la parole est donc le résultat de l'excitation des cavités supra-glottiques (nasale et/ou orale) par une source acoustique. La nature apériodique de la source vocale peut s'ajouter ou se substituer à sa nature périodique.

Le larynx Le larynx est le lieu de naissance de la composante quasi-périodique du signal de parole. Cet organe essentiel de la parole est une source vocale devant être alimentée en air pour émettre des sons. Le larynx est situé dans la région moyenne du cou (voir figure 1.3 page suivante). Sa position dépend de l'âge, du sexe et de l'individu. Structure déformable, il est constitué de nombreux muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée. Sa cavité interne peut se décomposer en trois parties : le vestibule, l'étage moyen et l'étage inférieur.

Les fausses cordes vocales sont situées de part et d'autre du larynx, on les appelle bandes vestibulaires. Les *cordes vocales* (vocal folds), dont deux photographies sont proposées sur la figure 1.4 page 15, sont situées juste en dessous des bandes vestibulaires. Les *cordes vocales* sont deux « lèvres » situées en travers du larynx. L'ouverture qu'elles forment en s'écartant ou en se rapprochant est appelée *glotte*. Pendant la respiration, la voix chuchotée ainsi que lors de la phonation de sons *non voisés*, les cordes vocales laissent passer librement l'air à travers le larynx. Au contraire, les sons *voisés* résultent d'une vibration périodique des cordes vocales. Le principe de vibration des cordes vocales, illustré sur la figure 1.5 page 16, est le suivant :

1. Les cordes vocales sont d'abord complètement fermées ce qui fait accroître la pression de l'air en amont de celles-ci (figure 1.5(a) page 16).
2. Lorsque la pression de l'air est suffisante, elle force les cordes vocales à s'ouvrir (figures 1.5(b) à 1.5(d) page 16).
3. La vitesse de l'air augmente, entraînant une diminution de pression (figures 1.5(e) à 1.5(g) page 16).

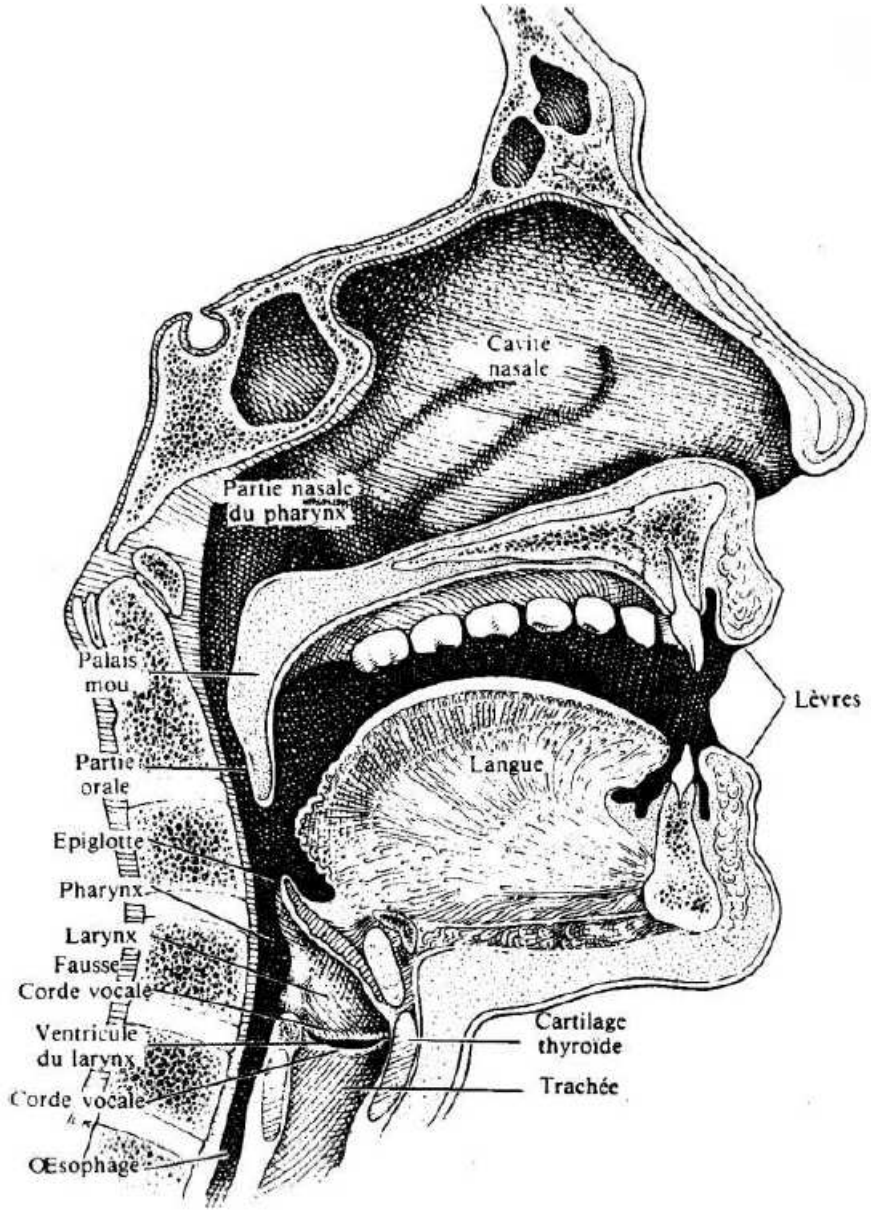
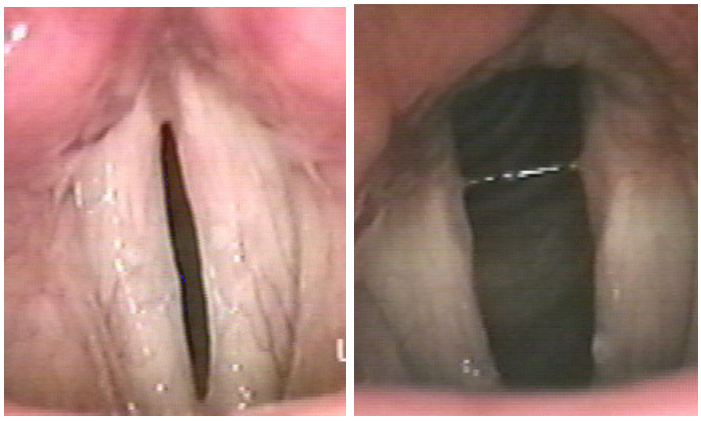


FIG. 1.3 – Coupe sagittale présentant les principaux acteurs de la phonation (rapporté par (Miller, 1986)).



(a) (b)

FIG. 1.4 – Vue des cordes vocales (extrait du site web du centre pour la parole de l'université de Pittsburgh, `http://www.pitt.edu/~crosen/voice/normcords.html`). L'ouverture formée par les cordes vocales est appelée *glotte*. Ces deux lèvres situées en travers du larynx, en vibrant de manière périodique, sont à l'origine de la production des sons voisés. Lors de la production de parole (a), les cordes vocales sont rapprochées. Au contraire, en (b), lors de la respiration, elles laissent passer l'air librement.

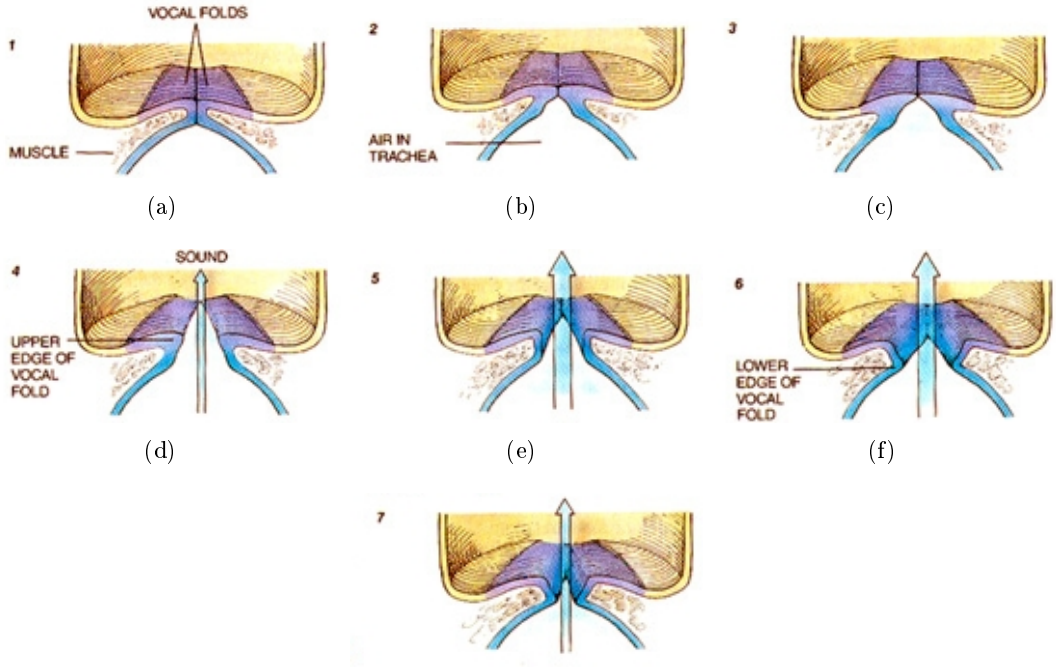
4. La diminution de pression engendrée par cette brève ouverture permet aux cordes vocales de s'accoler et d'obstruer à nouveau le passage de l'air.

Un signal périodique est ainsi produit au niveau du larynx. La fréquence de vibration et la hauteur des sons produits dépendent de la tension des cordes vocales et de leur masse.

Articulation Le conduit vocal, constitué du conduit oral et du conduit nasal, est un lieu important dans le processus de fabrication de la parole. En particulier, c'est dans le conduit oral que se forme la majorité des mouvements articulatoires. Les principaux articulateurs sont par ordre d'importance la langue, les lèvres, le voile du palais et les mandibules. La langue, composée d'un nombre important de muscles, est l'articulateur le plus mobile.

Les sons de la parole peuvent être regroupés en classes, notamment en fonction de leur mode articulatoire. On distingue généralement trois classes principales : les voyelles, les semi-voyelles et les liquides, ainsi que les consonnes. La discipline qui étudie la classification des sons en fonction d'indices articulatoires et acoustiques constitue la phonétique (Calliope, 1989; Boite *et al.*, 2000).

La phonologie s'intéresse à l'organisation des sons d'une langue afin de former un



(a) (b) (c) (d) (e) (f) (g)

FIG. 1.5 – La vibration des cordes vocales. En (a), les cordes vocales sont accolées et obstruent le conduit vocal. En (b) et (c), la pression de l'air augmente. En (d), (e) et (f), lorsque la pression exercée sur les cordes vocales est suffisante, ces dernières s'écartent et laisse passer l'air. Cette ouverture provoque une diminution de la pression de l'air et les cordes vocales vont pouvoir à nouveau s'accoler en (e). (extrait du site web du centre pour la parole de l'université de Pittsburgh, `http://www.pitt.edu/~crosen/voice/anatomy2.html`).

énoncé. Elle introduit le phonème comme la plus petite unité distinctive, c'est-à-dire qu'elle permet de distinguer le sens des mots. Par exemple, les mots « cote » (/kɔt/) et « côte » (/kot/) sont différenciés par l'utilisation respective des phonèmes /ɔ/ et /o/. Le phonème est une unité abstraite qui n'est pas définie sur une base acoustique ou articulatoire. On distingue en général 37 phonèmes en français. Un *phone* est un son qui correspond à la réalisation acoustique d'un phonème. Les variantes articulatoires d'un phonème constituent ses *allophones*. Le mot « rat » peut, par exemple, être prononcé avec un /r/ roulé, grasseyé ou normal qui se notent respectivement [r], [ʀ] et [ʁ] en utilisant la notation phonétique.

Pour résumer, le processus de phonation comporte trois étapes essentielles :

1. la génération d'une énergie ventilatoire dont le but est de mettre en mouvement les cordes vocales et/ou de générer des bruits ;
2. la vibration des cordes vocales qui donne naissance aux sons voisés (plus de 80% du temps de phonation) et/ou apparition de bruits d'explosion ;
3. la réalisation d'une gestuelle articulatoire au niveau des cavités supra-glottiques (conduit vocal et fosses nasales).

Pour le lecteur intéressé, de plus amples explications sont consultables dans le livre de *Calliope* concernant la parole et son traitement automatique (Calliope, 1989).

1.3 La prosodie

Dans ce paragraphe, nous décrivons de manière succincte la prosodie du français.

1.3.1 Définition de la prosodie

Di Cristo (2000) propose une définition assez complète de la prosodie : *« **La prosodie** (ou la prosodologie) est une branche de la linguistique consacrée à la description (aspect phonétique) et à la représentation formelle (aspect phonologique) des éléments de l'expression orale tels que les accents, les tons, l'intonation et la quantité, dont la manifestation concrète, dans la production de la parole, est associée aux variations de la fréquence fondamentale (F_0), de la durée et de l'intensité (paramètres prosodiques physiques), ces variations étant perçues par l'auditeur comme des changements de hauteur (ou de mélodie), de longueur et de sonie (paramètres prosodiques subjectifs). Les signaux prosodiques véhiculés par ces paramètres sont polysémiques et transmettent à la fois des informations para-linguistiques et des informations linguistiques déterminantes*

pour la compréhension des énoncés et leur interprétation pragmatique dans le flux du discours. »

D'après cette définition, la prosodie traite des éléments de l'expression orale qui se manifestent physiquement par des variations de F_0, de durée et d'intensité. Ces éléments de l'expression orale transmettent notamment des informations sur le sens d'un énoncé. Sur l'exemple suivant, à l'écrit, une marque syntaxique (« ? ») permet la distinction entre les deux phrases et il n'y a pas d'ambiguïté sur leur sens :

- Le train arrive à midi.
- Le train arrive à midi ?

À l'oral, la situation est différente et il est nécessaire pour se faire comprendre de transposer la marque interrogative de l'écrit dans le message oral. On peut noter que la nature des sons, les phonèmes ne changent pas dans les deux exemples. C'est un autre procédé qui va permettre de modifier le sens de la phrase et c'est l'intonation qui est utilisée. Le sens de la phrase dépend donc de l'intonation et plus généralement de la prosodie. On voit sur cet exemple que la prosodie est un procédé non univoque. Autrement dit, un même énoncé peut être prononcé avec des prosodies différentes. Ces différences de prosodie influent sur le sens de l'énoncé, c'est pourquoi la prosodie paraît essentielle à la compréhension et au naturel de la parole.

Elle dépend non seulement du niveau de la syntaxe mais aussi de la sémantique. De plus, selon (Vannier, 1999), la prosodie a la particularité d'être à la fois universelle et spécifique à une langue. Autrement dit, chaque langue possède sa propre prosodie, même si elle partage certaines propriétés avec d'autres langues. Pour une même langue, on note également l'existence d'une diversité intra-locuteur et interlocuteur qui peut par exemple être liée à l'état d'esprit du locuteur ou encore à son origine socioculturelle.

Dans le domaine du traitement de la parole, les paramètres prosodiques prennent une importance particulière (Calliope, 1989). Pour la synthèse de la parole, ils contribuent à une meilleure intelligibilité du signal synthétique. Pour ce qui est de la reconnaissance, ils peuvent être utiles pour lever des ambiguïtés possibles et déterminer ainsi la structure d'une phrase énoncée.

Nous étudierons plus spécifiquement dans le paragraphe 1.5 page 23 les principaux paramètres de la prosodie que sont la durée, l'intensité et la fréquence fondamentale. Comme l'indique (Calliope, 1989), de manière simplifiée, on peut considérer que l'information prosodique se résume essentiellement à l'évolution de la fréquence fondamentale qui, de ce fait, apparaît comme un élément prépondérant de la prosodie.

1.3.2 Intonation

L'article de Delattre (Delattre, 1966) établit une classification des différentes intonations possibles dans un énoncé. Pour établir les principaux types d'intonation du français, Delattre a utilisé des extraits de conversations, de pièces de théâtre et de conférences. Le résultat de cette analyse met en évidence l'existence de dix types d'intonation de base (figure 1.6 page suivante).

Avec ce modèle d'intonation, il existe quatre niveaux d'intonation : basse, moyenne, haute et aiguë. Cette modélisation met en jeu les trois modalités suivantes : interrogation, exclamation, affirmation. En particulier, Delattre montre qu'en faisant des substitutions entre les intonations de base dans une phrase de même contenu, on obtient des changements de sens. Cela montre notamment le rôle important de l'intonation pour la compréhension du message oral.

D'après Mertens (1993), *« Le mot* ton *désigne le ou les niveaux de hauteurs observés dans une syllabe donnée. Le ton coïncide donc avec la partie de la courbe mélodique qui se rattache à une seule syllabe ».* L'intonation d'un énoncé se présente comme une succession de tons. Dans ses travaux, Mertens, distingue les quatre niveaux de hauteur haut, bas, infra-bas et suraigu.

1.3.3 Accent

En français, l'unité porteuse de l'accent est la syllabe et l'accentuation, comme l'écrit Mertens (1992), peut être définie de la manière suivante : *« Une syllabe est dite accentuée quand elle ressort sur son entourage par sa force particulière, par un contraste d'intensité subjective. »*

On distingue généralement trois classes de langues suivant leur comportement par rapport à l'accentuation :

- les langues à accent libre (cas de l'anglais) : on ne peut pas déterminer sa place à l'avance et il y a autant de possibilités de placement de l'accent que de syllabes (dans le cas où l'unité accentuable est la syllabe) ;
- les langues à accent déterminé, ou « fixe » (cas du français) : l'application des lois de placement de l'accent suppose le décompte préalable des syllabes. En règle générale, on détermine la place de l'accent en partant de la fin d'un mot ;
- les langues à tons : les variations tonales y sont utilisées à des fins sémantiques pour distinguer plusieurs significations linguistiques.

Deux grandes catégories d'accents sont distinguables en français : l'accent final (primaire) et l'accent non-final (secondaire) associé aux fonctions linguistiques (focalisation)

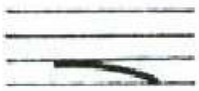
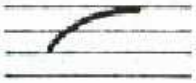
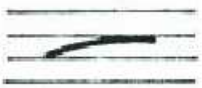
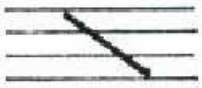
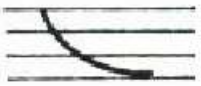
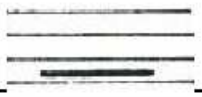
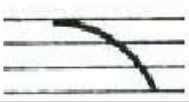
		représentations graphiques
déclaratives	finalité	2-1 
	continuation majeure	2-4 
	continuation mineure	2-3 
	implication	2-4_ 
	commandement	4-1 
interrogatives	question	2-4+ 
	interrogation	4-1 
parenthétiques	parenthèse	4-1 
	écho	4-4 
	exclamation	4-1 

FIG. 1.6 – Les dix intonations de base de Delattre. Il s'agit d'un inventaire de formes de base proposées par Delattre sur une échelle qui possède quatre niveaux. Celles-ci sont identifiées selon des critères fonctionnels et perceptifs. (extrait de (Delattre, 1966)).

ou para-linguistiques (expressivité). Pour Di Cristo (1998), sur le plan lexical, l'accent français n'est pas distinctif ni pour les mots ni pour les morphèmes.

Plusieurs dénominations existent pour parler de l'accent final du français : accent logique, objectif, tonique, normal ou interne. La plupart des études s'accordent à dire que l'accent primaire est assigné à la dernière syllabe pleine (qui ne contient pas de schwa) du dernier item lexical d'un groupe accentuel. Il se caractérise principalement par un allongement important de la durée.

L'accent secondaire est optionnel. De manière générale, sur le plan acoustique, l'accent secondaire se traduit par une augmentation de la fréquence fondamentale et de l'intensité. Il obéit à des contraintes de nature rythmique, pragmatique et expressive. Trois types d'accents secondaires peuvent ainsi être distingués :

- *L'accent rythmique* est lié à la réalisation d'un décompte syllabique et permet la mise en relief de la mélodie. Il peut être associé à l'augmentation de l'intensité.
- *L'accent pragmatique,* également qualifié d'accent énonciatif ou d'accent de focalisation, permet de mettre en relief une partie de l'énoncé.
- *L'accent expressif ou emphatique,* exprime l'attitude du locuteur à l'égard de ce qu'il dit.

1.3.4 Intensité et débit

Le rôle de l'intensité est strictement communicatif. L'enjeu pour un locuteur est de montrer s'il désire conserver la parole ou bien la céder à son interlocuteur, de montrer une certaine insistance sur une partie de son discours ou encore de montrer son adhésion plus ou moins forte à l'énoncé.

Comme le note Zellner (1998), même si le débit est généralement exprimé en unités de parole par unité de temps, par exemple en nombre de syllabes par seconde pour le cas du français, augmenter ou diminuer le débit d'une phrase de manière globale ne produit pas un effet naturel. En effet, le débit est influencé par plusieurs facteurs qui sont notamment les pauses, l'allongement ou le raccourcissement des segments, l'ajout de sons. Bien que le dernier point puisse sembler inattendu, il correspond notamment à un phénomène que l'on pourrait qualifier « d'hyper-articulation » et qui provoque un ralentissement du débit de la parole. Les variations de durée peuvent par exemple traduire l'hésitation du locuteur, une certaine incertitude ou encore une émotion (Beller *et al.*, 2006). Concernant le débit du français, il se situe entre 4 et 7 syllabes par seconde.

1.4 Fonctions de la prosodie

En 2003, Fónagy a écrit une synthèse des différentes fonctions de l'intonation (Fónagy, 2003). Dans ce paragraphe, nous allons simplement lister les fonctions principales de la prosodie.

1.4.1 Structuration de l'énoncé

Dans les langues à accent libre, la prosodie permet la distinction entre homonymes. En anglais par exemple, elle permet de distinguer un nom d'un verbe (en gras, position de l'accent lexical) :

- **seg**ment (un segment)
- seg**ment** (segmenter)

Dans les langues à accent fixe, la position de l'accent lexical est la même pour tous les mots. En français, l'accent lexical intervient sur la dernière syllabe. Il permet de distinguer les frontières des mots (fonction démarcative).

De manière générale, la prosodie permet de déterminer la structure d'un énoncé. Par exemple, certaines ambiguïtés syntaxiques peuvent apparaître :

« La petite brise la glace »

Cette phrase peut prendre deux sens différents selon que le verbe est « briser » ou « glacer ». Ici, la prosodie nous aide à déterminer quel est le sens adapté à la situation. Par exemple, une pause du locuteur après « La petite » permet de focaliser l'attention sur le sujet et de le délimiter. Ainsi, cela permet de lever l'ambiguïté syntaxique de la phrase. Dans ce cas, on peut aisément identifier que le verbe est « briser » et non « glacer ». L'expression « une tasse de thé russe » est un autre exemple d'ambiguïté pouvant être levée grâce à l'intonation. Est-ce la tasse qui est russe ou bien le thé ? On voit alors clairement apparaître la fonction de structuration que possède la prosodie.

1.4.2 Focalisation

La focalisation est un moyen d'insister sur certains mots. Dans l'exemple suivant, l'accentuation apporte un sens particulier à la phrase :

- **Je** vais terminer
- Je **vais** terminer
- Je vais **terminer**

Ici, le premier cas montre une insistance sur le sujet pour montrer qu'il est important que ce soit *moi* qui termine. La suivante traduit le fait que cela sera terminé mais qu'il faut encore du temps. Enfin, la dernière assure que la tâche sera bel et bien terminée, c'est une certitude. Ces différentes accentuations apportent donc des nuances de sens qui traduisent une volonté du locuteur par rapport au message qu'il souhaite transmettre.

1.4.3 Modalité

La prosodie, et plus particulièrement la mélodie, est liée au mode de la phrase. On peut distinguer quatre modes : affirmatif, interrogatif, impératif et exclamatif. Elle permet, par exemple, d'identifier une question sans qu'il y ait besoin d'indices syntaxiques tels qu'une inversion sujet/verbe. Ainsi, la phrase « tu vas bien » peut être produite selon les modes interrogatif, affirmatif ou encore exclamatif.

1.4.4 Fonctions non linguistiques

La prosodie peut apporter des informations sur l'état psychologique du locuteur : calme, énervé, triste, etc. Aussi, elle varie très largement suivant la provenance géographique, le niveau social et permet d'identifier un individu en tant que membre d'un groupe social ou culturel. L'exemple le plus frappant est sûrement l'accent régional comme celui du sud ou du nord de la France.

De plus, la prosodie permet de véhiculer l'attitude du locuteur envers l'interlocuteur et celle vis-à-vis de l'énoncé (adhésion plus ou moins forte). La manière de parler à un enfant peut facilement être différenciée de celle utilisée pour parler à un adulte. Le type d'intervention orale fait également appel à des prosodies différentes que ce soit pour une narratrice de contes populaires ou pour le discours d'un homme politique. La prosodie est liée à la stratégie de communication du locuteur.

1.5 Paramètres de la prosodie

Dans cette partie, nous allons présenter les trois principaux paramètres de la prosodie que sont la fréquence fondamentale, l'intensité et la durée.

1.5.1 Fréquence fondamentale

La mélodie de la voix résulte de la vibration des cordes vocales, et se traduit acoustiquement par l'évolution de la fréquence de vibration laryngienne en fonction du temps. D'un point de vue acoustique, on parle de *fréquence fondamentale* (F_0). Celle-ci est une

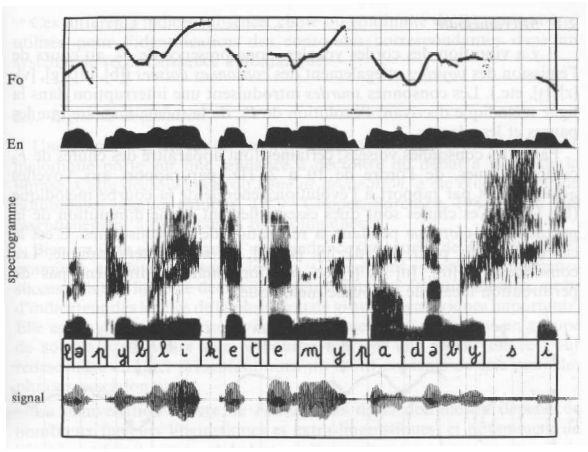


FIG. 1.7 – Spectrogramme de la phrase : « Le public est ému par Debussy » (extrait de (Calliope, 1989)). F_0 correspond à l'évolution de la fréquence fondamentale et E_n représente l'évolution de l'énergie. Concernant la fréquence fondamentale, on peut observer son absence sur certaines parties du signal (zones non voisées). De plus, on peut voir la forme d'onde particulière d'une courbe de fréquence fondamentale.

estimation de la fréquence laryngienne. Pour avoir une idée de l'ordre de grandeur de cette fréquence, nous pouvons rappeler les chiffres suivants :

- de 70 à 250 Hz pour les hommes
- de 150 à 400 Hz pour les femmes
- de 200 à 600 Hz chez les enfants.

Cependant, pour un même locuteur, les variations de la valeur de la fréquence fondamentale peuvent être considérables.

Une courbe mélodique n'est pas continue. La fréquence fondamentale existe au cours de l'émission des voyelles et également des consonnes voisées ([b], [d], [g], ...). Les consonnes sourdes introduisent une interruption dans la ligne mélodique décrivant l'évolution de F_0 (figure 1.7), de la même manière que les pauses et les silences. Certaines consonnes voisées font apparaître des chutes de F_0 caractéristiques, de l'ordre de 10Hz à 20Hz. Celles-ci sont dues essentiellement à une chute de la pression sous-glottique. Ces variations de la courbe mélodique définissent la micro-mélodie.

De manière générale, on observe une déclinaison de la ligne mélodique. Celle-ci est due essentiellement à des facteurs physiologiques. Elle peut être utilisée pour repérer les fins de phrases ou les groupes syntaxiques.

1.5.2 Intensité

L'intensité est perçue comme la force sonore de la voix. Son niveau est lié au fonctionnement des systèmes respiratoires et phonatoires et à la pression sous-glottique.

Sur le plan expérimental, l'intensité est influencée par la sensibilité de l'enregistrement, ce qui rend difficile l'étude et la comparaison de l'intensité à travers différents locuteurs.

1.5.3 Durée

Les variations de la durée sont analysées sur la durée des phonèmes ou des syllabes et sur les pauses. La hauteur des voyelles et la nasalité (lorsque l'onde produite est dirigée vers le conduit nasal, par exemple : [an], [on]) influencent positivement la durée.

L'ensemble des facteurs temporels est perçu comme le débit de la parole.

1.6 Caractérisation de la voix

Lors d'une communication téléphonique, il est aisé de reconnaître notre interlocuteur par la seule écoute de sa voix. L'indice qui nous permet cette identification est ce que l'on appelle le timbre de la voix. Ce terme est assez difficile à définir objectivement et la définition donnée par l'*American Standards Association* emploie même le terme de sensation : *« ... that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar ».* Le timbre reposerait donc sur un jugement subjectif de l'auditeur. Ceci montre une influence de l'auditeur sur les critères et la stratégie mise en place pour discriminer deux voix par rapport à leur timbre. Un autre point serait que le timbre est indépendant de la force sonore et du pitch.

Dans la suite, nous allons présenter quelques enjeux et applications issus de la problématique de caractérisation de la voix, puis nous présenterons quelques résultats qui reflètent la difficulté de caractériser l'identité d'une voix. Nous terminerons ce paragraphe par un bilan des principaux paramètres acoustiques utilisés dans ce domaine.

1.6.1 Enjeux et applications

La parole fait intervenir les trois niveaux linguistique, para-linguistique et extra-linguistique. Le niveau para-linguistique regroupe les facteurs qui caractérisent le locuteur à un niveau comportemental, par exemple ses émotions. Le niveau extra-linguistique est caractérisé par les propriétés physiologiques et physiques de l'appareil phonatoire

d'un locuteur. Kuwabara et Sagisaka (1995) distinguent la dimension socio/psychologique de la dimension physiologique. La première est à rapprocher du niveau paralinguistique et comprend des facteurs tels que l'âge le statut social, le dialecte et la communauté d'appartenance.

On peut noter que les facteurs impliqués sur les plans para-linguistique et extralinguistique peuvent varier non seulement entre deux locuteurs mais également pour un même locuteur. Comme l'indique Atal (1976), la problématique est alors de déterminer quels paramètres permettent d'expliquer la variabilité inter-locuteur de la voix, tout en étant insensibles à la variabilité intra-locuteur. Ainsi, le lien entre les facteurs qui influencent la perception de la voix et les paramètres acoustiques doit être déterminé. Il définit également les caractéristiques idéales de ces paramètres :

- La représentation de l'information dépendante du locuteur doit être efficace.
- L'acquisition du paramètre doit être facile.
- Le paramètre doit se montrer stable à travers le temps.
- Il doit apparaître naturellement et fréquemment dans la parole.
- Un imitateur ne doit pas pouvoir le reproduire.

De nombreux travaux de recherche existent sur la caractérisation de l'identité d'une voix. Les applications de ses travaux se situent au niveau de la biométrie ou encore de la synthèse de la parole. Dans le premier cas, il s'agit d'identifier une personne par les caractéristiques de sa voix. Les conséquences juridiques de cette application ne sont pas anodines et il convient d'être prudent (Bonastre *et al.*, 2003). Pour ce qui concerne la synthèse de parole, connaître les caractéristiques qui font l'identité d'une voix permettrait d'améliorer le naturel des voix de synthèse mais aussi de diversifier les voix de synthèse en modifiant les paramètres acoustiques importants.

1.6.2 Études perceptives

Pour identifier les caractéristiques propres à la voix d'un locuteur, on peut s'intéresser à la manière dont les auditeurs distinguent les voix les unes des autres. Ainsi, Hollien *et al.* (1982) montrent que le fait d'être familier avec un locuteur permet de l'identifier de manière assez fiable. Dans le cas contraire, une accoutumance est nécessaire pour atteindre un score plus élevé que le hasard. Ils montrent également que si un locuteur déguise sa voix, cela provoque une certaine confusion pour tous les auditeurs, même pour ceux qui sont familiers de la voix du locuteur en question.

Concernant les facteurs qui permettent d'identifier une voix, de nombreuses études dont les résultats sont souvent contradictoires existent. Certaines montrent l'importance de paramètres de nature supra-segmentale (débit et mélodie pour (Voiers, 1979; Atal,

1972)), d'autres présentent le F_0 moyen, la pente spectrale de l'onde de glotte et les trois premiers formants comme des facteurs prépondérants (Matsumoto *et al.*, 1973).

Necioglu *et al.* (1998) établissent une distinction entre voix d'hommes et voix de femmes, et classent les critères retenus par ordre d'importance perceptive. Pour les premières, la valeur médiane du pitch, puis la longueur du conduit vocal seraient les facteurs dominants tandis que pour les voix de femmes, il s'agirait de la valeur médiane du pitch, suivie de la pente spectrale de la source et de la durée moyenne des segments voisés.

En outre, ces différents travaux permettent de montrer qu'il existe des difficultés pour identifier les caractéristiques de la voix à travers des expériences perceptives. Dans celles-ci, l'auditeur joue un rôle central puisque chaque auditeur va appliquer des stratégies de reconnaissance différentes. Notamment, les auditeurs modifient leur stratégie suivant qu'il faut identifier une voix d'homme ou de femme (Singh et Murry, 1978). Les travaux de Schmidt-Nielsen et Crystal (2000) montrent que la tâche de vérification du locuteur est difficile pour un être humain et que les performances sont très variables suivant l'auditeur.

1.6.3 Principaux paramètres acoustiques

Dans le cadre d'une étude sur la reconnaissance automatique du locuteur, Atal (1976) dresse un bilan des principaux paramètres acoustiques utiles à la caractérisation de la voix :

- **Intensité.** Il s'agit de l'un des paramètres les plus faciles à obtenir. Pour des signaux non stationnaires comme la parole, l'intensité est définie en fonction du temps par $E(t) = \int_{t-T/2}^{t+T/2} s^2(\tau)d\tau$, où T est choisi arbitrairement. Sa valeur est habituellement comprise entre 10 et 30ms. Les variations de l'intensité de la parole sont causées par la variation de la pression sub-glottique et la forme du conduit vocal. Elle est liée à des caractéristiques dépendantes du locuteur.
- **Pitch.** La réalisation acoustique du pitch est la fréquence fondamentale, F_0. Les variations temporelles du pitch représentent une caractéristique importante de la parole et des travaux ont montré leurs importance pour la reconnaissance automatique du locuteur (Atal, 1972).
- **Spectre à court-terme.** Il s'agit d'une représentation à trois dimensions de la structure temps-fréquence du signal de parole. Le spectre à court-terme offre une description complète des caractéristiques acoustiques du signal. Cette information semble efficace pour la reconnaissance automatique du locuteur même si elle n'est pas très compacte.

- **Coefficients de prédiction.** La prédiction linéaire est une méthode efficace pour représenter les propriétés spectrales du signal de parole. Avec cette méthode, un échantillon est vu comme une combinaison linéaire des p échantillons passés.
- **Fréquence et bande passante des formants.** Les fréquences centrales des formants sont définies comme les fréquences de résonance du conduit vocal et sont dépendantes du locuteur. La principale difficulté consiste à estimer de manière efficace les formants à partir du spectre à court-terme.
- **Coarticulation nasale.** En parole continue, la forme du conduit vocal à un instant t dépend non seulement du phonème courant mais également des phonèmes voisins. Le phénomène de coarticulation résulte de l'influence du contexte phonémique sur le mouvement des articulateurs. Des travaux montrent que la coarticulation pendant la production de nasales permet de différencier les locuteurs.
- **Corrélation spectrale.** Un degré significatif de corrélation existe entre les spectres à court-terme à différentes fréquences. Ces corrélations sont obtenues par des moyennes de spectres sur le long terme et elles varient de manière consistante d'un locuteur à un autre.
- **Débit d'élocution et événements temporels.** La durée de certains événements dans la parole est différente d'un locuteur à un autre. Doddington (1985) propose de mettre en relation les événements de deux locuteurs par une déformation non linéaire de l'axe du temps.

Comme le notent Kuwabara et Sagisaka (1995), il n'existe pas de paramètre unique pour contrôler l'identité d'une voix. Au contraire, chaque paramètre acoustique joue un rôle dans l'identité de la voix. Necioglu *et al.* (1998) constatent également que c'est l'utilisation conjointe des différents paramètres qui peut donner de bons résultats dans le cadre de l'identification du locuteur.

1.7 Conclusion

Dans ce chapitre, nous venons de décrire la parole, tout d'abord sur le plan acoustique puis sur le plan physiologique. Du point de vue segmental, la parole peut être vue comme une succession de sons, de segments qui possèdent une structure particulière. Un niveau supra-segmental, celui de la prosodie, intervient à une échelle plus grande et constitue en quelque sorte la mélodie de la parole. Sur le plan physiologique, la production de parole fait intervenir de nombreux muscles et organes qui composent ce phénomène complexe.

Dans la troisième partie de ce chapitre, nous avons présenté de manière succincte ce

qu'est la prosodie. La prosodie apparaît comme essentielle pour la parole notamment pour son intelligibilité ou son naturel. Dans la suite de ce document, nous prendrons pour hypothèse que la fréquence fondamentale est un facteur prosodique prépondérant et nous nous focaliserons sur celui-ci. De plus, la quatrième partie de ce chapitre fait apparaître que les paramètres prosodiques contribuent également à la caractérisation de l'identité d'une voix.

Chapitre 2

La transformation de voix et ses applications

Dans ce chapitre, nous allons nous intéresser à la transformation de voix sous un angle général ainsi qu'à deux champs applicatifs possibles : la synthèse de la parole et la reconnaissance du locuteur.

Dans une première partie, le principe de la transformation de voix est présenté ainsi que les méthodes permettant de modifier les paramètres acoustiques d'un signal de parole afin de changer l'identité perçue par un auditeur. Dans une deuxième partie, nous détaillons le premier domaine d'application qu'est la synthèse de parole. Cette deuxième partie expose l'architecture d'un système de synthèse de la parole à partir du texte ainsi que les différentes étapes qui se succèdent dans un tel système. Enfin, la troisième partie décrit le domaine de la reconnaissance du locuteur qui partage les mêmes préoccupations scientifiques que la transformation de la voix, à savoir déterminer les paramètres les plus pertinents caractéristiques de l'identité d'une voix.

2.1 La transformation de voix

2.1.1 Principe

L'objectif de la transformation de voix est de modifier les caractéristiques de la voix d'un locuteur *source* pour qu'elle soit perçue comme celle d'un locuteur *cible*. Cette définition se place sous l'angle de la perception et cela implique qu'il ne s'agit pas que la voix transformée soit identique en tout point à la voix cible. Ce cadre de travail nécessite l'étude de paramètres acoustiques qui caractérisent l'identité de la voix et dont nous avons parlé au paragraphe 1.6 page 25.

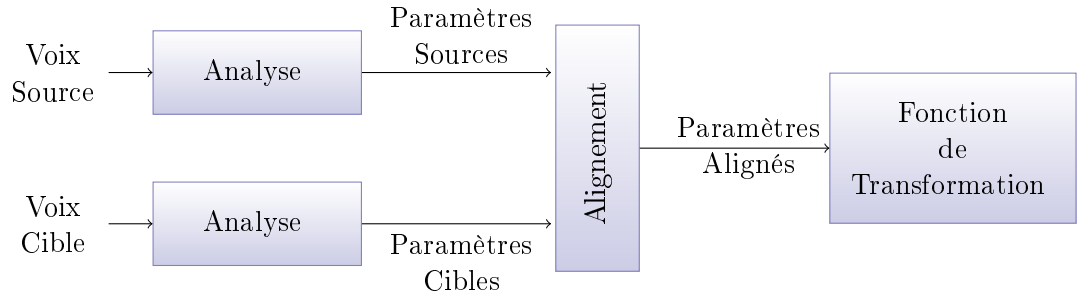


FIG. 2.1 – Phase d'apprentissage d'une fonction de transformation de la voix. Les signaux des voix source et cible sont analysés, puis alignés afin de calculer les paramètres de la fonction de transformation.

Dans ce paragraphe, nous allons décrire les grandes lignes de la construction et de l'utilisation des systèmes de transformation de la voix. Un état de l'art plus détaillé des systèmes de transformation de la voix est présenté dans (Kain, 2001).

La construction d'un système de transformation de la voix requiert des signaux de parole issus des deux locuteurs source et cible. Une phase d'apprentissage est réalisée et elle permet d'estimer les paramètres d'une fonction de transformation par analyse des signaux de parole source et cible comme le montre la figure 2.1. Une étape d'alignement temporel des signaux de parole, ou de manière équivalente des paramètres décrivant l'évolution de ce signal, est nécessaire afin de mettre en correspondance les vecteurs acoustiques source et cible. Lors de l'utilisation du système, le signal de parole issu du locuteur source est analysé afin d'estimer les valeurs des paramètres qui sont ensuite placées en entrée de la fonction de transformation (voir figure 2.2 page ci-contre). Celle-ci fournit en sortie les paramètres modifiés qui sont utilisés pour la synthèse du signal transformé. Au paragraphe 2.1.4 page 34, une description de fonctions de transformation usuelles est présentée.

Ce processus fait donc clairement apparaître une fonction d'analyse du signal qui repose sur un modèle de parole, une fonction de transformation qui modifie les paramètres issus de la voix source, et une fonction de synthèse qui permet de générer le signal transformé en appliquant les nouveaux paramètres.

2.1.2 Apprentissage de la transformation

Analyse du signal La majorité des systèmes de transformation de la voix s'appuient sur la modélisation de l'enveloppe spectrale du signal. Abe *et al.* (1988) utilisent les coefficients de prédiction linéaire, *LPC*, pour représenter le signal de parole. D'autres

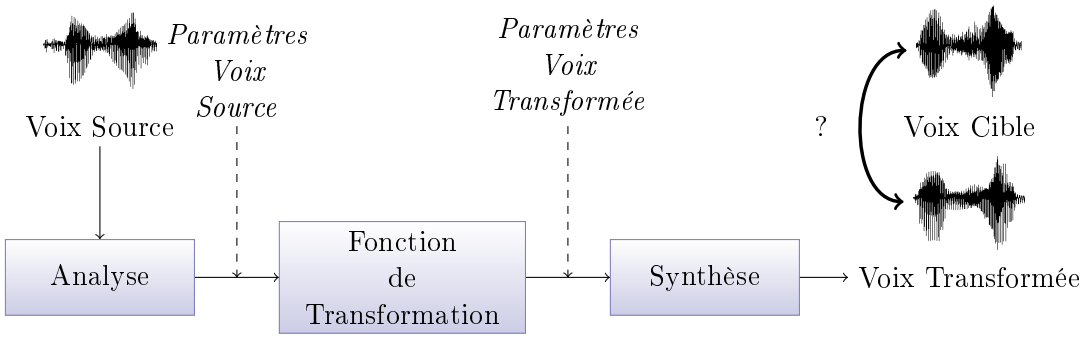


FIG. 2.2 – Phase d'utilisation d'une fonction de transformation de la voix. Lorsque l'on souhaite transformer les paramètres acoustiques d'une phrase source, la première étape est l'analyse du signal grâce au modèle de parole afin d'obtenir les paramètres à fournir en entrée de la fonction de transformation. Celle-ci donne en sortie, un jeu de paramètres modifiés qui permet de synthétiser le signal transformé.

représentations paramétriques peuvent également être employées : Kain et Macon (1998) et Arslan (1999) proposent de mettre en œuvre une représentation plus robuste avec les coefficients LSF obtenus à partir des coefficients *LPC* ; Stylianou *et al.* (1998) modélisent l'enveloppe spectrale à l'aide des coefficients cepstraux ; Mizuno et Abe (1995) s'appuient sur une représentation des formants.

Appariement des vecteurs source et cible Cette étape d'analyse permet d'obtenir pour chaque locuteur des séquences de vecteurs qui contiennent les paramètres du modèle de parole choisi. La création du corpus d'apprentissage est alors réalisée en regroupant par couples les vecteurs acoustiques source et cible. Un vecteur source et un vecteur cible sont assemblés si ceux-ci correspondent à la réalisation du même son. Cet assemblage peut être réalisé grâce à un alignement calculé par déformation non uniforme de l'axe temporel en appliquant un algorithme de type *DTW, Dynamic Time Warping.* Ce type d'appariement nécessite que les locuteurs aient prononcé le même texte, ce qui est un frein au développement des systèmes de transformation de la parole. Depuis quelques années, des travaux portant sur l'apprentissage d'une fonction de transformation à partir de données non alignées apparaissent (Mouchtaris *et al.*, 2004; Duxans *et al.*, 2006).

Apprentissage des paramètres de la fonction de transformation Ces données assemblées sont ensuite utilisées pour estimer la fonction de transformation dont l'objectif est de capturer le lien entre les caractéristiques de la voix source et celles de la

voix cible. En pratique, elle peut être mise en œuvre par des tables de correspondance, *mapping codebooks* (Abe *et al.*, 1988; Arslan, 1999), des mélanges de lois gaussiennes (Stylianou *et al.*, 1998; Kain et Macon, 1998), des réseaux de neurones (Narendranath *et al.*, 1995) ou encore des HMM (Duxans *et al.*, 2004).

2.1.3 Transformation d'une phrase

Il s'agit de trouver les paramètres qui permettent de décrire le signal de parole qu'aurait produit le locuteur cible, ou du moins s'en approcher le plus possible. La fonction de transformation estimée lors de la phase d'apprentissage est utilisée pour modifier les caractéristiques acoustiques d'une phrase prononcée par le locuteur source. Cette phrase est tout d'abord analysée grâce au même modèle que celui utilisé lors de l'apprentissage. La fonction de transformation permet de prédire les caractéristiques du signal cible qui sont ensuite utilisées par un modèle de synthèse.

Les paramètres prosodiques de la phrase, tels que le F_0, l'énergie et le débit, sont en général simplement ajustés pour qu'ils correspondent à la prosodie moyenne du locuteur cible. Dans les systèmes de transformation actuels, la prosodie n'est pas transformée de manière fine, et l'objectif de cette thèse est justement d'étudier et de proposer une stratégie de transformation de la prosodie plus élaborée que celles existantes.

2.1.4 Fonctions de transformation

Abe *et al.* (1988) proposent une méthode de transformation qui repose sur une table de correspondance entre les vecteurs du modèle *LPC* de la voix source et ceux de la voix cible et nomment cette méthode *mapping codebook*. Les entrées de la table sont des vecteurs issus de la quantification des espaces acoustiques source et cible. Un problème fondamental de cette technique est que l'ensemble des vecteurs cible possibles est discret ce qui introduit des discontinuités sur le signal de parole reconstruit. Des améliorations ont été apportées pour palier ce problème, notamment en introduisant un lissage entre les vecteurs acoustiques.

Certains travaux introduisent la notion de transformation locale : l'espace acoustique du locuteur source est partitionné en classes et une fonction de transformation est associée à chaque classe. Valbret *et al.* (1992) proposent deux fonctions de transformation locales : une régression linéaire multiple et un alignement dynamique fréquentiel (DFW). De manière similaire, Mizuno et Abe (1995) mettent en place un ensemble de règles de transformation pour chaque sous-espace obtenu par quantification vectorielle. Ce type de transformation qui manipule de manière continue une représentation du

signal de parole est capable de produire une infinité de vecteurs cibles. Cependant, il peut tout de même exister des discontinuités dans le signal de parole. Celles-ci sont dues au nombre fini de fonctions de transformation ainsi qu'au « saut » qui se produit à la frontière de deux classes.

Des fonctions de transformation continues ont également été proposées. Narendranath *et al.* (1995) fondent leur approche sur un réseau de neurones qui est utilisé pour capter la transformation entre les trois premiers formants de chaque locuteur. Stylianou *et al.* (1998) modélisent statistiquement les relations entre les enveloppes spectrales de deux locuteurs qui prononcent le même texte grâce à un modèle à mélange de lois gaussiennes, GMM, appris sur les données du locuteur source. Des variantes de cette approche existent, par exemple, Kain et Macon (1998) proposent d'utiliser un GMM conjoint source-cible. Mesbahi *et al.* (2007) montrent que les fonctions de transformation par GMM donnent des phrases converties qui ont le défaut d'être trop lisses (effet d'over-smoothing). De plus, le nombre de paramètres de ces modèles est important et il peut se produire un effet d'over-fitting si le nombre de vecteurs d'apprentissage est trop faible.

Plus récemment, Duxans *et al.* (2004) ont introduit la notion de séquence dans la fonction de transformation par l'intermédiaire de HMM. Il s'agit de la généralisation de l'approche GMM à des séquences de vecteurs acoustiques. Tous les états des HMM utilisés sont connectés et la distribution de probabilité d'une émission pour chaque état suit une loi gaussienne. De la même manière que pour les GMM, on peut apprendre un HMM pour la source ou bien un HMM conjoint. Dans ce même article, ils proposent également d'ajouter des informations phonétiques dans la fonction de transformation par le biais d'un *CART, Classification and Regression Tree.*

Les approches que nous venons de décrire nécessitent que les deux locuteurs aient prononcé les mêmes phrases. Des travaux récents tentent d'éliminer cette contrainte par la création de paires d'unités source-cible qui sont alignées à partir des données disponibles non alignées. L'approche proposée par Duxans *et al.* (2006) consiste à appliquer un algorithme de sélection d'unités pour créer ces paires alignées afin d'employer les méthodes de transformation de voix classiques. Une autre approche est d'apparier les espaces source et cible par une méthode hiérarchique descendante (Mesbahi *et al.*, 2008). Le principe est de construire deux arbres en parallèle et d'apparier les feuilles de l'arbre source avec celles de l'arbre cible. On peut noter que ces premiers travaux pour la transformation de la voix à partir de données non alignées donnent des résultats encourageants.

Nous verrons par la suite que dans bien des cas, les techniques étudiées pour la

transformation de la prosodie s'inspirent de celles déjà mises en place pour le segmental. Concernant le point particulier des techniques reposant sur des corpus non alignés, nous proposerons une méthode de transformation de ce type appliquée à la prosodie au chapitre 7 page 143.

Dans la suite de ce chapitre, deux applications possible de la transformation de voix sont présentées : la synthèse de parole à partir du texte et la reconnaissance du locuteur.

2.2 Synthèse de la parole à partir du texte

Dans Calliope (1989), la définition suivante est donnée : *« L'objectif de la synthèse de la parole est de produire des sons de parole à partir d'une représentation phonétique du message »*.

Cette définition inclut donc à la fois la synthèse de parole à partir du texte et la synthèse de parole à partir de concepts. Ces deux types de synthèse de parole diffèrent par la nature de l'entrée du système de synthèse. Le cas de la synthèse par concepts, qui suppose que le système ait accès à la sémantique du message, n'est pas traité dans cette partie.

Les applications de la synthèse de parole sont multiples. On peut par exemple citer le domaine de la communication homme/machine, pour lequel l'utilisation de la voix en entrée et en sortie du système apporte convivialité et simplicité à l'utilisateur. Dans le domaine des télécommunications, la lecture automatique de courriels ou la consultation d'annuaires sont deux services proposés par certains opérateurs. La synthèse de la parole peut également être utilisée pour l'aide aux personnes souffrant d'un handicap, que ce soit dans le cas de personnes ayant perdu l'usage de la parole ou bien pour les non-voyants ou mal-voyants. Cette liste d'applications possibles ou existantes n'est pas exhaustive et on pourrait l'étendre en ajoutant par exemple les jouets parlants et l'aide à la navigation.

Nous nous focalisons ici sur la synthèse de parole à partir d'une entrée textuelle, dont nous allons décrire l'architecture générale, avant d'étudier les différentes étapes de traitement nécessaires. La suite de cette partie est largement inspirée de l'ouvrage de Boite *et al.* (2000), aussi le lecteur intéressé pourra s'y référer avantageusement pour plus de précisions. L'article de Klatt (1987) et le livre Calliope (1989) sont également des références à consulter pour avoir une vue d'ensemble du domaine de la synthèse de parole.

Un système de synthèse de la parole à partir du texte peut être décomposé en deux blocs principaux comme le montre la figure 2.3. L'entrée textuelle subit tout d'abord

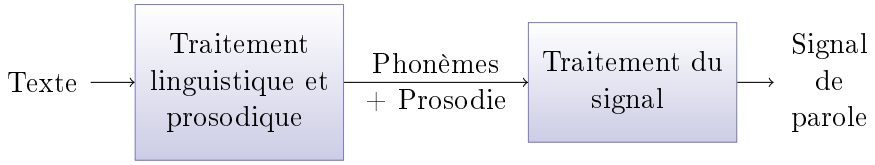


FIG. 2.3 – Architecture générale d'un système de synthèse de la parole à partir du texte.

des traitements de nature linguistique et prosodique. Ces derniers ont pour objectif de fournir la liste des phonèmes du texte à synthétiser ainsi qu'une prosodie adaptée à ce texte. Ces informations sont ensuite utilisées pour générer le signal de parole.

Le traitement linguistique et prosodique est constitué de la succession de trois grandes étapes :

- l'analyse morphosyntaxique qui permet d'obtenir les liens de dépendances syntaxiques entre les mots ainsi que la nature des mots à prononcer ;
- la phonétisation automatique dont le rôle est de fournir la transcription phonétique du message à synthétiser ;
- la génération de la prosodie qui doit trouver une intonation et un rythme acceptable pour la phrase.

Nous allons maintenant décrire ces trois points avant d'introduire différentes méthodes de synthèse du signal de parole.

2.2.1 Analyse morphosyntaxique

2.2.1.1 Pré-traitement

Le texte en entrée d'un système de synthèse de la parole peut avoir des origines très différentes qui influent sur sa forme et sa présentation. Par exemple, les textes d'une pièce de théâtre, d'un SMS, d'un courriel et d'un article de journal ne respectent pas les mêmes conventions d'écriture. Ainsi, le texte d'une pièce de théâtre contiendra des marques permettant d'attribuer la parole à un personnage ou à un autre. Un SMS contient des formes abrégées très particulières, comme par exemple « a 2m1 » pour « à demain ». On peut trouver des suites de caractères représentant des « smileys » dans un courriel, qui ne doivent pas être prononcés, mais qui transmettent plutôt une émotion ou un état d'esprit. Enfin, le titre d'un article est souvent écrit en majuscules sans qu'il ne s'agisse de plusieurs phrases ou d'une suite d'acronymes.

L'étape de pré-traitement est donc essentielle pour mettre le texte à synthétiser sous une forme exploitable par les structures de données internes au système de synthèse de

la parole. Cette normalisation du texte d'entrée doit notamment permettre de repérer les différentes entités qui apparaissent sous une forme contractée (abréviations, nombres, dates, etc.) ainsi que les limites de phrases.

Bien que les acronymes soient généralement notés avec un point comme séparateur de leurs lettres (« S.N.C.F. »), ils peuvent apparaître également sans le séparateur, ce qui ne pose pas de problème à un lecteur mais est très délicat pour une machine. Aussi les règles de prononciation peuvent être différentes, il est donc important de repérer correctement ces entités. Par exemple, pour « S.N.C.F. », chaque lettre est prononcée séparément tandis que « A.S.S.E.D.I.C. » est prononcé [asedik].

Les fins de phrases doivent également être détectées et distinguées des acronymes qui utilisent le même symbole. La même ambiguïté se pose avec les nombres qui possèdent un point comme séparateur des multiples de mille.

Concernant les nombres, les dates et les heures, il s'agit de les transcrire sous leur forme textuelle. Par exemple, la date suivante « 21/06/2006 » devra être remplacée par « vingt et un juin deux mille six ». Il en va de même pour les abréviations qui sont prononcées sous leur forme pleine, comme « c-à-d » qui sera remplacé par « c'est-à-dire ».

2.2.1.2 Analyse morphologique

L'analyse morphologique a pour tâche d'étudier les différentes manières de construire les mots. Il s'agit de proposer toutes les natures possibles pour chaque mot en fonction de sa graphie. La nature d'un mot s'exprime en termes de *morphèmes* qui représentent des unités élémentaires de sens, par exemple :

« soit » : morphème « être »
+ morphème du conditionnel présent

Ce sont des unités abstraites dans la mesure où un même morphème peut apparaître sous une multitude de formes appelées *allomorphes*. Le morphème du pluriel peut, par exemple, s'exprimer sous la forme de « s », de « x » ou de « nt ».

L'ensemble des morphèmes se divise en deux classes : les morphèmes grammaticaux et les morphèmes lexicaux. Les premiers renvoient à une catégorie grammaticale exprimant le nombre, le temps, le genre, le mode, l'aspect ou les connexions logiques tandis que les seconds renvoient à un concept empirique ou abstrait, comme des noms, des verbes et des adjectifs. Le nombre des morphèmes lexicaux est a priori infini.

La morphologie se divise en trois branches selon le mode de construction des mots :

- morphologie *inflexionnelle* : elle tient compte des caractéristiques telles que le genre, le nombre, le mode, le temps, la personne, etc. Cette opération ne modifie pas la catégorie syntaxique de la racine.

- morphologie *dérivationnelle* : elle s'intéresse à la construction de mots appartenant à des catégories syntaxiques différentes à partir d'un morphème de base.
- morphologie *compositionnelle* : elle étudie l'assemblage de plusieurs morphèmes pour former de nouveaux mots, i.e. *« porte + avions = porte-avions »*.

Les opérations d'inflexion et de dérivation sont réalisées par l'ajout de préfixes et de suffixes. L'analyse morphologique d'un mot est réalisée à l'aide d'un dictionnaire.

L'étude de la morphologie des mots est importante pour la synthèse de la parole du fait de son influence sur la prononciation des mots ainsi que sur la prosodie qui leur est associée.

2.2.1.3 Analyse syntaxique

À l'issue de l'analyse morphologique, toutes les natures possibles pour chaque mot sont disponibles. En effet, les mots étant évalués indépendamment les uns des autres, cela peut conduire à des ambiguïtés sur leur nature et par là même sur leur prononciation et leur prosodie. Par exemple, on trouve fréquemment des ambiguïtés nom/verbe en français (« couvent », « président », etc.).

Un premier objectif de l'analyse syntaxique est donc de faire intervenir le contexte des mots afin de réduire la liste des natures possibles pour chaque mot. Sur l'exemple « les poules du couvent couvent », l'analyse morphologique du mot « couvent » ne permet pas de déterminer s'il s'agit d'un nom ou d'un verbe. Par contre, si on considère ce mot dans le contexte de « du », on peut alors déduire qu'il s'agit d'un nom commun.

Un second objectif est de déterminer l'organisation hiérarchique du texte. Cette étape procède à l'examen de l'espace de recherche résultant de l'analyse contextuelle afin d'établir un découpage du texte en groupes de mots. Il est à noter que l'organisation hiérarchique d'un texte ainsi que sa prosodie dépend du niveau syntaxique mais également des niveaux sémantiques et pragmatiques. Idéalement, il faudrait tenir compte de ceux-ci pour pouvoir générer une prosodie adéquate. On peut tout de même noter que la plupart des systèmes TTS reposent sur de la parole lue, et que dans ce cas précis, la prosodie dépend essentiellement de la syntaxe. De manière récente, de plus en plus d'études portent sur la synthèse de parole expressive (Erickson, 2005).

On distingue généralement les analyseurs syntaxiques probabilistes (n-grammes, réseau de neurones), qui s'appuient sur les probabilités de transition entre les catégories syntaxiques des mots, et les analyseurs non-probabilistes qui reposent sur un ensemble de règles.

2.2.2 Phonétisation

La phonétisation a pour rôle de trouver la transcription phonétique d'un texte en s'appuyant sur les informations morphosyntaxiques obtenues à l'étape précédente. Elle repose sur l'usage d'un dictionnaire qui associe une prononciation à chaque mot. Cependant, un même mot peut être prononcé de plusieurs manières différentes, notamment en fonction de son contexte, ce qui rend l'utilisation d'un dictionnaire insuffisante. En particulier, les problèmes suivants peuvent survenir :

assimilation Ce phénomène est lié à des contraintes articulatoires qui impliquent la modification de certains traits phonétiques d'un phonème. Calliope (1989) indique qu'il s'agit d'un transfert d'une caractéristique d'un son vers un son immédiatement voisin. C'est le cas, par exemple, pour le trait de voisement et la nasalité. On obtient alors la modification d'un son en un autre, comme dans « médecin » qui peut se prononcer [mɛdəsɛ̃] ou [mɛtsɛ̃].

liaisons Lorsque la dernière consonne d'un mot, qui est suivi immédiatement d'une voyelle, est prononcée, alors il se produit une liaison. C'est le cas, par exemple, pour « ils aiment » (/ilzɛm/). Les liaisons peuvent être obligatoires, facultatives ou interdites. L'absence d'une liaison obligatoire est perçue comme une erreur de prononciation comme dans « mes amis ». Les liaisons facultatives dépendent essentiellement du style d'élocution (« Nous allons à Marseille ») et sont couramment omises.

homographes hétérophones Les homographes hétérophones sont des mots qui possèdent la même orthographe mais qui se prononcent différemment suivant leur contexte, leur fonction. C'est le cas de « couvent » qui peut être un nom ou un verbe. L'ambiguïté peut être levée par la catégorie grammaticale du mot.

mots inconnus Deux cas peuvent être distingués : les mots nouveaux et les noms propres. Le premier cas tient au fait que les langues sont en évolution perpétuelle et d'après Boite *et al.* (2000), on estime à 12000 le nombre de mots apparus en français entre 1980 et 2000. Concernant les noms propres, la principale difficulté est que leur prononciation dépend souvent de leur origine géographique supposée (Klatt, 1987). L'utilisation exclusive d'un dictionnaire ne peut pas résoudre ces problèmes.

e muet (schwa) L'élision d'un « e » ou son maintien est assez variable. En théorie, ce phénomène n'apparaît que si la disparition du « e » ne provoque le rapprochement

que d'au plus deux consonnes. Il dépend également de contraintes rythmiques et de variantes régionales.

On distingue deux classes de phonétiseurs selon qu'ils reposent sur un dictionnaire ou bien un ensemble de règles. Dans la seconde approche, des règles définissent la phonétisation des mots les plus courants et un dictionnaire d'exceptions de taille réduite est utilisé pour les autres mots. Plus de détails, ainsi que des références sont disponibles dans (Boite *et al.*, 2000).

2.2.3 Génération de la prosodie

Nous avons vu au paragraphe 1.3 page 17 que la prosodie possède une influence importante sur l'intelligibilité du signal de parole ainsi que sur le sens et la compréhension du message oral. Sur le plan acoustique, la prosodie se traduit par des variations de fréquence fondamentale, de durée et d'intensité. Dans le cadre de la synthèse de parole, il est important, pour le naturel et l'intelligibilité de la parole synthétique, de générer une prosodie de qualité. Il s'agit ici de déterminer la structure prosodique et la place des accents dans la phrase, de manière à générer des profils rythmique et mélodique adaptés.

2.2.3.1 Prédiction de la durée

Comme le note Boite *et al.* (2000), la plupart des systèmes TTS actuels cherchent à prédire la durée des phonèmes. Selon Goubanova et King (2008), il est préférable de prédire la durée des phonèmes plutôt que celle des syllabes ou des *Inter-Perceptual Center Groups - IPCG* (Barbosa et Bailly, 1994). Une première raison est qu'il existe un nombre restreint de phonèmes ce qui implique une quantité de données nécessaires moins grande que pour les syllabes. Une deuxième raison est qu'un modèle de durée syllabique doit permettre de retrouver la durée des phonèmes qui composent la syllabe. Au contraire, la prédiction de la durée des phonèmes permet de prendre en compte directement les facteurs linguistiques qui influencent leur durée sans devoir effectuer une transition avec les syllabes. Néanmoins, comme le note Boite *et al.* (2000), les effets syllabiques et lexicaux devraient également être pris en compte dans les systèmes TTS.

Une méthode classique consiste à assigner aux phonèmes leur durée moyenne *intrinsèque* (calculée sur un corpus), et de modifier ces valeurs en faisant intervenir des facteurs *cointrinsèques* (durées intrinsèques des phonèmes voisins) et linguistiques de type multiplicatif ou additif.

D'autres techniques permettent de prédire la durée des phonèmes par le biais d'un

apprentissage automatique. On peut notamment citer l'utilisation d'arbres de régression et de classification (Riley, 1990; Malfrère *et al.*, 1998), de réseaux de neurones (Boeffard et Emerard, 1997) ou encore l'utilisation récente de réseaux bayésiens (Goubanova et King, 2008).

2.2.3.2 Prédiction de la fréquence fondamentale

Comme nous l'avons mentionné précédemment, sur le plan acoustique, l'intonation se résume principalement à l'évolution de la fréquence fondamentale, F_0. Dans le cadre de la synthèse de la parole, la prédiction du F_0 est en général précédée d'une étape de modélisation ou de stylisation pour laquelle de nombreuses techniques existent (cf. chapitre 3 page 53). Cette étape de modélisation permet de fournir un modèle de l'évolution du F_0 et ainsi de réduire l'ensemble des phénomènes intonatifs à un jeu de paramètres restreint. Il s'agit alors, le plus souvent, de prédire les paramètres d'entrée du modèle de prosodie à partir d'informations morphosyntaxiques. On peut noter l'existence d'approches par réseaux de neurones (Traber, 1992; Morlec *et al.*, 1995), par arbres de régression et de classification (Dusterhoff *et al.*, 1999) ou encore par sélection d'unités prosodiques existantes (Malfrère *et al.*, 1998).

2.2.4 Synthèse du signal de parole

Dans la première partie de ce paragraphe, nous passons en revue les principaux modèles de production de la parole. Ces modèles de production sont utilisés dans les systèmes de synthèse de la parole afin de représenter et de générer le signal de parole. Ainsi, dans la deuxième partie de ce paragraphe, nous présentons les deux méthodes de synthèse du signal de parole qui peuvent être distinguées : la synthèse par règles et la synthèse par concaténation d'unités acoustiques.

2.2.4.1 Modélisation du signal de parole

Modèle articulatoire Ce type de modèle de la parole est le premier apparu, il repose sur un modèle articulatoire de la parole qui tente de décrire le conduit vocal et de reproduire les mécanismes naturels de la production du signal de parole (Breen, 1992; Coker, 1976). Les paramètres de ces modèles ont une interprétation physiologique directe. Les systèmes de synthèse articulatoire constituent également un outil pour l'étude des mécanismes de production de la parole. Selon Shadle et Damper (2001), ces systèmes possèdent un fort potentiel dans le long terme.

Modèle formantique Les formants et anti-formants correspondent aux résonances et aux anti-résonances du conduit vocal. Ils apparaissent sous la forme de pics et de vallées sur l'enveloppe spectrale d'un signal de parole et sont caractérisés par leur fréquence centrale, leur largeur de bande et leur amplitude (voir 1.1 page 9).

La synthèse par formants consiste à générer un signal de parole à partir de l'évolution temporelle des paramètres qui caractérisent les formants. Contrairement au modèle articulatoire, la modélisation par formants utilise seulement l'information contenue dans le signal de parole. Néanmoins, la détection des formants dans le signal de parole n'est pas aisée, notamment lorsqu'ils sont proches ou se croisent au cours du temps.

À la synthèse, chaque formant est réalisé par un résonateur du second ordre. Les résonateurs peuvent être connectés en série ou en parallèle (Klatt, 1980; Holmes, 1973).

Modèle par prédiction linéaire Il repose sur l'hypothèse source/filtre selon laquelle le conduit vocal agit à la manière d'un filtre en atténuant ou renforçant certaines fréquences. Le modèle de production de parole consiste en un filtre AR excité soit par un train d'impulsions quasi-périodique, pour les sons voisés, soit par un bruit blanc, pour les sons non voisés (Makhoul, 1975).

L'estimation des paramètres du filtre est effectuée de façon à minimiser l'erreur quadratique moyenne entre le signal original et le signal prédit sur une fenêtre donnée. Le signal de parole est supposé stationnaire sur cette fenêtre de l'ordre de 30ms.

Modèle hybride harmonique/stochastique Cette modélisation, encore appelée harmonique+bruit, suppose que le signal de parole est la superposition d'une composante harmonique et d'une composante de bruit. La première est principalement associée aux sons voisés tandis que la seconde est associée aux sons non voisés ainsi qu'à certaines phases transitoires. On trouve ainsi le modèle à excitation multibande (Griffin, 1988), le modèle HNM (Stylianou *et al.*, 1995), et ceux proposés par Abrantes *et al.* (1991) ou d'Alessandro *et al.* (1998).

Modèle HMM Le modèle HMM, présenté par Masuko *et al.* (1996), modélise simultanément l'évolution du spectre du signal, du pitch et de la durée des phonèmes. Il prend en compte la dynamique de ces paramètres en considérant les dérivées première et seconde des coefficients cepstraux et du F_0. Cette modélisation d'abord appliquée au japonais a également été mise en œuvre pour l'anglais (Tokuda *et al.*, 2002).

2.2.4.2 Synthèse par règles - Synthèse par concaténation d'unités

La synthèse par règles consiste à modéliser les transitions entre phonèmes sous la forme de règles. Le modèle formantique de parole est le plus couramment utilisé pour ce type de synthèse. Dans ce contexte, les règles définissent l'évolution temporelle des paramètres de formants pour réaliser les transitions entre phonèmes. Le principal avantage de cette technique est la faible quantité de données nécessaires. Cependant les règles sont difficiles à établir et leur nombre peut être très élevé (Calliope, 1989).

La synthèse par concaténation d'unités consiste à accoler des segments acoustiques de base et pré-existants. Cette approche est fondée sur l'utilisation d'une base de données constituée de segments de parole. Les segments de parole utilisés doivent prendre en compte les phénomènes de coarticulation. On peut considérer que la parole est une succession de zones transitoires qui s'intercalent entre des zones de stabilité acoustique. Les diphones permettent de tenir compte de la zone de transition entre deux phones en considérant que la partie centrale d'un phone est stable. Un diphone est donc constitué de deux demi-phones et s'étend du centre de la réalisation du premier phone à celui de la réalisation du deuxième. Des unités de plus grande taille, comme les triphones ou les quadriphones, peuvent également être définies. La généralisation de ces unités conduit à l'utilisation d'unités de taille variable afin de minimiser les effets de la concaténation.

Les unités à concaténer sont en général extraites dynamiquement à l'aide d'un algorithme de sélection d'unités. Celui-ci cherche à minimiser le coût de sélection global qui tient compte du contexte phonétique et prosodique ainsi que du degré de continuité entre les unités sélectionnées.

La qualité de la parole synthétique dépend essentiellement de la variabilité contextuelle des unités contenues dans la base de données ainsi que de l'algorithme de sélection d'unités utilisé. Des modifications spectrales et prosodiques sont nécessaires au niveau des zones de jonctions entre unités afin d'approcher une transition naturelle entre les sons.

2.2.5 Synthèse de la parole et transformation de la voix

Lors de la création d'une voix de synthèse, une grande quantité de données doit être utilisée. Ces données audio sont acquises dans une salle d'enregistrement en faisant appel à un locuteur qui va prêter sa voix au système. Une fois cette étape fastidieuse d'enregistrement effectuée, il est encore nécessaire de traiter les données du corpus ainsi créé afin de l'analyser, de l'annoter, de le segmenter, d'extraire le F_0, etc. Ces différentes étapes sont souvent effectuées ou du moins vérifiées de manière manuelle.

Ce processus de création d'une nouvelle voix que nous venons de décrire est long, coûteux et doit être reproduit à chaque fois que l'on souhaite créer une nouvelle voix. Dans ce contexte, on voit assez bien l'intérêt d'une méthodologie de transformation de voix. Faisons l'hypothèse que l'on dispose d'une voix de référence pour laquelle tous les traitements précédent ont été effectués. Il serait alors possible de créer une nouvelle voix par transformation de cette voix de référence en utilisant un corpus de taille réduite plus facile à acquérir.

Une telle méthodologie permettrait donc de diversifier les voix de synthèse existantes et de les personnaliser de manière plus aisée. Bien entendu, on peut y trouver des enjeux commerciaux importants (réduction des coûts, nouvelles applications possibles, etc.) mais des enjeux scientifiques sont également présents. En effet, savoir transformer l'identité d'une voix signifie connaître les paramètres qui font cette identité. Nous allons ainsi voir dans le paragraphe suivant, un autre champ d'application de la transformation de voix pour lequel l'objectif est de reconnaître automatiquement l'identité d'un locuteur par sa voix.

2.3 Reconnaissance du locuteur

2.3.1 Définition

La reconnaissance du locuteur est un terme générique qui recouvre l'ensemble des tâches où il s'agit de différencier des individus par les caractéristiques de leur voix. Généralement, on distingue deux classes de systèmes : la vérification ou authentification du locuteur et l'identification du locuteur. La première consiste à accepter ou rejeter l'identité prétendue d'un locuteur en déterminant si l'échantillon fourni par ce locuteur est suffisamment proche de l'échantillon de référence associé à cette identité. La deuxième classe, l'identification du locuteur, consiste à trouver parmi une population de N locuteurs lequel est le plus proche de l'échantillon fourni par un locuteur inconnu. Rosenberg (1976) note que le problème d'identification est plus complexe que celui de vérification. L'argument qu'il avance est que, dans le premier cas, la probabilité de se tromper est une fonction croissante de la taille de la population N alors que dans le second, cette probabilité est indépendante de N. Dans ce paragraphe, nous allons nous focaliser sur les systèmes de vérification du locuteur qui ont été plus largement étudiés et dont des applications réalisables à court-terme existent déjà.

Il existe également différents types de systèmes de reconnaissance du locuteur suivant les hypothèses utilisées. Ainsi, certains systèmes sont dits dépendants du texte dans le sens où le texte utilisé pour la comparaison est fixé par le système. Une autre hypothèse

couramment employée est que le locuteur est coopératif, c'est-à-dire qu'il ne tente pas d'altérer les caractéristiques de sa voix. La qualité de la prise de son est également très importante puisque le microphone capte non seulement la voix du locuteur mais également les bruits de l'environnement.

Les applications de la reconnaissance du locuteur sont nombreuses : sécurisation des cartes de crédit, contrôle d'accès à des bases de données, paiements sécurisés distants, accès distants à un réseau d'ordinateurs, etc. Il est également nécessaire de prendre garde aux applications potentielles de l'identification de la voix dans le domaine judiciaire. Ainsi, un article récent, (Bonastre *et al.*, 2003), attire l'attention sur l'utilisation abusive de la reconnaissance du locuteur et prévient qu'à l'heure actuelle, aucun système ne peut identifier un individu à partir de sa voix avec une certitude absolue.

2.3.2 Sources d'erreurs de reconnaissance

La reconnaissance du locuteur apporte un certain confort pour l'utilisateur puisque les attributs biométriques ne peuvent pas être oubliés ou perdus. Néanmoins, la voix d'un individu est un attribut dynamique qui dépend à la fois du locuteur et de son environnement. Cela rend son utilisation plus complexe que d'autres attributs biométriques tels que les empreintes digitales ou la géométrie de la main qui reposent sur des caractéristiques physiques statiques.

Les sources d'erreurs de reconnaissance sont multiples :

- Erreur de lecture ou de prononciation.
- État émotionnel : l'état émotionnel d'un locuteur peut notamment altérer le timbre de sa voix, son débit d'élocution ou encore l'évolution du F_0.
- Maladie : les maladies affectant les voies respiratoires ou les organes de la phonation peuvent altérer de manière importante la voix.
- Temps : pendant l'intervalle de temps qui sépare l'enregistrement de la référence et celui de l'échantillon de test, les caractéristiques de la voix peuvent évoluer. L'âge est, par exemple, un facteur qui modifie les caractéristiques de la voix. Il est donc important de réactualiser les échantillons de référence pour chaque locuteur de manière régulière.
- Environnement : les modifications de l'environnement (microphone, position du microphone, bruit environnant, etc.) affectent la qualité de l'enregistrement.

(Campbell, 1997) note que ces facteurs sont généralement indépendants des algorithmes et sont résolus de manière plus efficace par d'autres moyens, par exemple avec le choix d'un microphone de meilleur qualité. Cette liste de sources d'erreurs permet de se rendre compte que la performance des systèmes de reconnaissance du locuteur sera

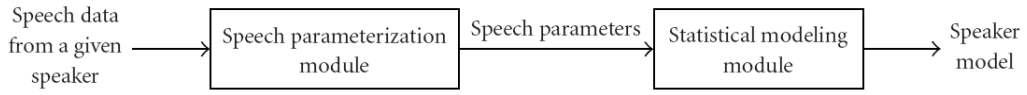


FIG. 2.4 – Phase d'apprentissage d'un système de vérification du locuteur. Le signal de parole est analysé afin d'en obtenir une paramétrisation. L'approche d'un modèle de la voix du locuteur est ensuite réalisée à partir des jeux de paramètres obtenus. Cette phase d'apprentissage correspond à une phase d'enrôlement de l'utilisateur (extrait de (Bimbot *et al.*, 2004)).

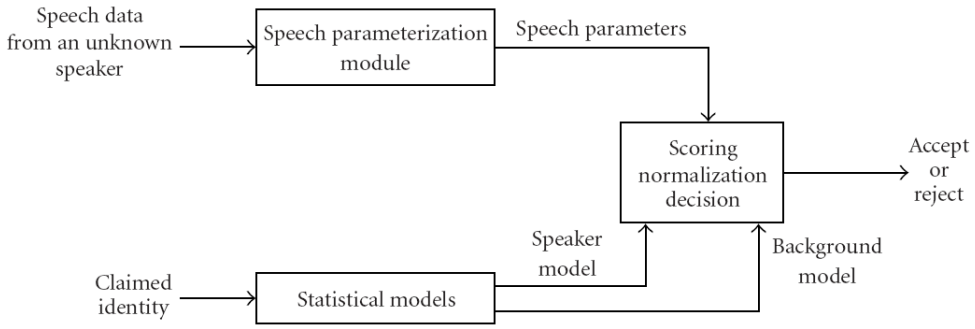


FIG. 2.5 – Phase de test d'un système de vérification du locuteur. Lors de l'utilisation du système, le signal de parole du locuteur inconnu est analysé avec le même module que lors de la phase d'apprentissage. Les paramètres obtenus sont ensuite utilisés afin d'effectuer une comparaison avec le modèle du locuteur qui possède l'identité proclamée. Un modèle des autres locuteurs, *background model*, peut être utilisé conjointement au modèle précédent pour le calcul du score de vraisemblance. La décision d'acceptation ou de rejet dépend du score obtenu par le locuteur inconnu et le seuil d'acceptation fixé dans le système (extrait de (Bimbot *et al.*, 2004)).

toujours limitée par l'erreur humaine (erreur de lecture ou de prononciation).

2.3.3 Architecture

L'approche générale d'un système automatique de vérification du locuteur comprend cinq étapes : acquisition du signal de parole, extraction des attributs caractéristiques, comparaison de l'échantillon avec le modèle de référence, décision d'acceptation ou de rejet et enregistrement d'un nouveau locuteur (Campbell, 1997). Tout comme pour la transformation de la voix, il est nécessaire d'identifier les caractéristiques propres de la voix d'un locuteur.

L'étape d'enrôlement correspond à la phase d'apprentissage sur les données d'un

nouveau locuteur. Cette dernière est schématisée sur la figure 2.4 page précédente extraite de (Bimbot *et al.*, 2004). La première étape est l'analyse du signal de parole afin d'en obtenir les paramètres caractéristiques. Ceux-ci sont ensuite utilisés afin d'apprendre un modèle de la voix du locuteur. La phase d'apprentissage est effectuée sur la base d'une identité vérifiée et pour chaque locuteur, on construit un couple modèle/identité.

La figure 2.5 page précédente présente le processus de test d'un système de vérification du locuteur. En entrée du système, on trouve l'identité proclamée par le locuteur ainsi qu'un échantillon de parole. Cet échantillon de parole est analysé en utilisant la même méthode que celle utilisée pour l'apprentissage des modèles dépendants du locuteur. Le modèle associé à l'identité proclamée est sélectionné ainsi qu'un autre modèle, appelé *Background Model*, qui représente « le reste du monde ». La dernière étape compare les paramètres de l'échantillon de parole avec les modèles retenus, calcule un score et prend la décision d'acceptation ou de rejet de l'identité. Typiquement, si le score entre l'échantillon et le modèle dépendant du locuteur est inférieur à un seuil fixé, alors l'identité est acceptée, dans le cas contraire elle est rejetée.

Nous venons ici de décrire le fonctionnement d'un système de vérification du locuteur. Dans le cas d'un système d'identification, le locuteur ne proclame aucune identité, et il faut effectuer une comparaison par rapport à chaque locuteur enregistré. On peut noter que le locuteur peut ne pas être enregistré dans le système, il faut dans ce cas être en mesure de le détecter. On remarque de manière intuitive que la complexité de ce système est plus importante que pour la vérification seule. De plus, dans le cas où la population des locuteurs est de grande taille, il est souhaitable de disposer de techniques permettant de limiter le nombre de comparaisons (Atal, 1976).

2.3.4 Choix des caractéristiques de la voix

Le choix des attributs qui caractérisent la voix d'un locuteur est une tâche primordiale pour la reconnaissance du locuteur. Nous avons vu au paragraphe 1.6 page 25 un certain nombre d'attributs acoustiques importants pour caractériser l'identité d'une voix. Cette liste proposée par Atal (1976) contient le F_0, l'intensité, les événements temporels, le spectre à court-terme du signal, les coefficients de prédiction LP, la fréquence des formants, la coarticulation nasale et la corrélation spectrale.

Atal (1976) note que la sélection des attributs dans un tel système doit être effectuée en utilisant un critère d'efficacité raisonnable. Idéalement, il doit rendre possible l'ordonnancement des attributs suivant leur efficacité. Ce critère doit également permettre d'estimer si les distributions de chaque locuteur pour l'attribut testé sont éloignées les

unes des autres dans l'espace de cet attribut. Le *F-ratio* repose sur cette idée et calcule le ratio entre la variance inter-locuteur et intra-locuteur pour l'attribut. Cependant, un inconvénient de ce critère est qu'il ne tient pas compte des corrélations qui peuvent exister entre les attributs.

Dans leur tutoriel récent sur les systèmes d'identification de la voix, Bimbot *et al.* (2004) mentionnent que la plupart des systèmes de vérification du locuteur reposent sur une représentation cepstrale du signal de parole. Ils utilisent donc les coefficients cepstraux que l'on peut extraire du signal par une transformée de Fourier rapide ou en appliquant un algorithme de prédiction linéaire. Ils mentionnent également l'importance de la dynamique des coefficients cepstraux captée par les coefficients delta et delta-delta associés, ainsi que l'évolution de l'énergie.

Des attributs de plus haut niveau permettent également de caractériser la voix d'un locuteur (Doddington, 1985). La langue, le sujet de discussion, le style de parole, les tournures utilisées, ou encore les expressions apportent des informations sur l'identité du locuteur. Malgré tout, ces informations sont assez difficilement accessibles pour les systèmes automatiques de reconnaissance du locuteur.

2.3.5 Reconnaissance dépendante/indépendante du texte

Comme le note Doddington (1985), le degré de contrôle sur l'acquisition du signal de parole est un paramètre très important. Ainsi, les systèmes de reconnaissance du locuteur peuvent être distingués suivant que le texte devant être prononcé par le locuteur est imposé ou non. On obtient alors des systèmes dépendants ou indépendants du texte.

Dans le premier cas, on dispose pour chaque locuteur enregistré d'un échantillon de parole correspondant au texte de contrôle. Lorsqu'un locuteur souhaite s'identifier, il prononce le texte imposé et proclame son identité. Il est alors nécessaire de comparer l'échantillon obtenu à celui ou ceux de référence dont le système dispose. Deux techniques peuvent être utilisées pour effectuer cette comparaison :

- **DTW.** Les échantillons sont stockés sous la forme de séquences de vecteurs acoustiques. Une ou plusieurs prononciations du texte de référence peuvent être mémorisées. La DTW, Dynamic Time Warping, permet d'obtenir le coût d'alignement entre le signal de test et celui de référence en calculant une déformation de l'axe temporel optimale.
- **HMM.** Pour chaque locuteur, un HMM est entraîné à partir de plusieurs prononciations du texte de référence. Le score est obtenu par l'algorithme de Viterbi qui permet de calculer la meilleure séquence d'états du HMM sachant la séquence de vecteurs acoustiques de l'échantillon de test.

Les systèmes dépendants du texte sont bien adaptés à l'hypothèse selon laquelle le locuteur inconnu souhaite être identifié et est donc coopératif. Dans le second cas, approche indépendante du texte, le contrôle sur le texte ne peut pas être maintenu soit parce que le locuteur n'est pas coopératif soit parce que la vérification doit être effectuée avec discrétion. Dans ce cas, le système ne connaît pas la séquence de mots qui va être prononcée ; cela implique qu'il est inutile de stocker des enregistrements directs de la voix des différents locuteurs. Bimbot *et al.* (2004) présentent une technique qui modélise la voix d'un locuteur par un modèle à mélange de lois gaussiennes (GMM). Ce modèle peut être appris directement sur les données d'enrôlement du locuteur ou bien par adaptation des paramètres d'un modèle indépendant du locuteur. Bimbot *et al.* (2004) citent également des alternatives à l'utilisation des GMM comme les réseaux de neurones et les SVM, *Support Vector Machines*.

Que le système soit dépendant ou indépendant du texte, il est nécessaire de savoir à partir de quel score (valeur de vraisemblance entre l'échantillon et le modèle du locuteur) l'échantillon de parole va être considéré comme appartenant au locuteur. La valeur du seuil de décision reste un problème ouvert dans le domaine de la vérification du locuteur. Ce seuil est notamment confronté à la variabilité intra-locuteur, à la variabilité inter-locuteur (dans le cas où le seuil de décision est indépendant du locuteur) et aux conditions environnementales. La normalisation du score a ainsi été introduite pour traiter la variabilité du score induite par ces différents facteurs. Nous n'allons pas détailler ici les différentes méthodes de normalisation du score, une liste de celles-ci est proposée dans (Bimbot *et al.*, 2004).

2.3.6 Reconnaissance du locuteur et transformation de la voix

La reconnaissance du locuteur est fondée sur l'utilisation des caractéristiques de l'identité de la voix. Dans la mesure où la transformation de la voix modifie ces mêmes caractéristiques afin de transformer la voix d'un locuteur pour qu'elle soit perçue comme celle d'un autre locuteur, ces deux domaines peuvent être considérés comme étroitement liés.

Même si ce n'est pas l'objectif visé par la transformation de voix, les techniques mises en œuvre dans ce domaine peuvent notamment être utilisées pour mettre à l'épreuve les systèmes de reconnaissance du locuteur. Par exemple, Bonastre *et al.* (2006) montrent qu'il est possible de tromper un système de reconnaissance du locuteur en modifiant les paramètres de la voix d'un imposteur de manière automatique. Cela montre que les avancées dans le domaine de la transformation de la voix peuvent apporter des éléments nouveaux pour la reconnaissance du locuteur.

2.4 Conclusion

Dans la première partie de ce chapitre, la transformation de la voix a été présentée. L'objectif de ces systèmes est de modifier les paramètres acoustiques d'une phrase issue d'un locuteur source de sorte qu'elle soit perçue comme si elle était prononcée par un locuteur cible. Dans cette partie, nous avons décrit les différentes étapes nécessaires à la transformation de la voix et présenté les méthodes existantes pour réaliser cette transformation. On peut noter que la transformation de la prosodie dans les systèmes actuels consiste simplement en l'ajustement des valeurs moyennes des paramètres prosodiques entre deux locuteurs. Ce constat constitue la base de ce travail de thèse dont l'objectif est de proposer une méthodologie de transformation de la prosodie.

Dans une deuxième partie, nous avons défini la synthèse de la parole à partir du texte et décrit l'architecture d'un tel système. Les différentes étapes, de l'analyse du texte à la génération du signal de parole, ont également été présentées. En particulier, nous pouvons retenir que les caractéristiques prosodiques du texte à synthétiser sont déterminées principalement à partir de données d'ordre syntaxiques même si elles dépendent également des niveaux sémantiques et pragmatiques.

L'architecture et les principes généraux du domaine de la reconnaissance du locuteur ont été décrits dans la troisième partie. Celle-ci est orientée vers la tâche de vérification du locuteur qui consiste à valider l'identité proclamée par un utilisateur à partir d'un échantillon de sa voix. De nombreuses sources d'erreurs, dont l'utilisateur lui-même, limitent les performances de ces systèmes.

Pour les deux domaines de la synthèse de parole et de la reconnaissance du locuteur, nous avons également discuté du lien qui existe entre ces domaines et la méthodologie de transformation de la voix. Ainsi, nous avons décrit l'impact que peut avoir la transformation de la voix sur la synthèse de la parole au niveau de la création et de la diversification des voix de synthèse. Concernant la reconnaissance du locuteur, la transformation de la voix pourrait apporter de nouveaux éléments pour caractériser l'identité d'une voix.

Chapitre 3

Modélisation de la prosodie : un état de l'art

Ce chapitre introduit les modèles classiques qui permettent de modéliser la prosodie et plus particulièrement la fréquence fondamentale, F_0. La modélisation de la prosodie est une étape très importante pour à la fois l'aspect explicatif des phénomènes prosodiques et l'aspect génératif, en particulier pour ce qui concerne les systèmes de synthèse de la parole.

Deux sous-domaines d'étude peuvent être dissociés au sein des travaux concernant la modélisation de la prosodie. Le premier s'attache à styliser les phénomènes observés, soit en utilisant des systèmes qui possèdent un ancrage linguistique, phonologique voire physiologique, soit par l'utilisation de modèles mathématiques fonctionnels sans aucun pouvoir explicatif des phénomènes sous-jacents. Le second point concerne la classification des contours mélodiques. Il a pour objectif de dégager d'un ensemble d'observations des classes représentatives du locuteur afin de ne conserver qu'une quantité restreinte d'informations pour représenter la prosodie de celui-ci.

Ainsi, la première partie de ce chapitre traite de la stylisation de la fréquence fondamentale qui est un paramètre prosodique considéré comme proéminent. Étant donné le nombre important de travaux dans le domaine de la stylisation de la prosodie, nous traitons ici les principaux modèles. La seconde partie de ce chapitre est quant à elle consacrée à la classification des contours mélodiques qui peut être considérée comme un prolongement de la stylisation.

3.1 Stylisation de la fréquence fondamentale

La stylisation peut être définie comme la simplification de l'observation d'un phénomène physique. Dans sa thèse, de Tournemire utilise cette définition et effectue une distinction entre stylisation et modélisation. Cependant, la simplification d'une courbe mélodique peut également être réalisée par l'intermédiaire d'un modèle comme c'est le cas pour MoMel (paragraphe 3.1.6 page 62). Dans ce document, nous considérons que le terme modélisation englobe la stylisation et est plus général.

La stylisation de la fréquence fondamentale est un sujet étudié depuis de nombreuses années et les modèles présentés comme une solution possible à ce problème sont nombreux. Dans la suite de cette section, nous allons présenter les principales approches qui permettent de styliser le F_0.

3.1.1 Tones and Break Indices

ToBI (Tones and Break Indices) est un modèle phonologique de l'intonation (Silverman *et al.*, 1992; Wightman, 2002). Ce système symbolique de transcription permet de représenter deux aspects de la prosodie présents dans la parole :

- les accents, qui contribuent à la proéminence relative d'un mot dans une phrase ;
- le phrasé, qui permet de créer des regroupements de mots.

Ainsi ToBI a été développé pour fournir un système de transcription unique pouvant être utilisé par de nombreux laboratoires de recherche afin d'annoter la structure prosodique d'une grande diversité de phrases.

La transcription ToBI d'une phrase nécessite au minimum l'enregistrement de la phrase, l'extraction du contour de F_0 associé et enfin un ensemble de symboles décrivant les événements prosodiques. Prenons par exemple le système ToBI pour l'anglais américain dont les conventions d'utilisation sont décrites dans (Beckman et Hirschberg). Il comprend quatre niveaux de description liés au temps. Trois niveaux sont obligatoires :

- le niveau orthographique (orthographic tier) qui décrit les mots de l'énoncé ;
- le niveau des indices de coupure (break index tier) qui indique les degrés de jonction entre les mots. Dans le système américain, cinq niveaux de coupure existent :
  - 0 : lorsque la frontière entre deux mots est inexistante ;
  - 1 : la frontière typique entre deux mots d'un énoncé ;
  - 2 : lorsque la frontière entre deux mots ne correspond pas à la frontière attendue ;
  - 3 : la frontière de groupe mineur ;
  - 4 : la frontière de groupe majeur ou groupe intonatif ;
- le niveau des tons (tonal tier) qui décrit l'évolution des accents de pitch, des

accents de phrase et des frontières de tons. Les deux tons de référence sont les tons haut (H) et bas (L).

Un quatrième niveau, le niveau « miscellaneous tier », permet de décrire des phénomènes additionnels et est facultatif. On peut annoter grâce à ce niveau des effets comme l'hésitation, le rire, l'inspiration et les autres effets spontanés de la parole. Un exemple de transcription est présenté sur la figure 3.1 page suivante. Sur cet exemple, on peut noter l'utilisation particulière de l'indice de coupure 2 qui marque ici une inadéquation entre l'absence de coupure entre « Quincy » et « could » dans le signal, alors qu'on devrait plutôt trouver une coupure entre ces deux mots. Dans cet exemple, le débit du locuteur peut expliquer cet effet de « liaison » perçu entre ces deux termes. La séquence des tons y est également notée. Le mot « Quincy » porte un accent H* et le pitch correspondant se situe dans la partie supérieure du registre du locuteur.

Le système ToBI pose un cadre général de description des énoncés qui doit être adapté pour l'utiliser avec une langue spécifique. De nombreuses versions existent et les langues suivantes sont prises en compte : anglais, allemand, japonais, coréen, grec, catalan, portugais, etc. Des travaux existent également sur l'étiquetage automatique d'énoncés selon le système ToBI, ainsi que sur la génération automatique de contours de F_0 à partir d'une description ToBI (Black et Hunt, 1996).

3.1.2 INTSINT (INternational Transcription System for INTonation)

Le système INTSINT décrit l'intonation à partir d'un ensemble limité de symboles conçus pour permettre une utilisation dans toutes les langues sans avoir à modifier le jeu de formes élémentaires des contours de F_0, (Hirst *et al.*, 1994; Louw et Barnard, 2004).

L'entrée du système INTSINT est une séquence de points cibles calculés grâce à l'algorithme de stylisation MoMel. Les symboles permettant de représenter les points cibles sont au nombre de huit et peuvent être répartis selon deux grandes catégories ci-dessous.

- Les tons absolus correspondent à la plage de variation du F_0 pour le locuteur :
  - T : Top
  - M : Mid
  - B : Bottom
- Les tons relatifs s'expriment de manière relative au point cible précédent :
  - H : Higher
  - S : Same
  - L : Lower

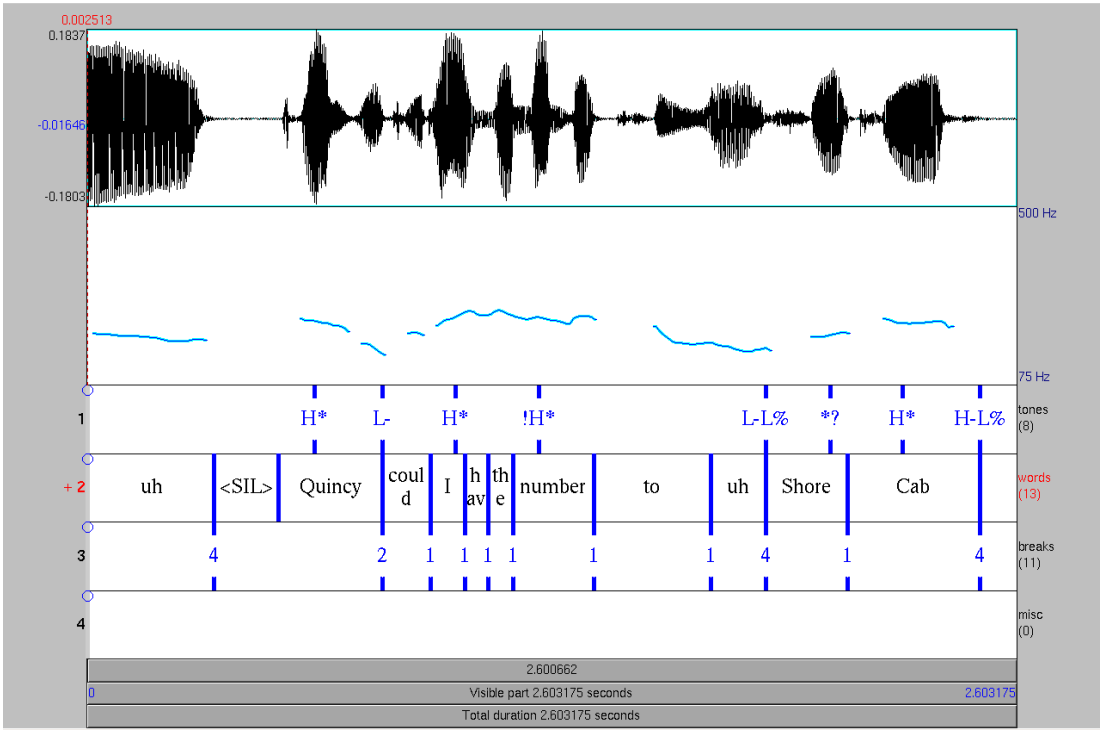


FIG. 3.1 – Exemple de transcription réalisée à l'aide de TOBI de la phrase « Uh Quincy could I have the number to uh Shore Cab ? ». De haut en bas, on trouve le signal, le F_0, la séquence des tons, la séquence des mots et la séquence des indices de coupure. On peut noter la présence de plusieurs frontières de groupes intonatifs délimités par le symbole 4. Dans cet exemple, l'indice de coupure 2 est utilisé pour mettre en avant que la disjonction entre « Quincy » et « could » n'est pas marquée alors qu'elle le devrait. En fait, ici, le locuteur parle de manière rapide ce qui peut transmettre à l'auditeur une impression d'urgence ou bien indiquer le désir de conserver la parole. La figure a été réalisée avec le logiciel Praat (Boersma et Weenink, 2008). (phrase extraite de (Beckman et Elam, 1997))

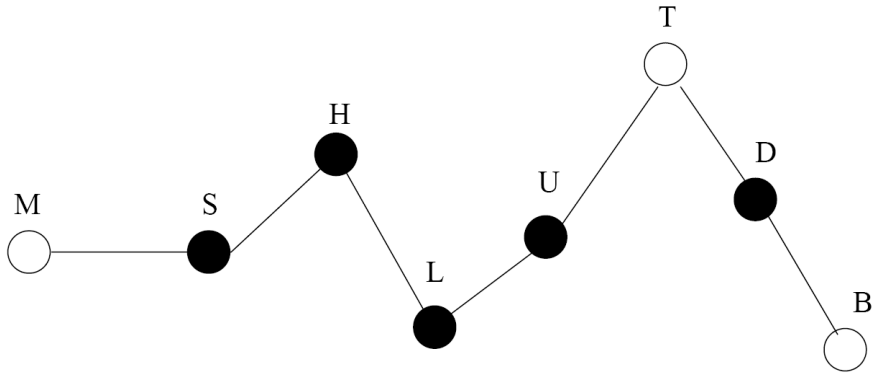


FIG. 3.2 – Exemple de séquence de symboles INTSINT. Les symboles T et B correspondent respectivement aux maximum et minimum pour la phrase. Le premier point cible de la phrase ou suivant une pause est toujours marqué M, sauf s'il est déjà marqué par T ou B (extrait de (Louw et Barnard, 2004)).

- U : Up-stepped
- D : Down-stepped

Les tons relatifs se décomposent à leur tour en deux ensembles qui correspondent aux tons non-itératifs, qui ne peuvent pas être répétés (H, S, L), et aux tons itératifs (U et D). Un ensemble de règles, présentées dans (Louw et Barnard, 2004), permet d'étiqueter les points cibles par des symboles INTSINT (figure 3.2).

Des expériences ont été menées sur 152 phrases en Zulu (Louw et Barnard, 2004). La transcription phonétique a été réalisée de manière automatique et corrigée manuellement. Les labels INTSINT sont alignés avec la voyelle centrale des syllabes. L'évaluation de l'étiquetage automatique de l'intonation est réalisée par le calcul d'une erreur RMS. Pour chaque phrase, la courbe mélodique est générée à partir des étiquettes INTSINT et est comparée à la courbe obtenue avec l'algorithme MoMel, présenté au paragraphe 3.1.6 page 62. Les résultats montrent une erreur RMS d'environ 9Hz en considérant toutes les syllabes, et d'environ 20Hz en considérant uniquement les syllabes accentuées.

3.1.3 PAINTE (Parametric INTonation Event)

Cette approche, décrite dans (Möhler et Conkie, 1998), représente les contours de F_0 par une fonction d'approximation qui est la somme de deux fonctions sigmoïdes, l'une croissante et l'autre décroissante avec un décalage temporel fixé (voir figure 3.3 page suivante). Chaque fonction sigmoïde peut être décrite par quatre paramètres modélisant son maximum, son amplitude, son alignement temporel et son inclinaison. L'alignement temporel ainsi que la valeur maximale peuvent être mis en commun pour les deux

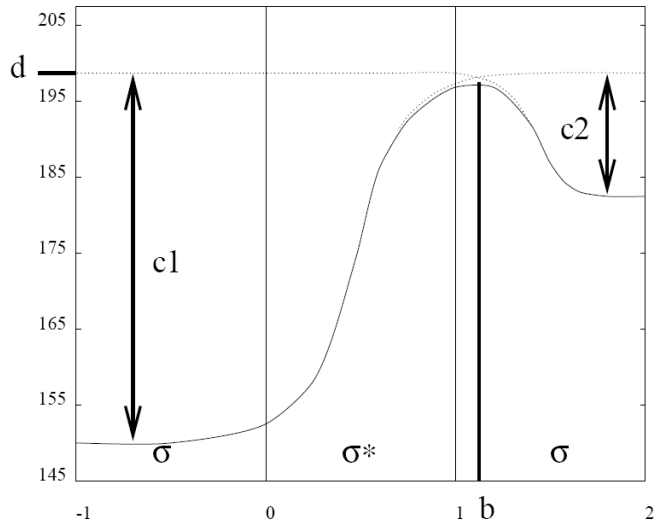


FIG. 3.3 – Fonction d'approximation d'un contour de F_0 avec le modèle PAINTE. Cette fonction est la somme d'une fonction sigmoïde croissante et d'une fonction sigmoïde décroissante décalée dans le temps. La longueur du support temporel des syllabes est normalisée (extrait de (Möhler et Conkie, 1998)).

fonctions sigmoïdes, réduisant ainsi le nombre de paramètres nécessaires à la description d'un contour à six paramètres :

- a1, a2 : inclinaisons respectives des sigmoïdes croissante et décroissante ;
- b : alignement temporel de la fonction qui représente l'instant où le maximum de la fonction est atteint ;
- c1, c2 : amplitude pour, respectivement, la fonction sigmoïde croissante et la fonction décroissante ;
- d : valeur maximale atteinte par le contour.

L'équation de la fonction du modèle s'écrit sous la forme suivante :

$$f(x) = d - \frac{c_1}{1 + \exp(-a_1(b - x) + \gamma)} - \frac{c_2}{1 + \exp(-a_2(x - b) + \gamma)} \tag{3.1}$$

où γ est un paramètre constant d'alignement des deux sigmoïdes de la fonction du modèle.

Dans (Möhler, 1999), Möhler montre qu'appliquer une quantification vectorielle sur les paramètres obtenus en stylisant un ensemble de syllabes permet d'obtenir un codebook composé de formes de contours représentatives de celles utilisées par le locuteur. Selon lui, cette approche permet également d'établir un lien entre les formes retenues et la phonologie.

3.1.4 Modèle RFC : Rise/Fall/Connection

Ce modèle, proposé par Taylor (Taylor, 1993, 1995), peut être situé entre le niveau de stylisation du F_0 et de celui d'une description phonologique. Ce modèle fait l'hypothèse qu'un contour de F_0 peut être représenté par une séquence d'éléments distincts qui peuvent être de trois types : montée (*rise*), descente (*fall*) et connexion (*connection*).

D'après les études portant sur les accents mélodiques de l'anglais, Taylor en distingue deux types : les accents de sommet (*peak accent*) et les accents de vallée (*trough accent*). La stylisation optimale du F_0 est estimée en respectant les quatre contraintes suivantes :

- un accent possède au maximum un élément de montée et un élément de descente ;
- les montées des frontières de syntagme sont modélisées grâce à un unique élément de montée ;
- les éléments non significatifs de l'intonation sont modélisés par des éléments de connexion ;
- entre un élément de montée et de descente, un seul élément de connexion peut être utilisé.

Un accent de sommet est modélisé en décrivant de manière séparée un élément de montée suivi d'un élément de descente. L'équation monomiale (3.2), obtenue empiriquement et proposée par Taylor, permet de décrire un élément de descente :

$$F_0 = \begin{cases} A_{\text{fall}} - 2.A_{\text{fall}}.(t/D_{\text{fall}})^2, & 0 < t < D_{\text{fall}}/2 \\ 2.A_{\text{fall}}.(1 - t/D_{\text{fall}})^2, & D_{\text{fall}}/2 < t < D_{\text{fall}} \end{cases} \tag{3.2}$$

Dans cette équation, A_{fall} représente l'amplitude de l'élément et D_{fall} la durée de l'élément de descente. Pour obtenir l'équation d'un élément de montée, il suffit de prendre la fonction symétrique par rapport à l'axe des ordonnées. La figure 3.4 représente dans sa partie droite, pour $t \geq 0$, la forme d'un élément de descente en appliquant l'équation (3.2) avec $A_{\text{rise}} = A_{\text{fall}} = 1$ et $D_{\text{rise}} = D_{\text{fall}} = 1$.

Les accents de vallées sont quant à eux représentés en utilisant la même fonction monomiale mais en positionnant en premier un élément descendant puis un élément montant.

L'algorithme de placement des éléments proposé par Taylor se présente en trois phases. La première consiste à placer grossièrement les éléments montants et descendants sur le signal de F_0 échantillonné toutes les 50ms. Un élément montant (resp. descendant) est détecté lorsque la pente du F_0 dépasse un seuil fixé. Les éléments adjacents sont regroupés sous la forme d'une section. À ce stade, les connexions ne sont pas encore placées entre les éléments. La deuxième phase est un processus de délétion des

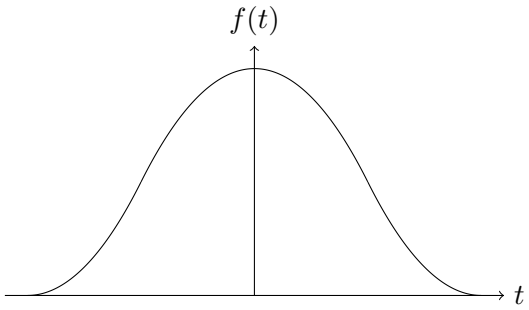


FIG. 3.4 – Prototype d'un accent de sommet construit par l'assemblage d'un élément de montée et d'un élément de descente avec $A_{\text{rise}} = A_{\text{fall}} = 1$ et $D_{\text{rise}} = D_{\text{fall}} = 1$ pour les deux éléments.

Set de données	**Nb. de phrases**	**Etiquetage manuel**	**Etiquetage automatique**
A	64	4.9Hz	4.7Hz
B	45	7.3Hz	5.4Hz
C	55	3.6Hz	4.2Hz
D	19	4.1Hz	4.3Hz
E	21	3.7Hz	3.8Hz
F	17	4.9Hz	3.9Hz

TAB. 3.1 – Erreur moyenne en Hertz entre les contours étiquetés manuellement et automatiquement (résultats extraits de (Taylor, 1995)).

sections trop courtes (par rapport à un seuil de longueur fixé) qui peuvent apparaître en raison de perturbations locales sur la courbe de F_0. La troisième et dernière phase de l'algorithme tente de positionner de manière optimale les frontières des éléments montants et descendants à partir du F_0 échantillonné toutes les 5ms. Pour cela, une fenêtre de recherche est définie autour de chaque élément déjà positionné. Dans chaque région, le contour optimal au sens de la distance euclidienne avec le F_0 est obtenu en évaluant toutes les possibilités de frontières.

L'évaluation réalisée dans (Taylor, 1995) porte sur six ensembles de données issues de locuteurs masculins et féminins pour l'anglais américain, l'anglais et l'irlandais. Les résultats du tableau 3.1 montrent une stylisation assez précise des contours de F_0 aussi bien dans le cas manuel que dans le cas automatique. À partir de ces résultats, Taylor conclut que la méthode de description RFC permet de styliser efficacement les contours de F_0. L'ancrage phonologique de ce modèle le rend utilisable pour la synthèse de contours de F_0 à partir d'une description linguistique.

3.1.5 Tilt

L'unité de base du modèle Tilt, proposé par Taylor (2000) est l'événement intonatif. Ces événements apparaissent de manière non consécutive. Deux types d'événements sont considérés :

- les accents (*pitch accents*), représentés par la lettre *a*. Ils correspondent à des mouvements de F_0 au niveau syllabique et permettent au locuteur de marquer un certain degré d'emphase au niveau d'un mot ou d'une syllabe.
- les frontières de ton (*boundary tones*), représentées par la lettre *b*. Elles apparaissent au niveau des frontières de groupes intonatifs et peuvent donner en outre un effet de continuation ou de questionnement.

Taylor décrit ce modèle comme une amélioration du modèle RFC pour lequel les paramètres de description d'un événement sont difficilement interprétables et manipulables. Par exemple, un événement RFC est décrit par deux jeux de paramètres (un pour le *rise* et un pour le *fall*) alors qu'il serait plus aisé de n'en utiliser qu'un seul. La représentation d'un événement Tilt s'appuie sur trois paramètres : la durée, l'amplitude et un paramètre sans dimension dénommé *tilt*. Le jeu complet de paramètres permettant de décrire aussi bien un accent qu'une frontière de tons est décrit ci-dessous.

- **Amplitude :** amplitude du mouvement de F_0 de l'événement ;
- **Durée :** en secondes du début à la fin de l'événement ;
- **Tilt :** paramètre sans dimension à valeurs dans $[-1, 1]$. Il décrit la forme de l'événement. Tilt est calculé relativement à la taille du rise et du fall. En particulier, la valeur $+1$ indique que l'événement est purement un rise, -1 pour un fall. La valeur 0 indique que le rise et le fall ont la même taille ;
- F_0 **position :** distance de F_0 entre la baseline et le milieu de l'événement ;
- **Time position :** emplacement temporel de l'événement. Il y a deux manières courantes de l'interpréter : soit comme la durée entre le début de la phrase et le milieu de l'événement, soit comme la position relative à la syllabe associée jusqu'au milieu de l'événement. Le point de référence dans la syllabe est généralement le début de la voyelle.

La première étape pour l'analyse automatique de l'intonation en utilisant Tilt est la détection des événements à partir du signal de parole (Taylor, 1998). Il s'agit en fait de repérer les zones du signal qui correspondent à ce que Taylor appelle des évènements intonationnels. Il utilise en particulier des HMM gauche-droite possédant trois états avec des densités continues pour détecter les accents, les frontières, les silences et les zones qui ne correspondent à aucun événement. La deuxième étape consiste à trouver les paramètres Tilt pour chaque événement en utilisant le lien qui existe entre les paramètres

RFC et les paramètres Tilt. Ainsi les paramètres RFC qui minimisent l'erreur entre la courbe stylisée et la courbe originale de F_0 de chaque événement sont calculés. Les paramètres Tilt sont ensuite obtenus à partir des paramètres RFC A_{rise}, D_{rise}, A_{fall} et D_{fall} de la façon suivante :

$$A_{\text{event}} = |A_{\text{rise}}| + |A_{\text{fall}}| \tag{3.3}$$

$$D_{\text{event}} = D_{\text{rise}} + D_{\text{fall}} \tag{3.4}$$

$$\text{tilt} = \frac{|A_{\text{rise}}| - |A_{\text{fall}}|}{2 * (|A_{\text{rise}}| + |A_{\text{fall}}|)} + \frac{D_{\text{rise}} - D_{\text{fall}}}{2 * (D_{\text{rise}} + D_{\text{fall}})} \tag{3.5}$$

Les paramètres de position du F_0 et l'amplitude de durée sont calculés directement. Pour une synthèse automatique du F_0, les paramètres Tilt sont convertis en paramètres RFC. Entre chaque événement, les valeurs de F_0 sont obtenues par interpolation linéaire.

Dans (Taylor, 1998), les résultats par rapport à l'erreur de stylisation de F_0 sur le corpus *DCIEM* donnent une erreur RMS de l'ordre de 7Hz pour le cas où le contour de F_0 est lissé. Dans cet exemple, un étiquetage automatique ou un étiquetage manuel des événements donnent des résultats similaires.

Enfin, les travaux de DusterHoff et al. (Dusterhoff et Black, 1997; Dusterhoff *et al.*, 1999) montrent que ce modèle peut être utilisé conjointement à un arbre de classification et de régression (CART) pour générer le F_0 à partir d'informations de haut niveau dans un système TTS.

3.1.6 Stylisation MoMel

L'algorithme MoMel (*MOdélisation MELodique*) s'intéresse à la modélisation de la composante macroprosodique du contour de F_0 qui reflète le choix d'un patron intonatif par le locuteur. Il représente le contour de F_0 par une courbe lisse et continue (Hirst et Espesser, 1993; Campione et Véronis, 2000), composée d'une fonction spline quadratique. Une telle fonction correspond à une série d'arcs de parabole reliés entre eux. Elle est continue et dérivable et peut être représentée en ne conservant que les points correspondant à des changements significatifs (passage par zéro de la tangente). Ces points significatifs sont dénommés points cibles.

Une description détaillée de l'algorithme est présentée dans (Campione et Véronis, 2000). Dans ce paragraphe, nous rappelons simplement les grandes lignes de l'algorithme MoMel, composé de quatre étapes :

1. Élimination des valeurs aberrantes du contour de F_0.
2. Estimation des points cibles. Un point cible est calculé pour chaque instant en

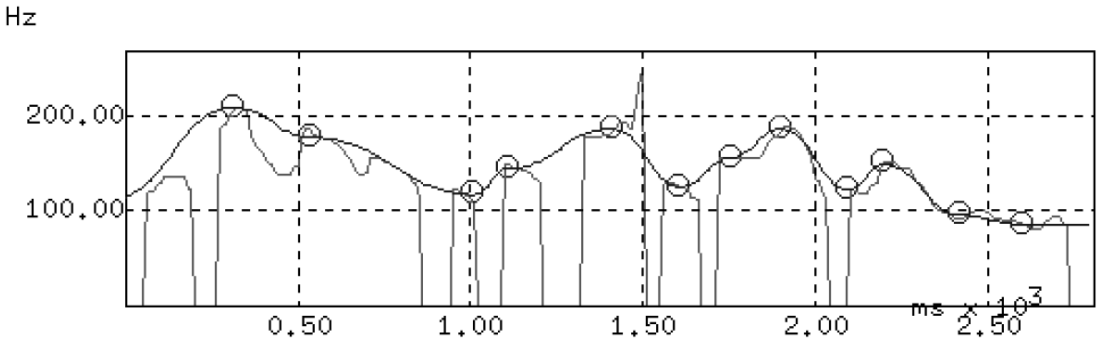


FIG. 3.5 – Courbe de F_0 et courbe spline quadratique obtenue avec MOMEL. Les points cibles sont représentés par des cercles (extrait de (Campione et Véronis, 2000)).

appliquant une régression modale sur une portion du contour centrée sur l'instant courant en appliquant une fenêtre d'analyse de longueur A.

3. Partition des points cibles candidats calculés. Elle s'effectue en appliquant une fenêtre de réduction glissante R, divisée en deux moitiés : gauche et droite. Une frontière de partition est insérée si la moyenne des point cibles candidats dans les deux moitiés diffère au-delà d'un certain seuil.
4. Estimation finale du point cible de chaque segment de la partition. Le point cible d'un segment est calculé comme étant la moyenne des points cibles candidats de ce segment après avoir éliminé les points cibles candidats trop éloignés de la moyenne.

À l'issue de cet algorithme, une séquence de points cibles décrivant le contour de F_0 est disponible. La figure 3.5 présente un exemple de stylisation d'un contour de F_0 ainsi que la courbe spline quadratique obtenue. On peut observer que le contour stylisé est un contour assez lisse et représente l'évolution générale du contour original.

Une méthode d'adaptation automatique à de nouveaux corpus de l'algorithme MOMEL est proposé dans (Mouline *et al.*, 2004). Dans cet article, Mouline *et al.* montrent que deux paramètres de l'algorithme liés à la taille des fenêtres sont fortement dépendants du corpus. Leur estimation automatique permet une réduction significative de l'erreur de stylisation.

3.1.7 Modèle de Fujisaki

3.1.7.1 Présentation

Le modèle de Fujisaki et Hirose (1984) est un modèle quantitatif du F_0 fondé sur les processus physiologique et physique de génération de la fréquence fondamentale. La

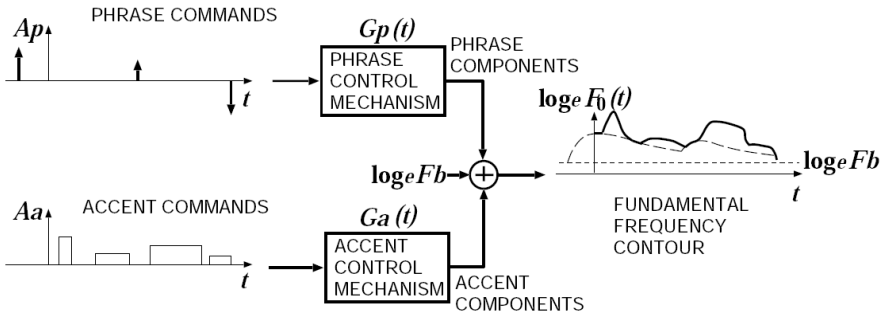


FIG. 3.6 – Diagramme fonctionnel du modèle de Fujisaki. Les composantes de groupe et d'accent s'ajoutent pour former le contour de F_0 (extrait de (Fujisaki, 2004)).

construction du modèle de Fujisaki (Fujisaki, 2004) repose sur l'existence de commandes de groupe et de commandes d'accent qui influent directement sur un modèle génératif. Les premières correspondent à un train d'impulsions. La composante de groupe est la réponse impulsionnelle d'un système linéaire du second ordre amorti. Les secondes correspondent à des fonctions « porte » où la composante d'accent est la réponse d'un système linéaire du second ordre à ces commandes. Pour finir, les composantes de groupe et d'accent sont additionnées pour former le contour de F_0. Le diagramme fonctionnel du modèle est représenté sur la figure 3.6.

Le modèle génère un contour de F_0 dans le domaine logarithmique avec les équations suivantes :

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{p_i} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{a_j} \left(G_a(t - T_{1j}) - G_a(t - T_{2j})\right) \quad (3.6)$$

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & \text{pour} t \geq 0 \\ 0, & \text{pour} t < 0 \end{cases} \quad (3.7)$$

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t)\exp(-\beta t), \gamma], & \text{pour} t \geq 0 \\ 0, & \text{pour} t < 0 \end{cases} \quad (3.8)$$

Dans ces équations, F_b indique la baseline du contour de F_0.

Commandes de groupe G_p dénote la réponse impulsionnelle du système linéaire des commandes de groupe. Les I commandes de groupe sont définies par leur magnitude A_{p_i} et leur instant de début T_{0i}. α représente la constante de temps du système linéaire pour les commandes de groupe et est supposé constant au moins au niveau d'une phrase. La forme d'une commande de groupe est illustrée pour plusieurs valeurs de magnitude

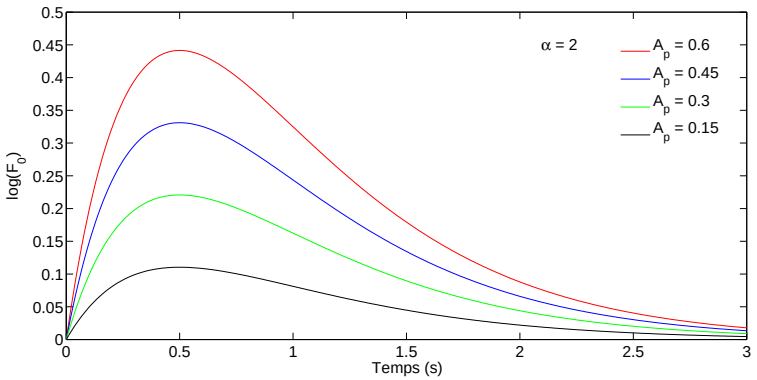


FIG. 3.7 – Forme des commandes de groupe pour différentes valeurs d'amplitude $A_p = 0.6, 0.45, 0.3$ et 0.15 avec $\alpha = 2$ (extrait de (Mixdorff, 1998)).

sur la figure 3.7.

Commandes d'accent G_a dénote la réponse du système linéaire des commandes d'accent. Une commande d'accent correspond à une fonction porte et est définie par son amplitude A_{a_j}, son instant de début T_{1j} et son instant de fin T_{2j}. La forme d'une commande d'accent est illustrée pour plusieurs valeurs d'amplitude et de durée sur les figures 3.8(a) et 3.8(b) page suivante.

3.1.7.2 Interprétation physiologique

La construction du modèle de Fujisaki repose sur les mécanismes physiologiques et physiques de fonctionnement du larynx et en particulier sur l'activité du muscle cricothyroïdien. Fujisaki dérive son modèle de la relation qui existe entre la tension T et l'élongation x d'un muscle :

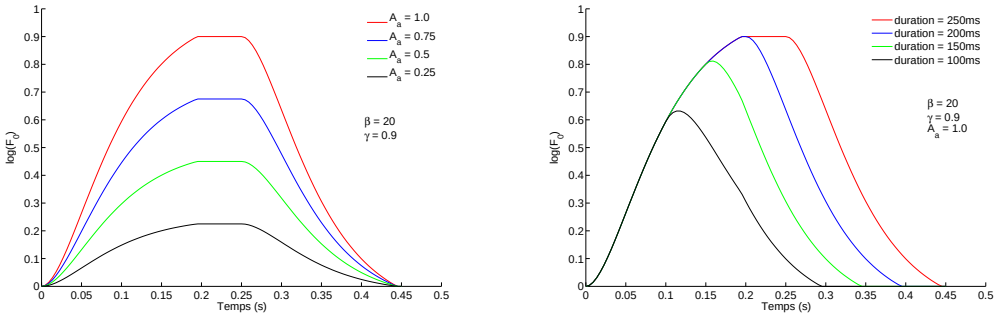
$$T = a(\exp(bx) - 1) \approx a\exp(bx) \gg 1. \tag{3.9}$$

La fréquence fondamentale F_0 de vibration d'une membrane élastique en fonction de la tension T est donnée par :

$$F_0 = c_0\sqrt{T/\sigma}, \tag{3.10}$$

où σ est la densité par unité de surface de la membrane et c_0 est une constante inversement proportionnelle à la taille de la membrane. En combinant les équations (3.9) et (3.10), on obtient :

$$log_e F_0 = log_e\{c_0\sqrt{T_0/\sigma}\} + (b/2)x. \tag{3.11}$$



(a) (b)

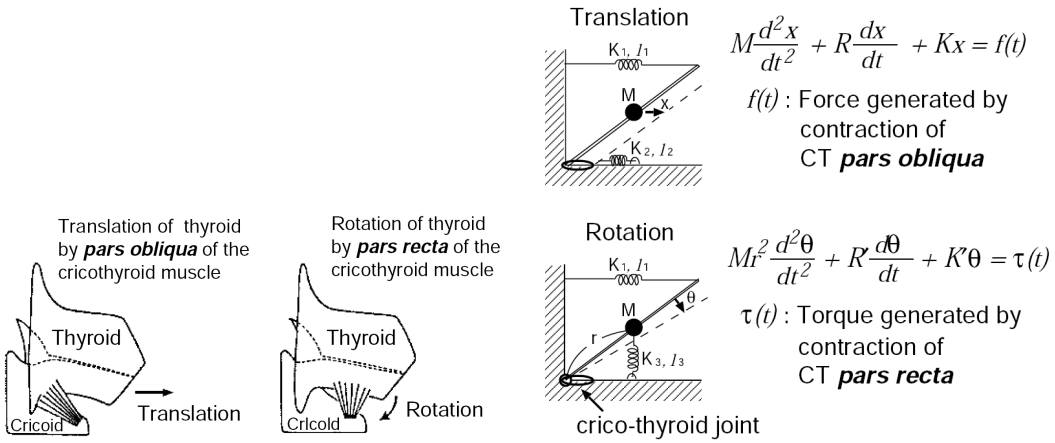
FIG. 3.8 – (a) Forme des commandes d'accent pour différentes valeurs d'amplitude $A_a = 1.0, 0.75, 0.5$ et 0.25 avec $\beta = 20$ et une durée de 250ms. (b) Forme des commandes d'accent en faisant varier la durée $250\text{ms}, 200\text{ms}, 150\text{ms},$ et 100ms avec $\beta = 20$ et $A_a = 1.0$ (extrait de (Mixdorff, 1998)).

Le terme constant $c_0\sqrt{T_0/\sigma}$ de l'équation (3.11) est réécrit F_b et indique l'existence d'une valeur de F_0 minimale lors de la phonation. Ainsi, l'évolution de la fréquence fondamentale $F_0(t)$ en fonction du temps est proportionnelle à l'élongation $x(t)$ plus une constante.

L'analyse de la structure du cartilage thyroïdien montre que le muscle cricothyroïdien possède deux degrés de liberté (figure 3.9(a) page suivante) : la translation horizontale du cartilage due à la partie oblique (*pars obliqua*) du muscle cricothyroïdien et la rotation due à la partie verticale de ce même muscle (*pars recta*). La translation et la rotation de la thyroïde peuvent être représentées par deux systèmes du second ordre (figure 3.9(b) page ci-contre) et provoquent une variation de la longueur des cordes vocales $x(t)$. L'activité de la partie oblique contribue à la translation de la thyroïde et provoque un changement de l'élongation $x_1(t)$. L'activité de la partie verticale provoque, quant à elle, un changement de longueur $x_2(t)$. L'équation (3.11) peut alors être réécrite :

$$log_e F_0(t) = log_e\{c_0\sqrt{T_0/\sigma}\} + (b/2)(x_1(t) + x_2(t)). \tag{3.12}$$

La composante de F_0 dépendant du temps s'exprime alors comme la somme de deux composantes elles-mêmes dépendantes du temps. De plus, le mouvement de translation possède une constante de temps bien plus grande que celui de rotation. Le premier va alors correspondre à un phénomène à long terme (groupe intonatif) tandis que le second aura une influence à court terme (accent).



(a) (b)

FIG. 3.9 – (a) Rôle de *pars obliqua* et *pars recta* du muscle cricothyroïdien dans la translation et la rotation du cartilage thyroïdien. (b) Équations de rotation et translation des mouvements du cartilage thyroïdien (extrait de (Fujisaki, 2004)).

3.1.7.3 Estimation des paramètres

L'inférence des paramètres du modèle de Fujisaki à partir du contour de F_0 est difficile puisqu'il n'existe pas de solution analytique au problème de l'inversion du modèle de Fujisaki. Ainsi l'extraction des paramètres du modèle de Fujisaki est réalisée grâce à un processus d'analyse-synthèse. De nombreux travaux traitent de ce sujet.

Mixdorff (2000) propose une méthode qui repose sur la décomposition du contour de F_0, stylisé avec une spline quadratique, en deux composantes. La séparation de ces deux composantes est réalisée à l'aide d'un filtre passe-haut. La première contient les fréquences les plus hautes, elle varie donc rapidement et correspond à la composante accentuelle du modèle de Fujisaki. La seconde contient les fréquences les plus basses et varie lentement : elle correspond à la commande de groupe. Les positions des commandes d'accent et de groupe sont repérées dans les deux composantes par la recherche des minima locaux. Les différents paramètres sont ensuite optimisés pour minimiser l'erreur quadratique moyenne à l'aide d'un algorithme de type *Hill Climbing*.

Narusawa *et al.* (2002) suggèrent d'extraire les paramètres en appliquant une procédure de pré-traitement qui stylise le contour de F_0 avec un polynôme continu de degré 3. La séquence des maxima et minima de la dérivée du polynôme donne les instants de début et fin des commandes d'accent. Les commandes de groupe sont ensuite estimées à partir du contour de F_0 duquel les commandes d'accent ont été soustraites.

D'autres méthodes existent également : van Santen *et al.* (2004) proposent d'utiliser

une décomposition en ondelettes pour repérer les commandes d'accents et de groupes, Sakurai *et al.* (2003) considèrent des réseaux de neurones ainsi que des arbres de régression, et Silva et Netto (2004) utilisent une approximation analytique des amplitudes des commandes d'accent et de groupe. Pour Agüero *et al.* (2004), il est nécessaire de tenir compte des contraintes liées aux groupes accentuels et intonatifs pour estimer les commandes d'accent et de groupe, à l'inverse des autres méthodes qui ne considèrent que le contour de F_0.

3.1.8 Systèmes dynamiques

La fréquence fondamentale peut être considérée comme étant une variable observée issue d'un système dynamique. À partir de cette hypothèse, Ross et Ostendorf (1999) proposent une modélisation conjointe de la F_0 et de l'énergie.

Un système dynamique représente une séquence d'observations $\mathbf{Y} = [\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_N]$ par l'intermédiaire d'une séquence d'états cachés $\mathbf{X} = [\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_N]$. La forme utilisée par Ross et Ostendorf pour le modèle à espace d'états discret est la suivante :

$$\mathbf{x}_{k+1} = F_j \mathbf{x}_k + u_j + w_k \tag{3.13}$$

$$\mathbf{y}_k = H_j \mathbf{x}_k + b_j + v_k \tag{3.14}$$

Les équations 3.13 et 3.14 sont appelées respectivement équation d'état et équation d'observation. $\{\mathbf{x}_k\}$ est un processus markovien et correspond au processus caché. $\{\mathbf{y}_k\}$ est un processus aléatoire à valeurs continues. Pour tout k, l'observation de $\mathbf{y}_k$ est conditionnée par $\mathbf{x}_k$ donc dépendante de $\{\mathbf{x}_k\}$. Les variables u_j et b_j représentent des entrées déterministes respectivement pour l'équation d'état (3.13) et l'équation d'observation (3.14). Les termes w_k et v_k sont des bruits gaussiens centrés et ne sont pas corrélés entre eux. Leur matrice de covariance est respectivement Q_j et R_j. L'état initial de ce système est $\mathbf{x}_0$ de loi gaussienne (de moyenne μ_{x_0} et de covariance Σ_{x_0}) indépendante du bruit de l'équation d'état et de l'équation d'observation. $\mathbf{y}_k$ est un vecteur de dimension 2 qui contient les valeurs de F_0 et d'énergie. De même, l'état $\mathbf{x}_k$ est un vecteur de dimension 2. L'indice j permet de sélectionner un jeu de paramètres approprié pour une région particulière du contour de F_0.

Dans leur article, Ross et Ostendorf indiquent que le jeu de paramètres adéquat est sélectionné en fonction de dépendances au niveau de la syllabe, du groupe intonatif et du segment. Les paramètres utilisés pour cette sélection sont la séquence d'étiquettes prosodiques, la structure de la syllabe, l'accent lexical et la durée du segment.

L'estimation des paramètres est réalisée par un algorithme de type EM, Expectation-Maximisation, (Dempster *et al.*, 1977). L'utilisation de l'algorithme EM permet

1. d'estimer une statistique sur le processus caché $\{\mathbf{x}_k\}$;
2. de considérer les zones non voisées comme des données manquantes.

Dans le cadre de la stylisation du F_0, ce second point suit l'hypothèse selon laquelle il existerait un geste mélodique continu pendant les zones non voisées. L'algorithme utilisé est dérivé de celui proposé par Digalakis *et al.* (1993) pour prendre en compte plusieurs jeux de paramètres lors de l'estimation (un jeu de paramètres j correspond à une région).

Dans (Ross et Ostendorf, 1999), les expériences sont menées sur un ensemble de phrases issues du locuteur *F2B* du *Boston Radio News Corpus*. En ce qui concerne l'erreur RMS de modélisation des contours de F_0, les résultats présentés sont assez difficilement comparables à ceux des autres modèles puisqu'ils sont calculés par rapport aux contours de F_0 normalisés. Mais ce modèle génératif est utilisable dans le cadre d'un système de synthèse de parole et les résultats de l'erreur RMS de prédiction des contours d'environ 30Hz sont comparables aux autres modèles de prédiction du F_0 de l'époque (Dusterhoff et Black, 1997; Dusterhoff *et al.*, 1999; Möhler et Conkie, 1998).

Une évaluation perceptive, qui compare ce modèle au système de synthèse de AT&T de 1994 utilisant le modèle source/filtre, a également été réalisée. Ce test, ayant pour objectif d'évaluer le naturel des contours de F_0 générés, montre une meilleure performance du système dynamique ainsi qu'une amélioration significative des résultats par rapport au système de AT&T.

3.2 Classification de contours de F_0

L'intérêt de la classification est de faire émerger un ensemble restreint de classes de contours de F_0 représentatives des contours produits par le locuteur, et de réduire le nombre de contours en choisissant des représentants significatifs. La classification de contours mélodiques s'inscrit dans le cadre de la modélisation et succède en général à une étape de stylisation.

Assez peu de travaux portent sur la classification de contours mélodiques. Dans la plupart des cas, la classification est utilisée pour faciliter la prédiction de la prosodie dans les systèmes de synthèse de parole. Un partitionnement des contours de F_0 est alors souvent associé à un arbre de classification ou de régression (Breiman *et al.*, 1984) permettant de prédire la classe du contour de F_0 à partir d'informations phonologiques, syntaxiques, etc.

La première étape, avant d'être en mesure d'effectuer une classification, est de rendre comparables les contours de F_0. L'échelle de temps de ces contours peut être le phonème, la syllabe, le groupe de souffle ou encore la phrase. La principale difficulté tient au fait que des contours de longueurs différentes peuvent représenter le même motif et ainsi appartenir à une même classe. Deux approches sont alors possibles :

- normaliser le support temporel des contours de F_0 ;
- utiliser un modèle capable de capter la variation de durée.

La première s'appuie d'abord sur une méthode arbitraire de normalisation suivie par une stylisation des contours mélodiques pour faciliter l'application d'une distance entre deux contours de F_0. La deuxième approche permet d'intégrer directement dans le modèle l'élasticité du support temporel en utilisant par exemple des chaînes de Markov à états cachés (HMM).

3.2.1 Normalisation et stylisation

Pour éviter le problème de longueur variable des contours de F_0, la première idée qui vient à l'esprit est de projeter tous les contours sur le même support temporel. Il s'agit donc d'une homothétie de l'axe du temps qui n'est pas identique pour tous les contours. Un ré-échantillonnage des contours est alors nécessaire pour obtenir le même nombre de points pour chaque contour.

Cette idée est mise en œuvre par Reichel (2007) pour éliminer l'influence du débit et de la structure de la syllabe. La longueur d'une syllabe est ainsi normalisée en projetant chaque constituant (onset, noyau et coda) sur un intervalle de longueur fixe. Les onsets ou codas manquants sont interpolés par rapport aux valeurs de F_0 du noyau et des syllabes voisines.

Une fois la normalisation du temps effectuée, il est nécessaire de choisir une représentation unique du contour de F_0. Ainsi, ré-échantillonner les contours de F_0 permet d'obtenir des vecteurs de taille fixe pour tous les contours. Une autre stratégie est d'appliquer un modèle pour styliser les contours de F_0. Chaque contour est alors représentable par un jeu de paramètres de taille fixe.

3.2.2 Classification

Même si en pratique de nombreuses méthodes issues de l'apprentissage automatique peuvent être utilisées pour obtenir un partitionnement des contours mélodiques, peu de travaux existent à ce sujet. On peut tout de même citer les travaux de Reichel (2007), Möhler et Conkie (1998) et de Yamashita *et al.* (2003), qui utilisent des méthodes de

type k-moyennes ou quantification vectorielle, ou encore les travaux de Tokuda *et al.* (1999) qui utilisent des MSD-HMM (Multi-Space probability Distribution HMM).

K-moyennes L'algorithme des k-moyennes est un algorithme de classification non supervisée ayant pour objectif de trouver k vecteurs $x_1, \ldots, x_k$ permettant de représenter au mieux un ensemble de n vecteurs $y_1, \ldots, y_n$. Un vecteur représentant une classe est choisi comme étant le représentant de cette classe au sens d'une distance. Cet algorithme fait donc l'hypothèse que le nombre k de classes est connu à l'avance. Une notion de distance dans l'espace des vecteurs à partitionner est également nécessaire. Le choix des centroïdes de classes initiaux est un point crucial qui influe de manière importante sur le résultat. En effet, comme cet algorithme ne garantit pas une solution optimale, la qualité de la solution dépend fortement de l'initialisation.

L'approche proposée par Reichel (2007) met en œuvre l'algorithme des k-moyennes pour partitionner un ensemble de contours mélodiques stylisés par une fonction polynômiale dont le support temporel est normalisé. Le nombre de clusters est sélectionné en utilisant l'index de Dunn qui mesure de validité d'un partitionnement prenant en compte la distance intra-cluster et inter-cluster. L'erreur RMS rapportée entre les contours originaux et régénérés est de 10.26Hz.

Quantification vectorielle La Quantification Vectorielle (QV), très largement utilisée en traitement d'images et traitement de la parole, est une méthode de compression de données. Cette technique, dont le sens général est l'approximation d'un espace continu par un espace discret, permet de représenter tout vecteur x de dimension k par un vecteur y de même dimension appartenant à un ensemble fini D, appelé dictionnaire (codebook). L'algorithme LBG, du nom de ses inventeurs (Linde *et al.*, 1980), est un algorithme de quantification vectorielle couramment utilisé, similaire à celui des k-moyennes.

La quantification vectorielle est utilisée dans le modèle INTSINT, proposé par Möhler et Conkie (1998), ainsi que dans les travaux de Yamashita *et al.* (2003), pour retenir un nombre restreint de formes de contours de F_0. Dans les travaux de ces derniers, un contour mélodique est simplement représenté par quatre segments de droite. L'erreur RMS entre les contours originaux et régénérés rapportée par Möhler et Conkie est de 25.1Hz pour un nombre de classes égal à 32 sur un corpus contenant des extraits du *Wall Street Journal*.

MSD-HMM Avec cette approche, la distribution des observations associée à chaque état d'un HMM correspond à un mélange de lois définies sur des espaces possiblement de dimensions différentes (Tokuda *et al.*, 1999). Dans le cas du F_0, on peut distinguer les zones voisées et les zones non voisées. Dans le cas des zones voisées, les valeurs de F_0 peuvent être modélisées par une variable gaussienne ; dans le cas contraire on peut représenter le fait que la zone est non voisée par un symbole discret. L'estimation des paramètres des MSD-HMM est réalisée grâce à un algorithme de type EM. L'utilisation des HMM présente l'avantage de prendre directement en compte le temps dans le modèle. Il n'est alors plus nécessaire de normaliser le support temporel des contours mélodiques. Masuko *et al.* (2002) présentent quelques exemples de génération de contours mélodiques suivant la taille du modèle. Ceux-ci montrent une dégradation des contours générés lorsque la taille du modèle diminue. De plus, pour un ensemble de 53 phrases hors de l'ensemble d'apprentissage, l'erreur RMS rapportée est de 0.62 octave. En dépit d'une erreur RMS assez importante, Masuko *et al.* avancent que le sentiment de manque de naturel n'est pas très prononcé.

3.3 Conclusion

Dans ce chapitre, nous avons détaillé les principales approches proposant une solution aux problèmes de stylisation et de classification de la fréquence fondamentale. Notamment, les différents systèmes de stylisation se distinguent par leur ancrage phonologique, linguistique ou encore physiologique comme celui de Fujisaki.

Dans la plupart des cas, les systèmes de stylisation de nature symbolique offrent une stylisation linguistique de la fréquence fondamentale. Au contraire, les modèles mathématiques permettent une précision arbitraire de la stylisation au détriment des capacités d'explication des phénomènes linguistiques. Pour la construction d'un système de synthèse de parole à partir du texte, il est bien entendu nécessaire d'établir un lien entre le texte et le modèle prosodique. Un compromis doit donc être réalisé entre une modélisation assez fine et une interprétation plausible.

Dans ce même cadre, il est important de disposer d'un modèle prosodique ayant une précision importante pour obtenir des contours de F_0 proches du naturel et ainsi améliorer le confort de l'auditeur. Cependant, il est également nécessaire de pouvoir prédire quels contours utiliser à partir d'une description linguistique, ce qui est une tâche délicate.

Chapitre 4

Transformation de la prosodie : un état de l'art

Un mécanisme de transformation de la prosodie consiste à modifier la prosodie d'une phrase d'un locuteur source pour qu'elle soit perçue comme si elle avait été réalisée par un locuteur cible. Cette définition générale est la transposition de celle de la transformation des caractéristiques acoustiques segmentales d'un locuteur au niveau supra-segmental.

La plupart des méthodes de transformation de la prosodie s'appuient sur des modèles de stylisation ou de classification de la prosodie que nous venons de décrire au chapitre 3 page 53. En effet, utiliser directement les contours de F_0 réalisés par le locuteur se révèle être difficile (contours de longueurs différentes, micro-mélodie, etc.) et suivant les hypothèses établies, il peut être nécessaire d'utiliser un modèle de stylisation pour représenter les contours. Notamment, une telle méthodologie a pour objectif de supprimer les différences de longueur entre les contours de F_0 dans le cas de la stylisation, ou de réduire l'espace mélodique des locuteurs à un ensemble de formes représentatives dans le cas de la classification.

Deux corpus sont considérés comme parallèles lorsqu'ils contiennent les mêmes phrases. Cela signifie, dans le cas de la transformation de prosodie, que les deux locuteurs ont prononcé les mêmes phrases. Cette question de parallélisme entre les corpus est importante dans la mesure où elle contraint fortement les applications potentielles des techniques de transformation. Pour chaque méthode que nous allons présenter, nous préciserons si elle nécessite des données parallèles ou non.

La première partie de ce chapitre est consacrée à la présentation des différentes méthodes de transformation du F_0 existantes. Dans la seconde partie, nous discuterons

de l'évaluation des méthodes de transformation qui s'avère souvent délicate.

4.1 Méthodes de transformation

4.1.1 Gaussian Normalization

La normalisation à l'aide d'une loi normale ou « méthode moyenne/variance » est une méthode de transformation globale au locuteur (Chappell et Hansen, 1998). Il n'est pas nécessaire d'effectuer une segmentation du signal de parole pour être en mesure de l'appliquer. Il faut simplement récupérer les valeurs de F_0 pour les locuteurs source et cible. Cette méthode de transformation fait l'hypothèse que les valeurs de F_0 sont distribuées selon une loi normale. Il s'agit ici d'ajuster les valeurs du F_0 source, F_0^s, pour les centrer sur la moyenne du F_0 du locuteur cible, F_0^c. De même, la variance du F_0 source est modifiée pour qu'elle corresponde à celle du F_0 cible.

La phase d'apprentissage de la fonction de transformation est très simple puisqu'il suffit de calculer la moyenne et la variance des valeurs de F_0 pour les locuteurs source et cible. Quatre paramètres sont donc suffisants pour pouvoir appliquer cette méthode.

En ce qui concerne la transformation, l'équation (4.1) décrit la fonction qui doit être évaluée pour chaque valeur F_0^s de la source pour obtenir la valeur transformée F_0^t :

$$F_0^t = \frac{F_0^s - \mu_s}{\sigma_s} * \sigma_c + \mu_c \tag{4.1}$$

où μ_s et σ_s représentent la moyenne et l'écart-type du F_0 pour le locuteur source, μ_c et σ_c représentent la moyenne et l'écart-type du F_0 pour le locuteur cible.

Le principal avantage de cette méthode est qu'il n'est pas nécessaire de disposer de corpus parallèles pour les locuteurs source et cible. Cela signifie que l'apprentissage de cette fonction peut être réalisé sans que les locuteurs aient prononcé les mêmes phrases. Néanmoins, cette méthode très simple ne permet pas de modifier la forme des contours de F_0. Ainsi, les différences de prononciation entre source et cible ne seront pas prises en compte. Il en résulte que la mélodie transformée est plus proche de la mélodie source que de la mélodie cible comme le montre la figure 4.1 page ci-contre.

4.1.2 Scatterplot ou Nth Order Conversion Function

Cette méthode, proposée par Chappell et Hansen, est une généralisation de la précédente. Ici, on ne considère plus que les valeurs de F_0 sont distribuées selon une loi normale. Dans le cas présent, on effectue une régression polynômiale sur les couples

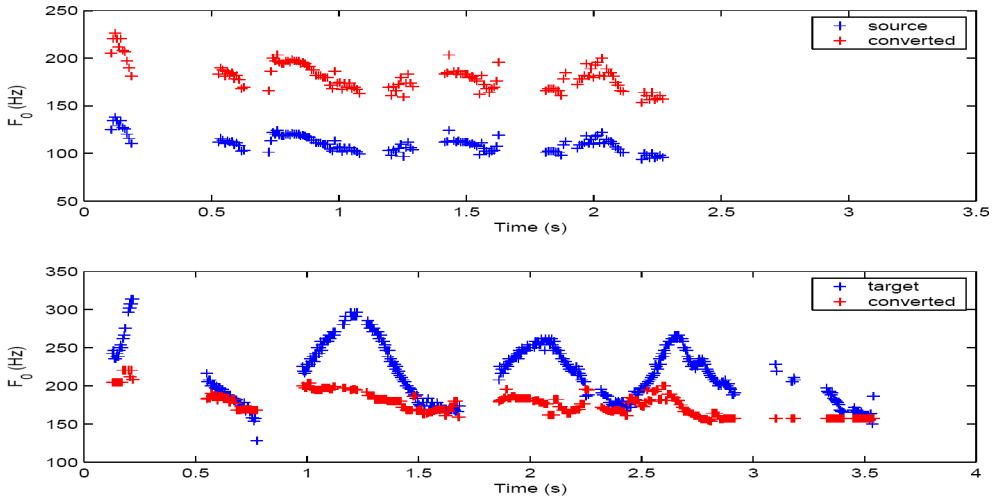


FIG. 4.1 – Exemple de transformation du F_0 pour la phrase « Trish saw hours and hours of movies on Saturday ». On peut noter la faible performance de la transformation puisqu'elle ne modifie pas la « forme » du contour de F_0. (extrait de (Inanoglu, 2003))

(F_0^s, F_0^c). La méthode de transformation proposée par Chappell et Hansen est effectuée au niveau des phones et nécessite des corpus parallèles. Un alignement des phones entre les phrases sources et cibles est effectué. Seuls les phones présents de manière commune pour la source et la cible sont pris en compte pour la transformation. Un « scatterplot[1] » est alors construit en prenant les couples de valeurs moyennes de F_0 au niveau des phones pour les deux locuteurs.

Une régression polynomiale est ensuite effectuée pour estimer le polynôme de degré n modélisant au mieux la relation entre F_0 source et cible au sens des moindres carrés. Ce polynôme est utilisé pour réaliser la transformation du F_0 d'une phrase du locuteur source. La figure 4.2 page suivante présente un exemple de régression polynômiale pour deux locuteurs.

Cette approche permet de modéliser une relation entre F_0 source et cible plus complexe que la précédente approche par normalisation gaussienne. Cette dernière peut être vue comme une régression par un polynôme de degré 1.

Concernant les résultats de l'évaluation perceptive qu'ils ont menée, Chappell et Hansen mettent en avant le fait que la prosodie de la phrase transformée par scatterplot est perceptivement différente de celle transformée par normalisation gaussienne. Cependant, la phrase modifiée est notablement plus proche de la phrase originale que de la phrase cible. Les résultats rapportés par Inanoglu (2003) montrent une erreur moyenne

[1] dans son sens strict, un *scatter plot* ou *scatter diagram* est un diagramme de dispersion.

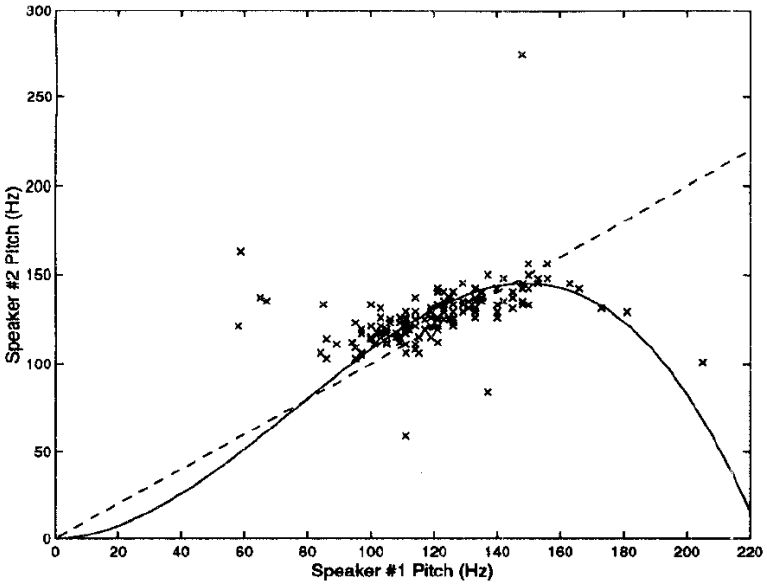


FIG. 4.2 – Exemple de scatterplot pour deux locuteurs. La courbe en pointillé représente la droite $x = y$ et sert de référence. La courbe en trait plein représente la fonction polynômiale de degré 3 estimée grâce aux couples de valeurs de F_0 source et cible. (extrait de (Chappell et Hansen, 1998))

de conversion des contours située entre 21.40Hz pour le meilleur cas et 39.77Hz pour le cas le plus défavorable. Dans cette expérience, 30 phrases sont utilisées pour l'apprentissage et 28 pour le test. L'erreur la plus forte apparaît lors de la conversion du F_0 depuis un locuteur de sexe masculin vers un locuteur de sexe féminin.

4.1.3 Mean-variance (Ceyssens)

La méthode proposée par Ceyssens *et al.* (2002) est fondée sur la méthode de normalisation gaussienne et effectue une paramétrisation des contours de F_0 au niveau de la phrase. Le principe est de modéliser, pour chaque contour, les paramètres moyenne et pente de manière déterministe, et les autres paramètres de manière probabiliste. Il s'agit ici de calculer pour chaque courbe un jeu de paramètres composé de l'offset P_o, de la pente P_s et de l'écart-type P_v du résidu de la soustraction entre le contour de F_0 et la droite de régression de ce même contour. À ces trois paramètres, il en ajoute un quatrième qui est la longueur du contour.

Les relations entre chacun des paramètres et la longueur des contours sont ensuite modélisées par une régression linéaire. Un nouveau jeu de trois paramètres est alors calculé pour modéliser chacune de ces trois relations. À partir des quatre paramètres

initiaux, on obtient donc un jeu de neuf paramètres pour chaque locuteur de la façon suivante.

- À partir des couples (offset, longueur) des contours de F_0, on calcule :
  - offset : P_{o_o}
  - pente (slope) : P_{o_s}
  - écart-type du résidu : P_{o_v}
- À partir des couples (pente, longueur), on calcule :
  - offset : P_{s_o}
  - pente : P_{s_s}
  - écart-type du résidu : P_{s_v}
- À partir des couples (écart-type, longueur), on calcule :
  - offset : P_{v_o}
  - pente : P_{v_s}
  - écart-type du résidu : P_{v_v}

On obtient alors pour chaque locuteur le jeu de paramètres P_{o_o}, P_{o_s}, P_{o_v}, P_{s_o}, P_{s_s}, P_{s_v}, P_{v_o}, P_{v_s}, P_{v_v}.

Lorsque l'on souhaite transformer un contour de F_0, celui-ci est modifié en lui donnant un nouvel offset, une nouvelle pente et une nouvelle variance autour de la droite de régression. Pour le contour source à transformer, il est donc nécessaire de calculer sa paramétrisation (P^s_o, P^s_s, P^s_v). Le calcul des valeurs transformées (P^t_o, P^t_s, P^t_v) est effectué en appliquant les équations (4.2), (4.3) et (4.4).

$$P^t_o = (P^c_{o_o} + L^s * P^c_{o_s}) + P^c_{o_v} \frac{P_o - (P^s_{o_o} + L * P^s_{o_s})}{P^s_{o_v}} \tag{4.2}$$

$$P^t_s = (P^c_{s_o} + L^s * P^c_{s_s}) + P^c_{s_v} \frac{P_s - (P^s_{s_o} + L * P^s_{s_s})}{P^s_{s_v}} \tag{4.3}$$

$$P^t_v = (P^c_{v_o} + L^s * P^c_{v_s}) + P^c_{v_v} \frac{P_v - (P^s_{v_o} + L * P^s_{v_s})}{P^s_{v_v}} \tag{4.4}$$

où L^s est la longueur du contour de F_0 à transformer. On note les paramètres sources et cibles avec en exposant les lettres s et c respectivement.

Cette méthode de transformation pourrait être étendue en modélisant plus finement le résidu autour de la droite de régression. En quelque sorte, une certaine récursivité au niveau du traitement du résidu pourrait être appliquée tant qu'il présente une importance perceptuelle. C'est ainsi que Ceyssens *et al.* définissent une transformation « parfaite » du F_0. Cependant, Ceyssens *et al.* ne donnent pas de résultats expérimen-

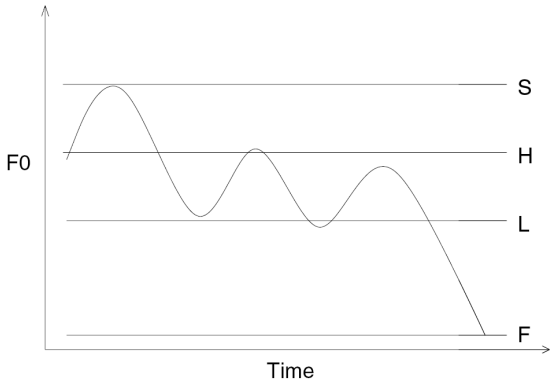


FIG. 4.3 – Emplacement des points de mesure sur un contour de F_0. (extrait de (Gillett et King, 2003))

taux montrant les performances de leur méthode.

4.1.4 Méthode de Gillett et King

L'approche de Gillett et King (2003) repose sur un jeu de paramètres limité et fondé linguistiquement. Le jeu de paramètres est issu des travaux de Patterson et est composé de quatre points cibles (mesures de valeur de F_0) sélectionnés dans chaque phrase :

- S : valeur de F_0 du pic initial de phrase (*sentence-initial high*),
- H : valeur maximale de pic d'accent non initial (*non-initial accent peaks*),
- L : valeur minimale de la vallée suivant un accent (*post-accent valleys*),
- F : valeur du F_0 minimum en fin de phrase (*sentence-final low*).

La figure 4.3 illustre les différents points cible sur un contour de F_0 fictif. Les points clés sélectionnés dans le contour de F_0 permettent de représenter de manière schématique la forme d'un contour de F_0 en prenant comme référence des points liés aux accents (maximum du pic d'accent, minimum de la vallée qui suit un accent). Pour chaque phrase, il n'y a qu'une seule valeur de S et F. Par contre, il y a un nombre variable d'accents. Pour chacun des quatre paramètres, leur valeur moyenne est calculée de manière à obtenir un jeu de paramètres $(\overline{S}, \overline{H}, \overline{L}, \overline{F})$ représentatif du locuteur.

La fonction de transformation $\mathcal{F}$ est définie comme une fonction linéaire par morceaux. Trois plages de valeurs de F_0 source sont distinguées dans la fonction. Cette fonction consiste sur chaque intervalle en un redimensionnement de la valeur de fréquence fondamentale source F_0^s et est définie par l'équation (4.5) :

$$\mathcal{F}(F_0^s) = \begin{cases} \overline{F^c} + \frac{(F_0^s - \overline{F^s})(\overline{L^c} - \overline{F^c})}{\overline{L^s} - \overline{F^s}} & F_0^s < \overline{L^s} \\ \overline{L^c} + \frac{(F_0^s - \overline{L^s})(\overline{H^c} - \overline{L^c})}{\overline{H^s} - \overline{L^s}} & \overline{L^s} < F_0^s < \overline{H^s} \\ \overline{H^c} + \frac{(F_0^s - \overline{H^s})(\overline{S^c} - \overline{H^c})}{\overline{S^s} - \overline{H^s}} & \overline{H^s} < F_0^s \end{cases} \tag{4.5}$$

Une étude perceptive pour l'évaluation de la performance et de la qualité de transformation a été réalisée par Gillett et King (2003). Cette évaluation montre une préférence pour cette méthode par rapport à la méthode de normalisation gaussienne présentée au début de ce chapitre. Cependant, leurs expérimentations reposent sur des corpus pour lesquels les paramètres des contours de F_0 sont extraits manuellement. L'extraction automatique de ces paramètres devient alors une difficulté à l'application de cette transformation. Enfin, des corpus parallèles ne sont pas requis pour mettre en œuvre cette méthode.

4.1.5 Table de correspondance

Cette approche, proposée par Chappell et Hansen (1998) sous le terme *codebook*, repose sur un alignement temporel dynamique (DTW) pour choisir, dans une table de correspondance source/cible, le contour source qui est le plus proche du contour de test au sens du coût de la DTW.

Sous l'hypothèse que les phrases prononcées par le locuteur source et le locuteur cible sont les mêmes, on peut alors remplacer le contour à transformer par le contour cible correspondant au contour source sélectionné. Ce principe met en œuvre une table de correspondance directe entre les contours sources et cibles correspondant aux mêmes phrases. Le système complet est présenté sur la figure 4.4 page suivante.

La transformation d'un contour de F_0 est réalisée par les étapes suivantes :

1. Sélection du contour source C_A, dans la base de données A, le plus proche du contour de test C_T par DTW ;
2. Alignement par DTW du contour C_B de la base de données B associé à C_A sur le contour source C_A pour obtenir le contour aligné C'_B ;
3. Alignement de C'_B sur le contour de test C_T par DTW ;
4. Modification du pitch de la phrase de test par TD-PSOLA.

Ce système réduit l'apprentissage de la fonction de conversion à la construction d'une table de correspondance entre les contours source et cible. L'approche proposée par Chappell et Hansen repose sur des contours de F_0 au niveau de la phrase mais d'autres déclinaisons de la méthode peuvent être envisagées au niveau des phones, des

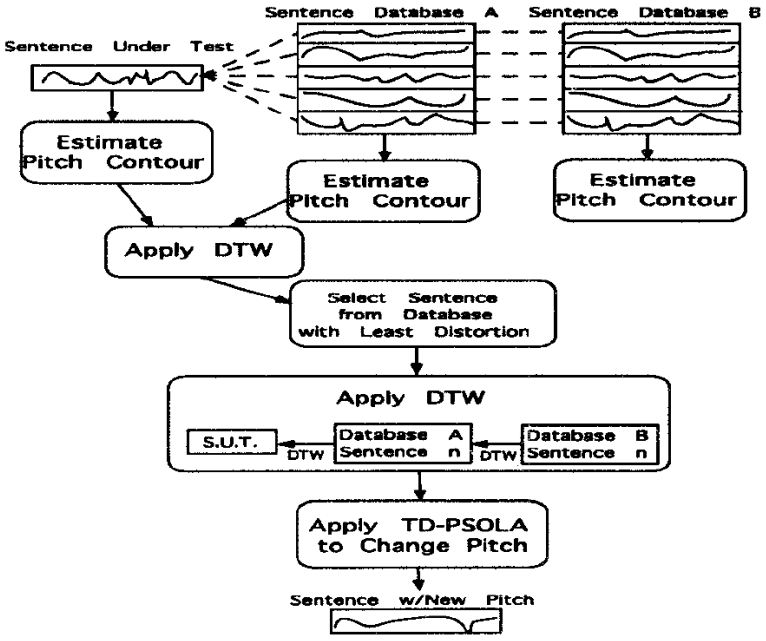


FIG. 4.4 – Diagramme de flot pour la transformation par codebook. (extrait de (Chappell et Hansen, 1998))

syllabes ou des mots par exemple. L'alignement par DTW pourrait également être amélioré en ajoutant des contraintes au niveau des frontières de phones ou de mots pour diminuer les effets liés au texte.

La construction de la table de correspondance est un point crucial de cette approche même si Chappell et Hansen ne mentionnent pas la méthodologie de construction de leur *codebook*. On peut faire l'hypothèse que tous les contours de la base d'apprentissage sont présents dans le codebook mais pour améliorer la représentativité des contours de la table de correspondance, une quantification pourrait être appliquée. De plus, comme l'indique Inanoglu (2003), si l'on souhaite transformer la prosodie d'une question, alors il est nécessaire que des contours de F_0 correspondant à des questions soient présents dans la table de correspondance.

Le principe de cette approche est que si le locuteur source produit un contour similaire à ceux qu'il a déjà produits, alors le locuteur cible fera de même. Pour relâcher cette hypothèse, Inanoglu propose une variante qui consiste, au lieu de ne retenir que le contour le plus proche du contour de test, à retenir les n contours sources les plus proches du contour de test. Le contour transformé est ensuite obtenu par une somme pondérée des contours cibles sélectionnés. Une variabilité plus grande des contours transformés est ainsi obtenue.

L'article récent de Helander et Nurminen (2007) met en œuvre une approche de type codebook au niveau de la syllabe. Leur codebook prend en compte des informations linguistiques et de durée associées à chaque syllabe. Le F_0 d'une syllabe est stylisé par les M premiers coefficients d'une DCT (Discrete Cosine Transform). Un arbre CART (Classification And Regression Tree) est utilisé pour prédire l'entrée du codebook à utiliser.

L'intérêt de cette méthode, en comparaison des méthodes précédentes, est d'utiliser des contours réels, issus de l'espace prosodique du locuteur cible. L'utilisation de contours au niveau de la phrase dans la table de correspondance permet alors d'obtenir des contours réellement produits par le locuteur et supprime les effets dus à la concaténation d'éléments plus courts. Cependant, du fait de la taille limitée de la table de correspondance, tous les contours possibles ne sont pas présents dans celle-ci. L'inconvénient par rapport à un modèle génératif est alors que la variabilité des contours mélodiques est restreinte aux contours présents dans la table de correspondance.

4.1.6 Fonction de transformation GMM

Le GMM est un modèle paramétrique classique utilisé dans de nombreuses applications de reconnaissance de formes. Pour la conversion de la prosodie, on peut se référer aux travaux de Inanoglu (2003) et de Kang *et al.* (2006). Un GMM fait l'hypothèse que les observations sont issues d'une densité de probabilité de la forme suivante :

$$p(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \mathcal{N}(\mathbf{x}; \mu_i, \Sigma_i) \tag{4.6}$$

où $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$ représente une densité de loi normale de dimension p, de moyenne α et de covariance Σ, définie par :

$$\mathcal{N}(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right] \tag{4.7}$$

Dans l'équation (4.6), les α_i sont des poids positifs dont la somme vaut 1. Les vecteurs d'observations $\{\mathbf{x}_t\}$ sont considérés deux à deux indépendants.

La méthodologie de transformation requiert des contours de F_0 source et cible alignés. Ils peuvent être alignés de différentes manières. Par exemple, une DTW (Dynamic Time Warping) peut être utilisée pour mettre en correspondance le F_0 trame par trame entre le locuteur source et le locuteur cible. Un alignement peut aussi être mis en œuvre pour construire des couples de F_0 moyen au niveau des phones.

L'utilisation d'un GMM pour la transformation du F_0 s'appuie sur la méthodologie

utilisée pour la conversion de voix sur le plan segmental par Stylianou *et al.* (1998) transposée au plan supra-segmental. Dans le cas présent, les vecteurs de données sont de dimension 1.

La relation entre les valeurs de F_0 source et cible est décrite par l'équation (4.8). Le GMM utilisé dans la suite, de paramètres ($\alpha_i, \mu_i, \sigma_i$ pour $i = 1, \ldots, m$), est appris sur l'ensemble des valeurs de F_0 source.

$$F_0^t = \mathcal{F}(F_0^s) = \sum_{i=1}^{m} P(C_i|F_0^s) \left[a_i + b_i \frac{(F_0^s - \mu_i)}{\sigma_{s_i}^2} \right] \tag{4.8}$$

où $P(C_i|F_0^s)$ est la probabilité d'appartenance d'une observation F_0^s à la classe C_i du GMM. L'estimation des paramètres de la fonction de transformation a_i et b_i pour chaque gaussienne i peut être réalisée par une estimation au sens des moindres carrés.

La méthode de transformation par normalisation gaussienne est appliquée au niveau de chaque gaussienne. La valeur de F_0 transformée est obtenue par une somme pondérée des différentes fonctions de transformation. Ainsi, la transformation du F_0 par GMM est une extension du cas particulier de normalisation gaussienne qui correspondrait à un GMM à une seule gaussienne.

Une restriction importante de ce modèle de transformation est qu'il est nécessaire de disposer de corpus parallèles. Cela signifie qu'il faut que les locuteurs aient prononcé le même texte et que la qualité de la transformation repose en partie sur la qualité de l'alignement des contours de F_0.

Les résultats rapportés par Inanoglu (2003) font état d'une erreur moyenne de conversion des contours située entre 21.37Hz dans le meilleur des cas et 38.54Hz dans le cas le plus défavorable. Le GMM utilisé possède deux composantes gaussiennes ce qui est très peu. De plus il avance que les résultats obtenus avec un tel GMM sont très similaires à ceux d'un scatterplot (voir paragraphe 4.1.2 page 74).

4.1.7 Transformation par arbre de régression et de classification

Les arbres de régression et de classification, CART, représentent une technique non paramétrique de classification hiérarchique de données. Ils peuvent notamment prendre en compte aussi bien des données numériques que nominales.

Cette technique, déjà mise en œuvre dans le contexte de la synthèse de parole pour la génération de contours mélodiques, peut également être utilisée pour la transformation de ces mêmes contours entre deux locuteurs.

Tao *et al.* (2006) proposent l'utilisation de cette technique pour modéliser la transformation d'une prosodie neutre vers une prosodie émotionnelle en intégrant des connais-

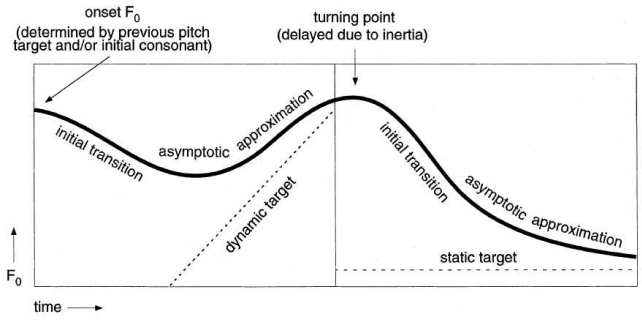


FIG. 4.5 – Illustration des *pitch targets* et de leur réalisation de surface. On peut observer comment le contour de F_0 de surface tend progressivement vers la cible. Ce modèle fait une analogie avec le délai nécessaire au système articulatoire pour se positionner dans la bonne configuration. (extrait de (Xu et Wang, 2001))

sances linguistiques.

Les contours de F_0 sont modélisés par le *Pitch Target Model* (Xu et Wang, 2001). Ce modèle fait l'hypothèse que le F_0 est une forme de surface dépendante des contraintes articulatoires, influencée par les mouvements du système articulatoire. Les mouvements du contour de fréquence fondamentale sont alors vus comme des mouvements cherchant à atteindre une cible particulière, tout comme la forme du système articulatoire est modifiée pour atteindre une position et une configuration particulière pour réaliser un son. La figure 4.5 illustre cette idée et montre l'évolution du contour de surface par rapport à deux cibles qui se succèdent. La première est une cible dynamique qui correspond à un mouvement de montée du F_0, tandis que la seconde est une cible statique correspondant à un niveau bas. Dans les deux cas, la cible n'est pas atteinte immédiatement, un délai dû à un effet d'inertie est observable.

En faisant l'hypothèse qu'une syllabe est définie sur l'intervalle $[0, D]$, ce modèle peut s'écrire sous la forme :

$$T(t) = at + b \tag{4.9}$$

$$F_0(t) = \beta \exp(-\lambda t) + at + b \tag{4.10}$$

$$0 \leq t \leq D, 0 \leq \lambda \tag{4.11}$$

$T(t)$ est la cible sous-jacente au contour de surface $F_0(t)$. Les paramètres a et b indiquent respectivement la pente et l'offset de la cible et permettent de décrire une cible statique ou dynamique. β est un paramètre qui mesure la distance par rapport à la cible à atteindre à $t = 0$. λ indique la vitesse à laquelle la cible est approchée. Plus λ est grand, plus la cible est approchée rapidement.

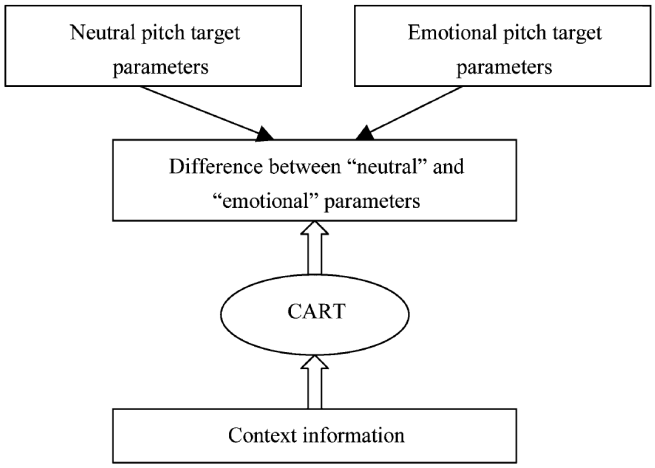


FIG. 4.6 – Système de transformation de la prosodie reposant sur CART proposée par Tao et al. (extrait de (Tao *et al.*, 2006))

Une syllabe est donc représentée par le jeu de paramètres a, b, β, λ. Ces paramètres doivent être transformés pour réaliser la modification de prosodie. L'approche proposée par Tao et al. est décrite par la figure 4.6. Les paramètres définis précédemment sont estimés pour toutes les syllabes avec une prosodie neutre et pour toutes celles ayant une prosodie comportant l'émotion désirée. Pour chaque couple de syllabes, la différence entre les paramètres neutres et émotifs est calculée. Cette paramétrisation utilise les couples de syllabe source et cible, ce qui introduit la contrainte de disposer de données parallèles. Néanmoins, elle permet de ne conserver que l'information utile pour la transformation qui répond à la question « que doit-on ajouter au jeu de paramètres correspondant à une prosodie neutre pour obtenir une prosodie émotionnelle ? ».

Ces vecteurs de différence entre les deux jeux de paramètres sont utilisés pour construire un arbre de régression et de classification, CART. Les attributs utilisés pour la prédiction du vecteur de différence dans CART sont de nature linguistique et permettent de prendre en compte des informations contextuelles à la syllabe : ton courant, précédent, suivant (5 catégories) ; type de phonème initial des syllabes courante et suivante (8 catégories) ; type de phonème final des syllabes courante et précédente (4 catégories) ; position dans la phrase et le Part-Of-Speech (30 catégories).

Tao *et al.* évaluent cette fonction de transformation de la prosodie en la comparant à une conversion par GMM. L'évaluation repose sur un test subjectif associé à une mesure de déviation de l'émotion perçue (DPE, Deviation of Perceived Expressiveness) qui consiste à demander à des testeurs d'annoter un ensemble de phrases selon l'émotion perçue (colère, tristesse, peur, joie) avec plusieurs degrés d'intensité (forte, moyenne

et faible). La mesure de déviation est ensuite appliquée pour évaluer la qualité de la prosodie transformée. Le second test appliqué consiste à mesurer l'erreur de prédiction des paramètres par rapport aux paramètres optimaux pour la prosodie cible.

Ces deux tests montrent que l'approche GMM (cf. paragraphe 4.1.6 page 81) est meilleure que l'approche CART même si le test subjectif ne révèle pas de différence importante. Les résultats présentés montrent également une incohérence avec le fait que les attributs linguistiques sont étroitement liés aux contours prosodiques réalisés. Selon eux, une explication possible est qu'un corpus plus grand serait nécessaire.

4.2 Évaluation de la transformation

L'évaluation de la fonction de transformation de prosodie est un aspect crucial pour tout système de transformation. Dans ce domaine, même s'il n'existe pas de standard d'évaluation, nous allons voir qu'un certain nombre de techniques sont couramment employées. En toute rigueur, un système de transformation de la prosodie doit, pour être complet, transformer toutes les caractéristiques supra-segmentales d'un locuteur pour les transposer dans l'espace d'un autre locuteur. Cependant, dans le cadre de cette thèse, nous allons nous focaliser sur la modification du F_0 qui apparaît comme un facteur prosodique proéminent.

Concernant l'évaluation de la fonction de transformation, deux stratégies complémentaires peuvent être envisagées : les évaluations subjective ou objective. La première est dite subjective dans le sens où le résultat dépend de l'opinion, du niveau d'expertise du testeur. Ce type d'évaluation, dans le domaine du traitement automatique de la parole et en particulier pour la synthèse de parole, est indispensable. En effet, l'objectif premier de la synthèse de parole est de créer de manière synthétique du signal de parole pour permettre la communication orale dans le sens machine vers Homme. Il est donc nécessaire que ce soit l'utilisateur final, l'Homme, qui évalue par exemple la qualité ou encore le naturel d'une voix de synthèse. Ce type d'évaluation ne peut pas, a priori, être remplacé par une évaluation objective qui doit être considérée comme complémentaire.

L'évaluation objective repose sur une mesure, le plus souvent une mesure d'erreur, qui doit être représentative de l'objectif de l'évaluation. Par exemple, une erreur quadratique moyenne (RMSE, Root Mean Square Error) faible sur le F_0 est-elle représentative de la qualité d'une transformation du F_0 ? On peut dire que si l'erreur est nulle, alors le contour transformé est parfait. Cependant, dans quelle mesure l'évolution de l'erreur est-elle liée à la dégradation de la qualité ? Ces questions montrent bien la complémentarité des approches objectives et subjectives.

4.2.1 Évaluation subjective

L'évaluation subjective d'un système de transformation doit faire appel à au moins deux types de données qui sont des données transformées et des données du locuteur cible. D'après, Ceyssens *et al.* (2002), il est nécessaire d'évaluer la transformation d'une seule caractéristique acoustique supra-segmentale à la fois, les autres étant simplement recopiées du locuteur cible vers le locuteur source. Ce point de vue pose donc la contrainte de disposer de phrases identiques pour les locuteurs source et cible. Cette approche permet néanmoins d'évaluer la contribution de chaque caractéristique indépendamment des autres. Dans le cas où on dispose de phrases communes pour les deux locuteurs, on peut, à la manière de Ceyssens *et al.* (2002), réaliser un test de préférence de type ABX qui consiste à décider si X ressemble plus à A ou B. Il s'agit d'une réponse binaire qui peut être étendue comme le fait Hanzlicek et Matousek (2007), en considérant une réponse graduelle avec 5 niveaux (comme A, plutôt comme A, pas de différence, plutôt comme B, comme B). Cependant, Helander et Nurminen (2007) utilisent une méthodologie de test permettant de ne pas se contraindre à utiliser des corpus disposant de phrases communes. Ce test est effectué pour déterminer laquelle des deux méthodes de transformation l'utilisateur préfère. La méthodologie est la suivante :

- le testeur écoute plusieurs phrases du locuteur cible pour se familiariser avec son style d'élocution,
- le test est réalisé en proposant, à chaque étape, une phrase avec chaque méthode,
- la question est : « quelle phrase est la plus proche du style d'élocution du locuteur cible ? »,
- les phrases de test sont différentes des phrases présentées pour le locuteur cible.

Cette méthodologie peut être mise en analogie avec des situations de la vie courante. Par exemple, lorsque l'on reçoit un coup de téléphone, on reconnaît la personne qui parle grâce à une certaine habitude par rapport à son timbre mais aussi son style d'élocution. La méthodologie d'évaluation précédente suit ce principe en présentant au testeur les phrases du locuteur cible en premier et seulement ensuite les différentes phrases de test.

Ces évaluations ont pour but d'évaluer la distance de la prosodie transformée par rapport à celle du locuteur cible. On peut également tenter d'évaluer la qualité et le naturel de la prosodie transformée en effectuant des tests de type MOS, Mean Opinion Score. L'objectif de ce type de test est de comparer plusieurs systèmes en notant leur performance par rapport à un critère sur une échelle, par exemple de 1 à 5. En général, on utilise au moins trois « systèmes » que l'on souhaite comparer : la prosodie naturelle du locuteur cible qui est la borne maximale que l'on cherche à atteindre, un système simple qui sert de borne minimale pour étalonner le test et le système que l'on souhaite

évaluer qui devrait se situer entre ces deux bornes.

Un test subjectif est réalisé en utilisant de la parole et contient donc à la fois les caractéristiques acoustiques segmentales et supra-segmentales. Ces caractéristiques sont étroitement liées et comparer la prosodie de deux voix qui possèdent une composante segmentale différente (deux voix) pose une difficulté. En effet, il faut prendre garde à ce que les résultats du test ne soient pas biaisés par le fait que les composantes segmentales sont trop différentes. Une manière de s'affranchir de ce problème, dans le cas où on dispose de phrases communes pour le test, est de « transplanter » les caractéristiques supra-segmentales non modifiées d'un locuteur vers l'autre, de manière à ne conserver le segmental que d'un seul locuteur à chaque étape du test.

4.2.2 Appliquer la prosodie au signal de parole

Comme nous l'avons évoqué précédemment, lorsque l'on dispose du signal de parole du locuteur source, il est nécessaire de pouvoir modifier ses caractéristiques supra-segmentales lorsqu'elles sont transformées ou bien tout simplement pour recopier celles du locuteur cible.

Une solution, pour modifier la durée et la fréquence fondamentale de manière indépendante et sans altérer le timbre, peut être trouvée dans les techniques dérivées de PSOLA, Pitch-Synchronous OverLapp and Add. La méthode PSOLA permet de modifier le rythme d'un signal de parole mais aussi d'apporter des modifications au spectre du signal de parole, notamment pour le F_0. Plusieurs méthodes dérivées de PSOLA existent, on peut citer FD-PSOLA (Frequency Domain PSOLA), TD-PSOLA (Time Domain PSOLA), etc. Un historique ainsi qu'une description des méthodes de la famille PSOLA sont présentés par Boeffard (2004).

4.2.3 Corpus parallèles ou non parallèles

À plusieurs reprises, nous avons utilisé les termes de corpus parallèles ou non parallèles. Nous rappelons ici que la définition retenue dans ce document est que deux corpus sont considérés comme parallèles lorsqu'ils contiennent les mêmes phrases. Cela signifie, dans le cas de la transformation de prosodie, que les deux locuteurs ont prononcé les mêmes phrases.

Cette contrainte forte est souvent imposée par le modèle dont l'apprentissage des paramètres nécessite un parallélisme des données entre le locuteur source et le locuteur cible. C'est par exemple le cas de la transformation par Codebook. Un inconvénient majeur de ce type de corpus est qu'il nécessite un coût de développement assez im-

portant. Relâcher cette contrainte permettrait de concevoir des applications de manière plus souple et cela représente donc un enjeu important d'un système de transformation de la voix et en particulier de la prosodie.

4.3 Conclusion

Dans ce chapitre, nous avons présenté plusieurs approches destinées à effectuer la transformation de la prosodie, et en particulier le F_0, entre un locuteur source et un locuteur cible.

Une première approche, très simple, repose sur une normalisation des valeurs de F_0 du locuteur source pour qu'elles soient dans le domaine des valeurs de F_0 produites par le locuteur cible. Nous avons également décrit des approches généralisant celle-ci. Nous avons ensuite abordé des approches plus complexes telles la transformation par GMM, les Codebooks ou encore les arbres de régression et de classification, CART.

La plupart des méthodes présentées nécessitent des données parallèles ce qui contraint fortement leur utilisation. On peut également noter qu'il n'existe pas, à ma connaissance, d'étude comparative de ces différentes méthodes.

Conclusion de la première partie

Dans cette première partie, nous avons présenté les principaux mécanismes impliqués dans le processus de production de la parole. Nous avons également décrit la parole sur le plan acoustique et ainsi défini à quel niveau interviennent les paramètres prosodiques tels que la durée ou la fréquence fondamentale. La prosodie varie largement d'un individu à l'autre mais aussi pour un même individu. Elle influence de manière importante la compréhension et le naturel de la parole.

La problématique de la transformation de voix a également été décrite : il s'agit de modifier les paramètres acoustiques d'une phrase issue d'un locuteur source afin qu'elle soit perçue comme prononcée par un locuteur cible. Dans cette partie, nous avons noté que la prosodie n'est pas transformée de manière fine dans les systèmes de transformation actuels. L'objectif de cette thèse est donc de proposer une méthode de transformation de la prosodie efficace.

Différents modèles de la prosodie ont été présentés et ils se distinguent par leur précision, complexité et capacité d'explication des phénomènes phonologiques sous-jacents. Dans le cadre d'un système de synthèse de la parole ou de transformation de la parole, la modélisation de la prosodie d'un locuteur constitue une étape préalable importante.

Pour le domaine particulier de la transformation de la prosodie, bien que ce domaine soit assez récent, un nombre assez important de méthodes ont déjà été proposées. La plupart d'entre-elles nécessitent des données source et cible alignées ce qui constitue une contrainte importante. Du point de vue des performances, la prosodie transformée est souvent très proche de la prosodie source. Enfin, une méthodologie d'évaluation des fonctions de transformation claire et bien posée est nécessaire afin de noter et de comparer les performances des différentes méthodes.

Deuxième partie

Contributions

Introduction à la deuxième partie

Deux objectifs principaux concernant ce travail peuvent être distingués. Le premier concerne l'amélioration du naturel des voix de synthèse qui repose en grande partie sur la génération d'une prosodie de qualité. Le second se focalise sur la diversification des voix de synthèse par des mécanismes de transformation de la prosodie. Ces deux aspects peuvent se traduire par l'étude des trois sous-domaines suivants : la stylisation, la classification et enfin la transformation de la prosodie. Les deux premiers points permettent de représenter la prosodie d'un locuteur et il est possible qu'un seul et même modèle les regroupe. Le troisième point, la transformation de la prosodie, se concentre sur la modification de la prosodie d'un locuteur source, qui peut être une voix de synthèse, en cherchant à la faire atteindre une cible donnée. Dans tous les cas, il faut garder à l'esprit que l'objectif est de générer une prosodie pour une phrase donnée.

En ce plaçant sous l'angle de l'analyse, la stylisation permet de rendre compte de la prosodie d'une phrase. Le modèle de stylisation, dans le cadre d'un système TTS (*Text-To-Speech*), doit donc pouvoir être génératif. Les paramètres d'un contour mélodique doivent pouvoir être prédits à partir des informations pré-existantes, extraites du texte.

Suite à une étape de stylisation, la classification des contours mélodiques permet de ne conserver qu'un jeu de formes élémentaires représentatif de l'espace prosodique de ce locuteur. Les modèles que l'on peut qualifier de modèles d'analyse de la prosodie, comme ToBI, proposent un jeu de formes restreint, construit manuellement à partir de l'observation d'un ensemble de phrases. Un premier objectif serait de construire cet ensemble de manière automatique et non supervisée. Un second objectif serait de pouvoir prédire les contours ou pour le moins la classe des contours mélodiques à générer pour une phrase. Dans ce cadre, on peut envisager l'utilisation des étiquettes de classe obtenues pour annoter automatiquement un corpus et utiliser ces valeurs comme attributs pour générer la prosodie en utilisant, par exemple, un arbre de prédiction.

La transformation de la prosodie dont le but est de modifier la prosodie d'une phrase, d'un locuteur, de manière à ce qu'elle paraisse être prononcée par un autre locuteur,

peut reposer sur une étape de classification. En effet, deux stratégies sont envisageables : soit on construit une fonction globale de transformation à la manière de la normalisation gaussienne, soit on construit une fonction de transformation en s'appuyant sur un partitionnement préalable de l'espace prosodique du locuteur. Dans ce cas, chaque fonction de transformation est spécifique à une portion de l'espace prosodique. La principale difficulté réside dans l'appariement des différentes zones des espaces prosodiques source et cible. N'ayant pas encore trouvé de solution satisfaisante à ce point, nous avons relâché cette contrainte et proposé une méthodologie de transformation utilisant des corpus source et cible non alignés.

Dans le chapitre 5 page ci-contre, nous proposons un modèle génératif permettant de styliser la prosodie. Ce modèle repose sur un outil mathématique, les B-splines, qui permet de tenir compte explicitement des non-linéarités de la courbe de F_0, notamment au niveau des transitions voisées/non voisées. Ce modèle permet d'obtenir une stylisation très fine des contours mélodiques. En contrepartie, le pouvoir explicatif d'un tel modèle est quasi-inexistant et l'étude de ce modèle nous a montré qu'il est difficile de l'utiliser afin de générer de la prosodie. En particulier, les variations de longueur des contours mélodiques posent des problèmes pour les comparer entre-eux. Dans le chapitre 6 page 125, nous proposons une méthodologie qui repose sur l'utilisation de HMM dont la séquence d'états cachés permet d'absorber les variations de longueur des contours mélodiques. Le partitionnement des contours mélodiques est ensuite construit de manière hiérarchique descendante.

Une méthodologie de transformation reposant sur une adaptation par régression linéaire est présentée dans le chapitre 7 page 143. L'apport de cette méthodologie consiste à transformer la prosodie entre deux locuteurs sans nécessiter de parallélisme entre les données source et cible. Cependant, nous verrons que perdre totalement la correspondance entre les phrases source et cible pose des problèmes. En effet, on peut dans ce cas vouloir faire correspondre des contours mélodiques source et cible qui peuvent ne pas avoir de correspondance naturelle. Il apparaît donc nécessaire d'incorporer des données linguistiques, phonologiques, syntaxiques permettant de faire la correspondance entre les contours source et cible par le biais d'informations qui conditionnent la réalisation d'un contour. Une autre lacune de cette méthodologie est qu'elle ne prend pas en compte la dynamique à long terme de la prosodie. Néanmoins, il est possible de l'intégrer par le biais de l'extension des vecteurs représentant la prosodie d'une syllabe ou en prenant en compte des informations de plus haut niveau comme mentionnées précédemment.

Chapitre 5

Stylisation du F_0 par un modèle B-Spline

Ce chapitre décrit une approche qui permet la stylisation de courbes de F_0 par un modèle B-spline. Ce dernier est caractérisé par une suite de nœuds auxquels sont associés des points de contrôle. Dans le paragraphe 5.1, nous discuterons du choix de ce modèle pour représenter des contours mélodiques. Le modèle B-spline est ensuite décrit dans le paragraphe 5.2 page 98 puis l'estimation de ses paramètres est présentée. Le nombre de paramètres du modèle est optimisé par une méthodologie MDL, Minimum Description Length, présentée au paragraphe 5.4 page 108. Les expériences, présentées au paragraphe 5.7 page 117 sont réalisées sur un corpus de parole du français et comparent trois critères MDL.

5.1 Introduction

L'estimation d'un modèle paramétrique d'une courbe à partir d'un ensemble de points observés du plan est un problème largement couvert par la littérature. Pour définir des estimateurs efficaces, il est toutefois nécessaire d'intégrer certaines hypothèses de régularité. Les splines sont très souvent mises en œuvre car elles offrent de nombreuses facilités algorithmiques et des propriétés de régularité, telles les splines cubiques naturelles qui sont optimales pour un facteur de régularisation proportionnel à la courbure. En contrepartie, la modélisation spline nécessite des hypothèses supplémentaires de régularité globale.

Les modèles B-splines offrent une alternative intéressante aux splines. D'une part, on peut montrer que les splines sont un cas particulier de courbes B-splines et d'autre

part, le modèle B-spline définit intrinsèquement une notion de régularité locale variable grâce à l'ordre de multiplicité de ses nœuds. Enfin sur un plan algorithmique, il existe comme pour les splines, des implantations efficaces de calcul d'une base de B-splines (Unser, 1999).

En faisant l'hypothèse de courbes expérimentales ayant des régularités locales variables, nous avons un double objectif : estimer les paramètres d'un modèle de courbe B-spline qui minimisent une erreur de reconstruction et rechercher la classe de modèles la plus parcimonieuse. Il est en effet inutile de chercher à minimiser une erreur de reconstruction si l'effort demandé en nombre de paramètres est disproportionné, le cas extrême étant celui où le nombre de paramètres est supérieur au nombre de points observés.

À propos du premier objectif, la problématique concerne le placement des nœuds du modèle B-spline. On distingue une stratégie de placement régulier (Figueiredo *et al.*, 2000) d'une stratégie de placement libre à nombre fixé de nœuds. Un placement uniforme se révèle souvent être un mauvais choix (Burchard, 1974). Dans le cas d'un nombre libre de nœuds, l'optimisation est relativement simple à mettre en œuvre. Toute la difficulté revient à insérer suffisamment de paramètres pour suivre la courbe. Un placement libre des nœuds permet quant à lui de reproduire pleinement les irrégularités de la courbe. Cependant, dans ce cas, le problème d'optimisation est non contraint, non linéaire et non convexe. D'une complexité combinatoire forte, il présente de nombreux minima locaux et de nombreux points stationnaires sur la surface d'erreur (Jupp, 1978). Ce thème est sujet à une littérature abondante où l'on distingue principalement deux types d'approches : l'une déterministe, l'autre probabiliste. La première fait appel à des changements d'échelle et des techniques de Gauss-Newton (Jupp, 1978; Schwetlick et Schütze, 1995; Lindstrom, 1999; Beliakov, 2004) ou encore à des algorithmes de type glouton pour l'insertion et la suppression des nœuds, (Cham et Cipolla, 1999), tandis que la seconde, approche probabiliste, repose sur des techniques de type MCMC (Markov Chain Monte-Carlo), d'algorithmes génétiques ou de recuit-simulé (Hansen et Kooperberg, 2002).

Dans ce chapitre, nous ne discutons pas du choix d'une meilleure stratégie de placement libre des nœuds. Nous considérons un algorithme de recuit-simulé (Kirkpatrick *et al.*, 1983) dont nous présentons les étapes nécessaires à l'optimisation d'une séquence de nœuds d'un modèle B-spline. Nous apportons une solution originale à la caractérisation des paramètres, notamment en ce qui concerne la prise en compte de la multiplicité des nœuds.

Quant au second objectif, il s'agit d'estimer un nombre optimal de paramètres pour

le modèle B-spline. Notre cadre théorique est celui du critère MDL, Minimum Description Length. Ce dernier a pour but d'établir un compromis entre la qualité du modèle, évaluée par l'erreur de reconstruction, et sa complexité, représentée par le nombre de ses paramètres.

Nos hypothèses méthodologiques sont les suivantes :

1. Poser un modèle de lissage de courbes ouvertes par un modèle B-spline avec un positionnement libre des nœuds.
2. Contourner la difficulté combinatoire d'estimation des paramètres du modèle B-spline par une approche statistique de type recuit-simulé.
3. Enfin, contrôler le nombre de degrés de liberté du modèle par un critère MDL.

Dans ce chapitre, nous nous intéressons essentiellement à l'articulation entre l'estimation des paramètres, selon un critère des moindres carrés, et la pénalité MDL (Cham et Cipolla, 1999; Figueiredo *et al.*, 2000). Nous démontrons pour le modèle B-spline une famille de bornes qui permettent de définir un critère MDL plus efficace que les bornes asymptotiques usuelles (Hansen et Yu, 2001).

Nous avons choisi de modéliser les contours de F_0 à l'échelle des syllabes. Ces courbes comportent de nombreux changements de régularité locale. La littérature est importante sur ce sujet et propose essentiellement des modèles splines. On peut notamment citer l'algorithme MoMel qui fournit une représentation des contours mélodiques via une approximation quadratique (Hirst *et al.*, 2000) et le modèle de Sakai et Glass (2003) composé d'une somme de splines cubiques naturelles. Dans (Barbot *et al.*, 2005), une étude comparative des capacités de modélisation des contours mélodiques par des splines interpolantes et des courbes B-splines régressives est présentée.

La suite de ce chapitre est organisée de la manière suivante. Au cours du paragraphe 5.2 page suivante, le modèle B-spline et ses paramètres, nœuds et points de contrôle, sont présentés. Pour une courbe observée, on détermine les points de contrôle optimaux au sens des moindres carrés. Dans le paragraphe 5.3 page 104, on considère le critère du maximum de vraisemblance -Maximum Likelihood Estimate ou MLE- et l'algorithme du recuit simulé pour un placement adéquat des nœuds. Au paragraphe 5.4 page 108, les critères MDL pour optimiser le nombre de nœuds du modèle B-spline sont décrits. Le protocole expérimental est décrit au cours du paragraphe 5.5 page 112. Les modèles B-spline et spline régressifs sont comparés dans le paragraphe 5.6 page 113. Le modèle B-spline, plus performant que le modèle spline, sera retenu. Les critères MDL proposés sont alors évalués pour ce modèle et les résultats sont donnés dans le paragraphe 5.7 page 117.

5.2 Modélisation B-spline

Dans cette section, on présente les courbes B-splines et leur capacité à styliser une courbe ouverte. Afin de comparer les atouts d'une modélisation par une courbe B-spline à celle obtenue par une spline, on rappelle la définition des fonctions splines et leur connexion avec les courbes B-splines. Par ailleurs, on détermine la courbe B-spline optimale selon un critère de moindres carrés.

5.2.1 Description du modèle

Dans un premier temps, on introduit les fonctions B-splines et les courbes B-splines associées. On présente leurs principales propriétés et on évalue l'impact de leurs paramètres (De Boor, 1976).

Soit une suite croissante de réels $\mathbf{t} = (t_0, \ldots, t_k)$. La valeur t_i est appelée un nœud et son ordre de multiplicité désigne le nombre de fois où il apparaît dans la suite $\mathbf{t}$. À partir de la suite de nœuds, on définit les fonctions B-splines de degré m par récurrence :

– les B-splines de degré 0 sont données, pour $i = 0, \ldots, k-1$, par

$$B_0^i(t) \;=\; \begin{cases} 1 & \text{si } t \in [t_i, t_{i+1}[\\ 0 & \text{sinon.} \end{cases}$$

– les B-splines de degré m sont des combinaisons quasi-convexes de deux B-splines de degré $m-1$. Pour $i = 0, \ldots, k-m-1$

$$B_m^i(t) \;=\; \frac{t - t_i}{t_{i+m} - t_i} B_{m-1}^i(t) + \frac{t_{i+m+1} - t}{t_{i+m+1} - t_{i+1}} B_{m-1}^{i+1}(t)$$

où, par convention, les quotients sont égaux à 0 si $t_i = t_{i+m}$ ou $t_{i+1} = t_{i+m+1}$.

On remarque que pour définir des fonctions B-splines de degré m, on doit disposer d'au moins $m+2$ nœuds, c'est-à-dire que $k \geq m+1$. De plus, la fonction B_m^i est construite à partir de $B_0^i, \ldots, B_0^{i+m}$ et a donc pour support $[t_i, t_{i+m+1}[$. Plus précisément, sur chacun des sous-intervalles, de $[t_i, t_{i+1}[$ à $[t_{i+m}, t_{i+m+1}[$, B_m^i est une fonction polynomiale positive de degré m. On rappelle également que les fonctions B-splines vérifient la propriété suivante, communément appelée « partition de l'unité » :

$$\sum_{i=0}^{k-m-1} B_m^i(t) \;=\; 1, \text{pour tout } t \in [t_m, t_{k-m}[. \tag{5.1}$$

Une illustration est proposée sur la figure 5.1(a) page 101. Pour des splines cubiques, en choisissant un vecteur de nœuds $\mathbf{t} = (0, 1, 2, 3, 4, 5, 6)$, on obtient une base B-spline

composée de trois fonctions. Chaque fonction de la base possède une influence sur un intervalle de cinq nœuds adjacents. La somme des trois fonctions de la base est représentée en tirets et vaut 1 uniquement au point d'abscisse 3.

On peut à présent définir une courbe B-spline g de degré m associée à un vecteur nœud $\mathbf{t}$ comme une combinaison linéaire des fonctions B-splines de degré m :

$$g(t) \;=\; \sum_{i=0}^{k-m-1} c_i B_m^i(t)\,, \tag{5.2}$$

où $c_0, \ldots, c_{k-m-1}$ sont des réels appelés points de contrôle. La courbe g a pour support $[t_0, t_k[$ et est définie en t_k par prolongement par continuité. D'après la propriété de partition de l'unité et la positivité des fonctions B-splines, on peut également définir une courbe B-spline comme une somme pondérée de points de contrôle.

Afin de mieux comprendre le rôle de chacun des paramètres dans la définition d'une courbe B-spline, on étudie indépendamment leur influence. Supposons que l'on fasse varier la position du i-ème point de contrôle c_i. Ce changement modifie le terme $c_i B_m^i(t)$ dans (5.2). La fonction B_m^i étant nulle en dehors de $[t_i, t_{i+m+1}[$, l'impact du changement de c_i ne se propage pas à l'extérieur du segment de courbe correspondant et reste donc localisé. En ce qui concerne l'effet des nœuds sur la courbe, le principal facteur est leur ordre de multiplicité. En effet, si les $m+1$ premiers nœuds sont affectés à t_0 et les $m+1$ derniers nœuds à t_k, la propriété de partition de l'unité est vérifiée sur l'intervalle $[t_0, t_k[$ en entier avec

$$B_m^0(t_0) \;=\; B_m^{k-m-1}(t_k) \;=\; 1\,.$$

Dans notre cas, pour des B-splines cubiques, i.e. $m = 3$, avec un ordre de multiplicité égal à 4 pour les nœuds extrêmes, on obtient l'exemple donné sur la figure 5.1(b) page 101. Sur cette figure, on observe que la somme des fonctions de la base B-spline, représentée en tirets, est égale à 1.

La courbe B-spline commence alors au point (t_0, c_0) et se termine en (t_k, c_{k-m-1}). De plus, si l'on considère un nœud interne t_i, différent des nœuds situés aux extrémités t_0 et t_k, plus son ordre de multiplicité m_i est élevé et moins la courbe B-spline est régulière au point t_i. En effet, si la courbe B-spline de degré m est de classe $\mathcal{C}^{m-1}$ entre deux nœuds consécutifs, elle est de classe $\mathcal{C}^{m-m_i}$ au nœud t_i. Par conséquent, si $m_i = m-1$, on observe un changement de courbure de la courbe en t_i, si $m_i = m$ la courbe n'est alors pas dérivable en t_i et si $m_i > m$, la courbe est discontinue en t_i. La figure 5.2 page 102 présente de manière séparée les différentes fonctions pour une base B-spline.

L'impact de l'ordre de multiplicité d'un nœud sur la régularité des fonctions de la base est visible.

Dans ce document, on considère des courbes B-splines cubiques, i.e. de degré $m = 3$, sur un intervalle $[a, b]$ avec une séquence de nœuds dont les nœuds extrêmes sont de multiplicité $m + 1$, c'est-à-dire $t_0 = t_m = a$ et $t_{k-m} = t_k = b$. On appelle les nœuds intermédiaires, de t_{m+1} à t_{k-m-1}, les nœuds internes. Afin de simplifier les notations, on désigne par l le nombre de nœuds internes, multiplicité comprise, et l'on a :

$$l \;=\; k - 2m - 1\,.$$

Enfin, une courbe B-spline de degré m étant déterminée, d'après (5.2), par la donnée du vecteur nœud $\mathbf{t}$ qui caractérise la base des B-splines et des points de contrôle, on note le vecteur des paramètres :

$$\theta \;=\; (\; a, b, \underbrace{t_{m+1}, \ldots, t_{m+l}}_{\text{nœuds internes}}, \quad \underbrace{c_0, \ldots, c_{l+m}}_{\text{points de contrôle}} \quad) \;=\; \left(a, b, \theta^l\right) \tag{5.3}$$

où θ^l désigne le vecteur des nœuds internes et des points de contrôle. On note alors g_θ la courbe B-spline associée. Un exemple de courbe B-spline associée à ses points de contrôle est représenté sur la figure 5.3 page 103.

À présent, on peut représenter matriciellement une suite de points d'une courbe B-spline de la forme g_θ. Soit $x_0, \ldots, x_{N-1}$ une suite de valeurs dans l'intervalle $[a, b]$, on définit la matrice $\mathbf{B}$ qui a $B_m^i(x_j)$ comme élément de la j-ème ligne et i-ème colonne, et on a, pour tout $j = 0, \ldots, N-1$

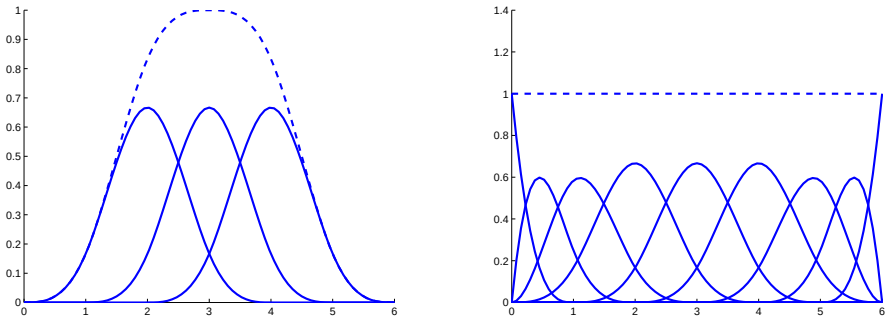
$$g_\theta\,(x_j) \;=\; \sum_{i=0}^{m+l} c_i B_m^i\,(x_j) \;=\; (\mathbf{Bc})_j \tag{5.4}$$

où $\mathbf{c}$ désigne le vecteur colonne des points de contrôle.

Avant de considérer l'estimation des différents paramètres du modèle B-spline g_θ que l'on a choisi, on rappelle dans le paragraphe suivant son lien avec les courbes splines généralement plus utilisées pour styliser des ensembles de points observés.

5.2.2 Discussion sur les splines

On introduit à présent la définition d'une fonction spline, sa relation avec les B-splines présentées précédemment et quelques travaux de modélisation spline des contours mélodiques qui correspondent à notre cadre expérimental.



(a) $\mathbf{t} = (0, 1, 2, 3, 4, 5, 6)$ (b) $\mathbf{t} = (0, 0, 0, 0, 1, 2, 3, 4, 5, 6, 6, 6, 6)$

FIG. 5.1 – Exemple de bases de fonctions B-splines cubiques.

Considérons une suite $\mathbf{u}$ strictement croissante de réels $u_0 < \ldots < u_{l+1}$. Une fonction spline de degré m sur $[u_0, u_{l+1}[$ est une fonction polynomiale de degré m sur chaque segment $[u_i, u_{i+1}[$ avec une régularité de classe $\mathcal{C}^{m-1}$ sur l'intervalle $[u_0, u_{l+1}[$. L'ensemble des fonctions splines de degré m associées à la suite $\mathbf{u}$ forme un espace vectoriel de dimension $l + m + 1$. Cet espace admet une base constituée de fonctions B-splines. Afin d'expliciter cette dernière, on définit à partir de $\mathbf{u}$ la suite de nœuds $\mathbf{t}$ en dupliquant m fois les valeurs extrêmes u_0 et u_l, i.e. $t_0 = t_m = u_0$ et $t_{m+l+1} = t_{m+2l+1} = u_l$, et en définissant les nœuds internes $t_{i+m} = u_i$ pour i de 1 à l. Les $(l + m + 1)$ fonctions B-splines de degré m caractérisées par $\mathbf{t}$ forment alors une base de l'espace vectoriel des splines. Par conséquent, une spline est une combinaison linéaire de fonctions B-splines et constitue un cas particulier de courbe B-spline très régulière, associée à une suite de nœuds internes tous distincts.

Les fonctions splines sont largement utilisées pour lisser une courbe issue de mesures physiques. Elles offrent pour cela des facilités algorithmiques et des propriétés de régularité permettant de satisfaire des contraintes d'interpolation, dérivation, courbure, d'erreur minimale. Par exemple, les splines cubiques naturelles sont généralement utilisées pour interpoler $l+2$ points avec des contraintes de dérivées secondes nulles à chaque extrémité de l'intervalle. Ces splines, dites « smoothing splines » (Unser, 1999), ont pour principale propriété d'avoir une courbure minimale, et donc de faibles oscillations, parmi les fonctions interpolantes de classe $\mathcal{C}^2$. Dans (Unser, 1999), ces différents aspects appliqués au traitement du signal et de l'image sont présentés, pour des séquences de nœuds équidistants.

En ce qui concerne la stylisation de la fréquence fondamentale, l'algorithme MoMel proposé par Hirst *et al.* (2000) a pour but de fournir une représentation d'une courbe

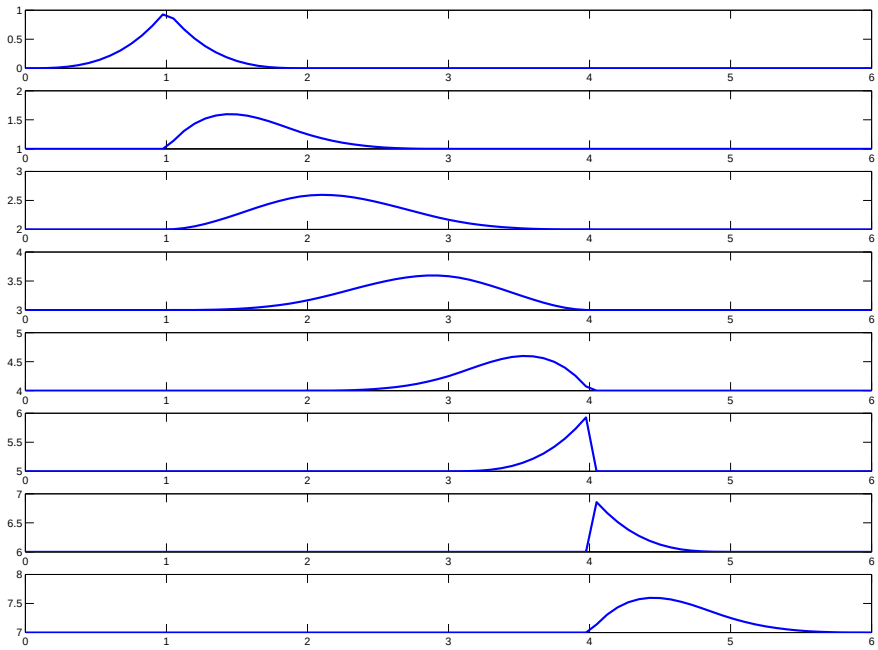


FIG. 5.2 – Influence de l'ordre de multiplicité des nœuds sur les fonctions de la base de B-splines cubiques avec $\mathbf{t} = (0, 1, 1, 1, 2, 3, 4, 4, 4, 4, 5, 6)$. De haut en bas, les fonctions B_3^0 à B_3^7 sont représentées.

mélodique via une approximation quadratique. L'algorithme estime des points cibles constituant les points stationnaires d'une spline de degré 2 dont le vecteur nœud est composé des abscisses des n points cibles et de leurs $(n-1)$ points médians. En effet, une spline quadratique est associée à $l = 2n-3$ nœuds internes ; elle est alors déterminée par les $2n$ contraintes données par ses n points stationnaires. En outre, on peut remarquer que les nœuds « médians » sont le lieu des points d'inflexion de cette spline. La qualité de cette approximation est évaluée selon une distance moyenne entre la courbe mélodique et la spline quadratique. Cette méthode d'interpolation permet une bonne approximation au voisinage des points cibles mais représente très partiellement les fluctuations de F_0 entre deux points cibles. Dans (Sakai et Glass, 2003), une approche plus globale est adoptée : la courbe F_0 est modélisée par la somme de deux fonctions $\mathcal{C}^2$, représentant respectivement les composantes de groupes intonationnel et accentuel. L'estimation du modèle est effectuée selon le critère des moindres carrés incluant une pénalité aux fortes courbures. Grâce à leur propriété de courbure minimale, le modèle correspond donc à

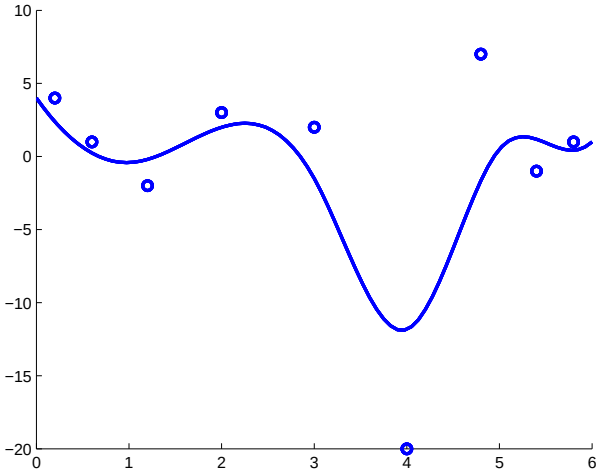


FIG. 5.3 – Exemple de courbe B-spline cubique avec ses points de contrôle et le vecteur de nœuds $\mathbf{t} = (0, 0, 0, 0, 1, 2, 3, 4, 5, 6, 6, 6, 6)$ et $\mathbf{c} = (4, 1, -2, 3, 2, -20, 7, -1, 1)$.

la somme de deux splines cubiques naturelles.

La modélisation à l'aide de courbes B-splines n'exclut pas une modélisation spline et permet de reproduire les fluctuations importantes, telles que les discontinuités, en incrémentant la multiplicité des nœuds. Les résultats expérimentaux présentés dans (Barbot *et al.*, 2005) confirment la meilleure capacité des courbes B-splines cubiques de type g_θ à représenter des contours mélodiques, selon le critère des moindres carrés, que les splines cubiques naturelles interpolantes. Dans le présent chapitre, on établit une comparaison entre les modèles régressifs B-splines et splines, c'est-à-dire estimés selon le critère des moindres carrés. Cela permet ainsi d'établir, s'il est nécessaire, un modèle intégrant une régularité locale variable pour styliser au mieux un contour mélodique.

5.2.3 Estimation des points de contrôle

Dans ce paragraphe, le vecteur nœud est supposé connu et on détermine la courbe B-spline, i.e. ses points de contrôle, minimisant l'erreur quadratique avec la courbe observée. Remarquons que pour un vecteur de nœuds internes deux à deux distincts, la courbe estimée est la spline de régression.

Soit un ensemble de mesures d'une courbe $\{(x_j, y_j), j = 0, \ldots, N-1\}$, où (x_j) est une suite croissante, on définit le vecteur colonne $\mathbf{y}$ ayant les coordonnées y_j. On souhaite obtenir la courbe B-spline g_θ de degré m telle que ses points $(x_j, g_\theta(x_j))$ minimisent l'erreur quadratique moyenne comparativement aux mesures. Pour un vecteur nœud $\mathbf{t}$

donné, où $t_0 = t_m = x_0$ et $t_{m+l+1} = t_{2m+l+1} = x_{N-1}$, on estime les points de contrôle tels que

$$\begin{aligned} \widehat{\mathbf{c}} &= \arg\min_{\mathbf{c}} \sum_{j=0}^{N-1} \left(y_j - (\mathbf{B}\mathbf{c})_j\right)^2 \\ &= \arg\min_{\mathbf{c}} \|\mathbf{y} - \mathbf{B}\mathbf{c}\|_2^2 \end{aligned}$$

d'après (5.4). Après dérivation, on obtient

$$\widehat{\mathbf{c}} = \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{y} \tag{5.5}$$

si la matrice $\mathbf{B}^T\mathbf{B}$ est inversible. Ceci est le cas lorsque les $m+l+1$ vecteurs colonnes de $\mathbf{B}$ sont linéairement indépendants. Pour cela, le vecteur $\mathbf{t}$ ne peut contenir de nœud de multiplicité supérieure à $m+1$, sous peine d'avoir une fonction B-spline nulle et donc une colonne de $\mathbf{B}$ nulle. De plus, le nombre N de lignes de cette matrice doit être supérieur ou égal au nombre de colonnes. Afin d'améliorer la stabilité numérique lors de l'inversion de $\mathbf{B}^T\mathbf{B}$, il est préférable que $\mathbf{B}$ dispose d'un nombre nettement plus important de lignes que de colonnes, permettant ainsi une meilleure différenciation de chaque fonction B-spline de la base et par conséquent de chaque colonne de $\mathbf{B}$. On suppose donc

$$N \gg m+l+1\,. \tag{5.6}$$

Ainsi, pour un vecteur de nœuds $\mathbf{t}$ fixé, la courbe B-spline de degré m optimale au sens des moindres carrés est décrite par la donnée des N points suivants :

$$x_j \rightarrow (\mathbf{B}\hat{\mathbf{c}})_j = \left(\mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{y}\right)_j\,. \tag{5.7}$$

5.3 Estimation des nœuds

Dans le paragraphe 5.2 page 98, on a présenté le modèle B-spline de degré m ainsi que l'optimisation des points de contrôle selon un critère des moindres carrés. Cette optimisation suppose le vecteur noeud $\mathbf{t}$ connu, c'est-à-dire le nombre l de ses nœuds internes ainsi que leurs positions respectives.

Pour cette partie, on suppose l connu, son optimisation sera par la suite traitée dans la partie 5.4 page 108, et on cherche à déterminer un placement optimal des nœuds. On commence par introduire le critère du maximum de vraisemblance pour la sélection de $\mathbf{t}$. On choisit ensuite une stratégie pour déterminer une solution $\widehat{\mathbf{t}}$ qui satisfait ce

critère. On présente alors l'algorithme du recuit-simulé pour traiter ce problème de nature combinatoire.

5.3.1 Maximum de vraisemblance

Soit $\{(x_j, y_j)\, j = 0, \ldots, N-1\}$ un ensemble d'observations, où (x_j) est une suite croissante. On note e_j l'erreur commise entre l'observation y_j à l'instant x_j et sa modélisation B-spline $g_\theta(x_j)$ définie par (5.3), dont les nœuds extrêmes sont donnés par $t_0 = t_m = x_0$ et $t_{m+l+1} = t_{2m+l+1} = x_{N-1}$,

$$y_j \;=\; g_\theta\left(x_j\right) + e_j\,.$$

Afin de simplifier les calculs, on suppose les erreurs $e_0, \ldots, e_{N-1}$ indépendantes et identiquement distribuées selon une loi gaussienne centrée de variance σ^2. Par conséquent, la log-vraisemblance du modèle s'écrit :

$$\log p\left(\mathbf{y}; \theta\right) = -\frac{1}{2\sigma^2}\left\|\mathbf{y} - \mathbf{B}\mathbf{c}\right\|_2^2 - \frac{N}{2}\log\left(2\pi\sigma^2\right)\,. \tag{5.8}$$

On remarque que $\widehat{\mathbf{c}}$, défini dans le paragraphe précédent, est l'estimateur du maximum de vraisemblance pour $\mathbf{c}$. L'estimateur $\widehat{\mathbf{t}^*}$ du maximum de vraisemblance pour le vecteur de nœuds internes $\mathbf{t}^*$ vérifie :

$$\begin{aligned}\widehat{\mathbf{t}^*} &= \arg\min_{\mathbf{t}^*}\left\|\mathbf{y} - \mathbf{B}\hat{\mathbf{c}}\right\|_2^2 \\ &= \arg\min_{\mathbf{t}^*}\left\|\mathbf{y} - \mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{y}\right\|_2^2\end{aligned} \tag{5.9}$$

où la matrice $\mathbf{B}$ dépend du vecteur de nœuds internes $\mathbf{t}^*$.

5.3.2 Positionnement libre des nœuds

Étant donné une séquence libre de nœuds internes, $\mathbf{t}^*$, à laquelle on ajoute $(m+1)$ fois les valeurs x_0 et x_{N-1}, la projection de la courbe observée sur la base B-spline est donnée par l'équation (5.7). On considère que les nœuds se situent à des instants d'observation, ils peuvent alors être représentés par un entier entre 0 et N-1. Il s'agit de trouver un optimum à la fonctionnelle (5.9).

Avant de choisir une stratégie pour déterminer un placement optimal des nœuds, nous avons réalisé des expériences informelles de manière à caractériser la complexité de l'espace de recherche. Pour $l > 2$, nous avons noté que la surface d'erreur du processus d'optimisation est chaotique. Il est donc illusoire de vouloir appliquer des techniques

d'optimisation fondées sur un gradient. Nous avons choisi de mettre en œuvre une stratégie d'optimisation globale par un algorithme de type Monte-Carlo (recuit-simulé) de manière à éviter les nombreux minima locaux de la fonction d'erreur.

5.3.3 Optimisation par recuit-simulé

Dans cette partie, nous proposons un algorithme d'optimisation globale pour répondre au problème du placement libre des nœuds en utilisant une stratégie de type recuit-simulé (Simulated Annealing, SA).

L'optimisation par recuit-simulé s'inspire du processus de cristallisation du métal depuis son état liquide jusqu'à son état solide pendant qu'il se refroidit (Kirkpatrick *et al.*, 1983). La température est diminuée de façon très lente pour permettre à la structure du métal de se stabiliser à chaque température. Le but de ce procédé, très similaire à un problème d'optimisation combinatoire, est de trouver l'agencement des atomes qui minimise l'énergie globale du système (Rutenbar, 1989). Pour une température donnée, il est nécessaire de faire évoluer le système vers un équilibre thermodynamique. Un algorithme de type recuit-simulé reprend ces principes où l'optimisation concerne l'exploration d'un espace de recherche continu ou discret. Pour mener cette exploration, une simulation de Metropolis Monte Carlo permet, à une température donnée, de converger vers un tirage aléatoire optimal des paramètres. La température est ici un méta-paramètre qui joue sur la variance des distributions de probabilité.

L'algorithme démarre donc à une température élevée par une simulation de Metropolis Monte Carlo. Un pourcentage relativement élevé des étapes aléatoires impliquant une augmentation de l'énergie est admis. À l'issue d'un nombre suffisant d'étapes, la température est abaissée. La simulation de Metropolis Monte Carlo peut alors continuer. Ce processus est répété jusqu'à l'obtention de la température finale. L'algorithme SA réalise le scénario précédent. Il est composé de deux itérations imbriquées. La boucle externe positionne une valeur de température et la boucle interne lance une simulation de Metropolis Monte Carlo à la température fixée. La décroissance de la valeur de température ne dépend pas de la fonction à optimiser. Contrairement à un algorithme glouton qui choisit toujours la meilleure configuration à l'étape courante, l'algorithme SA peut décider ponctuellement d'un mauvais choix pour élargir son espace de recherche et éviter de rester dans un minimum local, (Granville *et al.*, 1994).

La principale difficulté concerne la relation entre les paramètres du modèle B-spline et les distributions aléatoires de l'algorithme SA (Ingber, 1996). Après plusieurs expériences de mise au point, nous proposons la description suivante (Barbot *et al.*, 2005; Lolive *et al.*, 2006c) :

1. L'algorithme SA échantillonne un vecteur $\mathbf{r}$ d'entiers dans $\{1, \ldots, N-2\}^l$.
2. On définit $\mathbf{v}$ un vecteur tel que $v_i = x_{r_i}$. $\mathbf{r}$ est donc interprété comme un vecteur d'indices du vecteur $\mathbf{x}$.
3. Un vecteur de nœuds internes $\mathbf{t}^*$ est défini en triant les coordonnées de $\mathbf{v}$ par ordre croissant.
4. Un vecteur de nœuds $\mathbf{t}$ est défini en ajoutant $(m+1)$ fois x_0 et x_{N-1} aux extrémités de $\mathbf{t}^*$.
5. Pour i parcourant le vecteur nœud dans l'ordre croissant, on fusionne avec t_i les nœuds qui le succèdent si la distance entre t_i et ses suivants est inférieure à 5% de l'intervalle $[x_0, x_{N-1}]$. On remplace alors dans $\mathbf{t}$ les nœuds fusionnés par la valeur t_i pour augmenter sa multiplicité. On réitère cette étape en choisissant ensuite le premier nœud $t_{i'}$ tel que sa distance avec t_i soit supérieure à 5% de l'intervalle $[x_0, x_{N-1}]$.

Au cours de l'optimisation, l'algorithme du recuit simulé choisit aléatoirement un vecteur nœud et tente de calculer le coût de la solution proposée. Si une fois les fusions éventuelles réalisées, le vecteur nœud $\mathbf{t}$ trié contient des nœuds de multiplicité supérieure à $m+1$, le vecteur est non valide et rejeté. Dans ce cas, le recuit simulé génère un autre vecteur de nœuds. Au fur et à mesure, le recuit simulé va ainsi converger vers des solutions correctes tout en optimisant la fonction de coût.

Pour les fonctions splines, si le nombre de nœuds est trop important, le vecteur que doit générer le recuit simulé est très contraint : les nœuds sont des entiers distincts. Dans ce cas, le nombre de vecteurs rejetés est très important et le ratio entre le nombre de vecteurs incorrects et le nombre de vecteurs corrects est très élevé. Pour cette situation, le recuit simulé ne parvient pas à ajuster les distributions des différents paramètres et ne réussit pas à converger vers une solution correcte. Si les contraintes pour ce cas particulier sont telles que le recuit simulé ne semble pas adapté, il s'agit d'un cas de fonctionnement limite et en pratique cela ne concerne qu'un ensemble très restreint de vecteurs solutions.

La dernière étape concerne l'optimisation du nombre de nœuds internes. Avant de l'aborder, section 5.4 page suivante, résumons les principaux choix et stratégies appliqués au modèle B-spline. Pour l fixé, on commence par estimer un vecteur nœud $\widehat{\mathbf{t}}$ optimal selon un critère MLE à l'aide d'un algorithme du recuit-simulé. On dispose alors de la matrice $\mathbf{B}$ associée. Les points de contrôle $\widehat{\mathbf{c}}$ optimaux sont calculés au sens des moindres carrés à partir des points expérimentaux et de $\mathbf{B}$. Par la suite, on désigne par $\widehat{\theta^l}$ le modèle estimé, ou matriciellement par $\mathbf{B}\widehat{\mathbf{c}}$, dont l est le seul paramètre libre.

Dans la section suivante, on cherche le meilleur modèle, i.e. une valeur de l optimale, selon un critère MDL.

5.4 B-splines et MDL

Une analogie entre le principe MDL et le rasoir d'Occam est proposée par (Hansen et Yu, 2001). Lorsque plusieurs explications sont possibles, le principe d'Occam conseille de choisir la plus simple. De la même manière, le but du MDL est de trouver le meilleur modèle permettant d'expliquer les données de la manière la plus simple. La longueur de description des données étant donné un modèle et la longueur de description du modèle permettent d'établir ce compromis. Il s'agit ici d'estimer la structure du modèle B-spline, c'est-à-dire le nombre de nœuds internes et leur positionnement, à l'aide d'un critère MDL (Lolive *et al.*, 2006d,b).

5.4.1 Principe de solution

On considère un ensemble de mesures $\mathbf{y}$ d'une courbe. On note $\widehat{\theta^l}$ l'estimation selon le maximum de vraisemblance du jeu de paramètres θ^l, défini par (5.3) pour l nœuds internes. Appliquer un critère MDL consiste à déterminer le modèle, ou encore le paramètre $\widehat{l}$, qui minimise la longueur de description $L(\mathbf{y})$ des données $\mathbf{y}$. D'après (Hansen et Yu, 2001),

$$\begin{aligned} \widehat{l} &= \arg\min_l L(\mathbf{y}) \\ &= \arg\min_l L\left(\widehat{\theta^l}\right) - \log_2 p\left(\mathbf{y}; \widehat{\theta^l}\right) , \end{aligned} \tag{5.10}$$

où $L\left(\widehat{\theta^l}\right)$ et $-\log_2 p\left(\mathbf{y}; \widehat{\theta^l}\right)$ désignent les longueurs de description respectivement du modèle estimé et des observations étant connu ce dernier. L'expression (5.8) fait intervenir la variance σ^2 de l'erreur que l'on estime par la variance empirique $\widehat{\sigma^2}$ des erreurs :

$$\begin{aligned} \widehat{\sigma^2} &= \frac{1}{N}\sum_{j=0}^{N-1}\left(y_j - (\mathbf{B}\widehat{\mathbf{c}})_j\right)^2 \\ &= \frac{1}{N}\left\|\mathbf{y} - \mathbf{B}\widehat{\mathbf{c}}\right\|_2^2 \end{aligned}$$

où $\mathbf{B}$ et $\widehat{\mathbf{c}}$ sont obtenus à partir du vecteur nœud $\widehat{\mathbf{t}}$ qui comporte l nœuds internes. Par conséquent, l'écart-type σ est estimé par la racine carrée de la moyenne des carrés des

erreurs entre les observations et la courbe B-spline optimale, également appelée erreur RMS (Root Mean Square). En injectant cette estimation dans (5.8), l'expression (5.10) devient :

$$L(\mathbf{y}) \;=\; L\left(\widehat{\theta^l}\right) + \frac{N}{2} + \frac{N}{2}\log_2(2\pi) + N\log_2\left(RMS\right) \;.$$

On considère à présent la longueur de description du vecteur $\widehat{\theta^l}$. Celui-ci est composé de l nœuds internes et $(l+m+1)$ points de contrôle. On suppose que tous les nœuds et les points de contrôle sont respectivement de même longueur de description.

Les nœuds internes sont positionnés dans $]x_0, x_{N-1}[$ à des instants d'observation. Un nœud est alors représenté par un entier entre 0 et $N-1$ et $\log_2(N)$ bits suffisent à son codage (on n'utilise pas de codes préfixés).

Quant aux points de contrôle, ce sont des paramètres à valeurs réelles estimés à partir des N observations. Lorsque N est grand, la longueur de description d'un paramètre réel est généralement approchée par $\log_2(\sqrt{N})$ qui est asymptotiquement optimale (Rissanen, 1989). Cependant, chaque point de contrôle ayant une influence locale sur le modèle B-spline $g_{\widehat{\theta^l}}$, et le nombre d'observations N pouvant être faible, cette approximation ne semble pas adaptée (Figueiredo *et al.*, 2000). On considère alors qu'une longueur de description d'un point de contrôle suit une loi a priori uniforme sur un intervalle borné $[-\alpha, \alpha]$ (Lee, 2001). Pour une précision ε sur la description d'un point de contrôle, sa longueur est donnée par $\log_2(\alpha) + 1 - \log_2(\varepsilon)$. La détermination ainsi que la discussion quant aux choix de α et ε sont traitées au cours du paragraphe suivant.

En résumé, la longueur de description de $\widehat{\theta^l}$ est décrite par

$$L\left(\widehat{\theta^l}\right) = l\log_2(N) + (m+l+1)\left(\log_2(\alpha) + 1 - \log_2(\varepsilon)\right) \;.$$

5.4.2 Bornes théoriques sur les points de contrôle

Afin de déterminer une densité a priori sur les points de contrôle $\widehat{\mathbf{c}}$, on les considère comme des observations d'une loi uniforme sur $[-\alpha, \alpha]$. L'estimation du maximum de vraisemblance pour α est $\|\widehat{\mathbf{c}}\|_\infty = \max_i |\widehat{c}_i|$. Ainsi, modulo une constante indépendante de l, on obtient un premier critère MDL :

$$\begin{aligned} L(\mathbf{y}) \;\;=\;\; & (m+l+1)\left(\log_2\left(\|\widehat{\mathbf{c}}\|_\infty\right) + 1 - \log_2(\varepsilon)\right) \\ & + N\log_2\left(RMS\right) + l\log_2(N) \end{aligned} \tag{5.11}$$

dénommé ***critère (a)***.

Supposons $\widehat{\mathbf{t}}$ connu. Selon l'expression (5.5) des points de contrôle $\widehat{\mathbf{c}}$ associés, on

établit ci-dessous une borne de leur support et on considère la loi a priori uniforme associée.

Proposition 5.4.1 *Soit μ la plus petite valeur singulière de la matrice $\mathbf{B}$, on a $\mu > 0$ et $\|\widehat{\mathbf{c}}\|_\infty \leq \dfrac{\|\mathbf{y}\|_2}{\mu}$.*

Preuve. La matrice $\mathbf{B}$ étant de rang maximal $m+l+1$, sa plus petite valeur singulière μ est strictement positive. On note λ sa plus grande valeur singulière et le rapport λ/μ est appelé conditionnement de $\mathbf{B}$. On considère la norme matricielle subordonnée à la norme vectorielle $\|.\|_2$, et l'on a $\|\mathbf{B}\|_2 = \lambda$. Le conditionnement de $\mathbf{B}$ vérifie

$$cond(\mathbf{B}) \;=\; \|\mathbf{B}\|_2 \left\| \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T \right\|_2 .$$

Ainsi,

$$\left\|(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\right\|_2 \;=\; \frac{1}{\mu}$$

et on en déduit , à l'aide de l'expression (5.5)

$$\begin{aligned} \|\widehat{\mathbf{c}}\|_\infty &\leq \|\widehat{\mathbf{c}}\|_2 \;=\; \left\|(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{y}\right\|_2 \\ &\leq \left\|(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\right\|_2 \|\mathbf{y}\|_2 \\ &\leq \frac{\|\mathbf{y}\|_2}{\mu}\,. \end{aligned}$$

□

D'après la proposition 5.4.1, si l'on considère que les points de contrôle sont distribués selon une loi a priori uniforme sur $\left[-\frac{\|\mathbf{y}\|_2}{\mu}, \frac{\|\mathbf{y}\|_2}{\mu}\right]$, on obtient le critère MDL

$$\begin{aligned} L(\mathbf{y}) \;=\; & (m+l+1)\left(\log_2\left(\frac{\|\mathbf{y}\|_2}{\mu}\right) + 1 - \log_2(\varepsilon)\right) \\ & + N\log_2(RMS) + l\log_2(N) \end{aligned} \tag{5.12}$$

dénommé ***critère (b)***.

5.4.3 Influence de la précision des points de contrôle

Pour $\widehat{\mathbf{t}}$ donné, on étudie à présent l'influence du choix de la précision de description ε des points de contrôle sur le modèle reconstruit. Selon le choix de l'évaluation de l'erreur entre les courbes estimée et reconstruite, on détermine une précision ε minimale nécessaire.

Proposition 5.4.2 *Soit* $\widetilde{\mathbf{c}}$ *une approximation de* $\widehat{\mathbf{c}}$ *telle que* $\|\widetilde{\mathbf{c}} - \widehat{\mathbf{c}}\|_\infty \leq \varepsilon$ *alors :*

$$\|\mathbf{B}\widetilde{\mathbf{c}} - \mathbf{B}\widehat{\mathbf{c}}\|_\infty \leq \varepsilon.$$

Preuve. On rappelle que $\|\mathbf{B}\|_\infty$ est la norme matricielle de $\mathbf{B}$ subordonnée à la norme vectorielle $\|\cdot\|_\infty$ et vaut d'après (5.1)

$$\|\mathbf{B}\|_\infty = \max_{j=0,\ldots,N-1} \sum_{i=0}^{m+l} \left|B_m^i(x_j)\right| = 1\,.$$

Par conséquent, pour $\|\widetilde{\mathbf{c}} - \widehat{\mathbf{c}}\|_\infty \leq \varepsilon$, on a :

$$\|\mathbf{B}\widetilde{\mathbf{c}} - \mathbf{B}\widehat{\mathbf{c}}\|_\infty \leq \|\mathbf{B}\|_\infty \|\widetilde{\mathbf{c}} - \widehat{\mathbf{c}}\|_\infty \leq \varepsilon\,.$$

□

Ainsi, si l'on souhaite un écart maximal entre les points des courbes B-splines estimée et reconstruite inférieur à l'écart entre les données et les points de la courbe estimée, il suffit de fixer une précision $\varepsilon = \|\mathbf{y} - \mathbf{B}\widehat{\mathbf{c}}\|_\infty$ sur la description des points de contrôle.

Corollaire 5.4.1 *Soit* $\widetilde{\mathbf{c}}$ *une approximation de* $\widehat{\mathbf{c}}$ *telle que* $\|\widetilde{\mathbf{c}} - \widehat{\mathbf{c}}\|_\infty \leq \varepsilon$ *alors l'erreur RMS entre les courbes optimale* $\mathbf{B}\widehat{\mathbf{c}}$ *et reconstruite* $\mathbf{B}\tilde{\mathbf{c}}$ *est inférieure à* ε*.*

Preuve. On rappelle que pour tout vecteur $\mathbf{x}$ de $\mathbb{R}^N$, on a :

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{N}\|\mathbf{x}\|_\infty\,. \tag{5.13}$$

Pour $\|\widetilde{\mathbf{c}} - \widehat{\mathbf{c}}\|_\infty \leq \varepsilon$, l'erreur RMS entre les courbes $\mathbf{B}\widetilde{\mathbf{c}}$ et $\mathbf{B}\widehat{\mathbf{c}}$ vérifie alors :

$$\frac{\|\mathbf{B}\widetilde{\mathbf{c}} - \mathbf{B}\widehat{\mathbf{c}}\|_2}{\sqrt{N}} \leq \|\mathbf{B}\widetilde{\mathbf{c}} - \mathbf{B}\widehat{\mathbf{c}}\|_\infty \leq \varepsilon$$

d'après la proposition 5.4.2. □

Ainsi, pour obtenir une erreur RMS entre les courbes B-splines estimée et reconstruite qui reste inférieure à l'erreur RMS entre les données et la courbe estimée, il suffit de fixer une précision de description des points de contrôle

$$\varepsilon = RMS = \frac{\|\mathbf{y} - \mathbf{B}\widehat{\mathbf{c}}\|_2}{\sqrt{N}}.$$

5.4.4 Critères MDL pour les B-splines

Dans le paragraphe 5.4.2 page 109, on a présenté deux critères MDL, nommés (a) et (b), selon le choix de la longueur de description des points de contrôle. Pour chacun d'entre eux, définis par (5.11) ou (5.12), on distingue trois sous-critères selon la précision ε sur les points de contrôle. Le premier cas considère ε fixe. Pour les second et troisième cas, ε est fonction de l'erreur de reconstruction du modèle B-spline. On considère qu'il n'est pas nécessaire que l'erreur entre les courbes estimée et reconstruite soit plus faible que celle entre les courbes observée et estimée. D'après le paragraphe 5.4.3 page 110, on choisit donc $\varepsilon = RMS$ pour le second cas et $\varepsilon = \|\mathbf{y} - \mathbf{B}\widehat{\mathbf{c}}\|_\infty$ pour le dernier. On obtient les trois versions des critères (a) et (b) :

(a).1 et (b).1 : ε fixe pour toutes les courbes,

(a).2 et (b).2 : $\varepsilon = RMS$, donc variable pour chaque courbe

(a).3 et (b).3 : $\varepsilon = \|\mathbf{y} - \mathbf{B}\widehat{\mathbf{c}}\|_\infty$, donc variable pour chaque courbe.

5.5 Protocole expérimental

Les expériences ont été réalisées sur un corpus de 500 phrases du français choisies aléatoirement (soit environ 7000 syllabes) dans un ensemble d'environ 7000 phrases enregistrées. L'enregistrement du corpus a été réalisé dans un studio d'enregistrement professionnel. Le signal acoustique a été annoté puis segmenté en unités acoustiques. La fréquence laryngienne moyenne, F_0, a été analysée de manière automatique en s'appuyant sur la fonction d'auto-corrélation du signal de parole. Ensuite, un algorithme a été appliqué à la chaîne d'unités phonétiques de manière à repérer chaque syllabe. On choisit la syllabe comme support minimal d'un contour mélodique. L'objectif est de mesurer les performances de la stylisation non supervisée de contours mélodiques en appliquant différents critères de décision.

Le modèle $\widehat{\theta^l}$ possède $(2l + m + 1)$ paramètres pour l nœuds internes. Si le nombre de paramètres est plus important que le nombre de valeurs de F_0, il est alors plus économique de conserver les valeurs de la courbe. Lors de l'estimation d'un modèle B-spline, il est nécessaire de respecter la condition portant sur le nombre de nœuds internes l :

$$N \geq (2l + m + 1). \tag{5.14}$$

Les courbes ayant un nombre différent d'observations, une normalisation du nombre de paramètres est préférable pour ensuite calculer et comparer des moyennes sur le nombre

de d.d.l. (degrés de liberté) pour l'ensemble des courbes. Pour cela, nous avons introduit le nombre de d.d.l. normalisé défini comme le rapport entre le nombre de paramètres d'un modèle B-spline et le nombre N de points observés.

$$ddl_n = \frac{2l + m + 1}{N}. \tag{5.15}$$

D'après les équations (5.14) et (5.15), ddl_n varie entre 0 et 1. De plus, un nombre de d.d.l. normalisé à 1 correspond alors à l'ensemble des points de la courbe, c'est-à-dire à un modèle complet.

Toutes les expériences proposées font intervenir le calcul des valeurs moyennes de l'erreur RMS et des degrés de liberté des modèles B-splines. Ces moyennes sont présentées avec des intervalles de confiance à 99%. Compte-tenu de la nature des expériences, il n'y a qu'un seul échantillon expérimental de 7000 courbes et les intervalles de confiance sont donc établis par ré-échantillonnage à partir des distributions empiriques (méthodologie bootstrap).

Dans la première partie de ce chapitre, nous avons présenté l'estimation de modèles B-spline et spline grâce à un algorithme de type recuit-simulé. Nous avons aussi exposé le principe du MDL ainsi que plusieurs critères permettant d'estimer un nombre optimal de nœuds pour un modèle B-spline. Dans la suite, nous présenterons des résultats expérimentaux qui montrent qu'un modèle B-spline est plus efficace, comme régresseur, qu'un modèle spline. Dans un second temps, nous estimerons les performances des divers critères MDL proposés.

Les expériences sont organisées de la manière suivante :

1. Comparaison entre une modélisation par courbe B-spline et courbe spline
2. Optimisation du nombre de nœuds pour un modèle B-spline :
   - Relation entre la RMS et les degrés de liberté d'un modèle,
   - Sensibilité des critères MDL par rapport à ε,
   - Analyse des critères MDL proposés.

5.6 Comparaison entre B-splines et splines : résultats et discussion

La figure 5.4 page suivante présente un exemple d'estimation de contour mélodique pour un modèle B-spline et un modèle spline (Lolive *et al.*, 2006a). L'estimation des

paramètres est réalisée pour 4 nœuds internes. Ce contour mélodique présente une discontinuité à l'abscisse 9. Pour refléter cette discontinuité, le modèle B-spline estimé contient un nœud de multiplicité 4. Quant au modèle spline, il ne peut faire mieux que de placer 4 nœuds autour de cette discontinuité. On constate ici la plus grande capacité de modélisation des B-splines d'une courbe à régularité locale variable. Quant à l'erreur RMS, le modèle B-spline donne 0.59Hz et le modèle spline 4Hz.

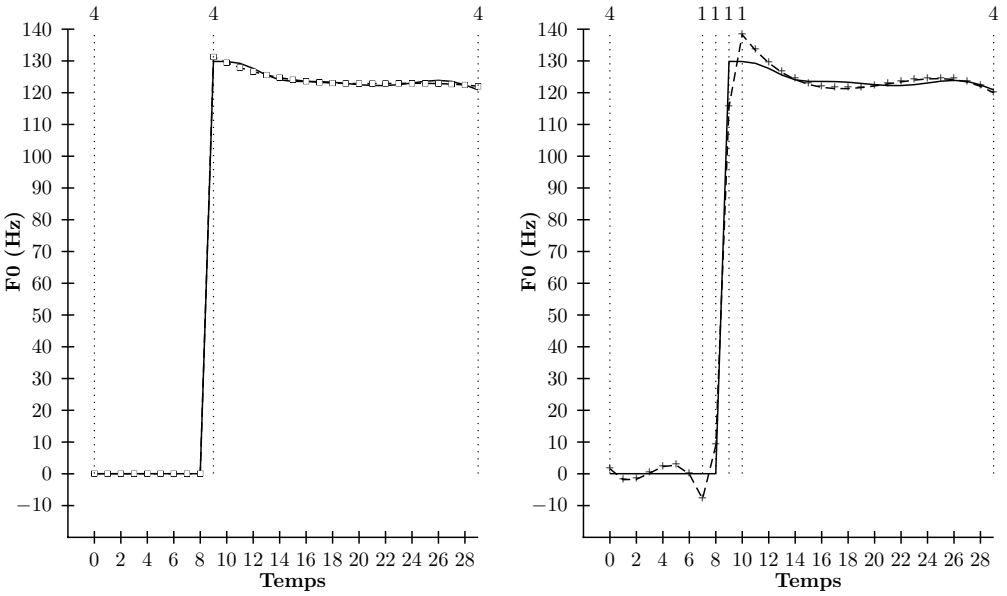


FIG. 5.4 – Exemple d'estimation d'un contour mélodique sur une syllabe pour un modèle B-spline (à gauche) et un modèle spline (à droite). La courbe originale est représentée en trait plein. Les courbes estimées sont tracées en traits pointillés avec à gauche la courbe B-spline et à droite la courbe spline. L'estimation est réalisée avec 4 nœuds internes soit 0.41 en d.d.l. normalisés. L'emplacement et la multiplicité des nœuds sont indiqués par les lignes verticales

Pour être en mesure de comparer les splines et les B-splines pour des degrés de liberté égaux, nous avons étudié le lien entre l'erreur RMS et le nombre de d.d.l. pour chaque modèle. La figure 5.5 page suivante permet de mesurer l'impact du nombre de paramètres sur la qualité d'estimation. Le nombre normalisé de d.d.l. présenté varie de 0 à 1. Toutes les courbes du corpus sont prises en compte. Lorsque le nombre de d.d.l. augmente, l'erreur RMS diminue. Cependant, pour les splines (courbe en pointillé), lorsque le nombre de d.d.l. normalisé est proche de 1, l'erreur RMS augmente. Il s'agit ici d'un cas limite de fonctionnement de l'algorithme de recuit simulé. En effet, comme les nœuds internes associés aux splines sont des entiers deux à deux distincts compris entre 1 et $N-2$, le nombre de vecteurs nœud valides est faible. Dans ce cas,

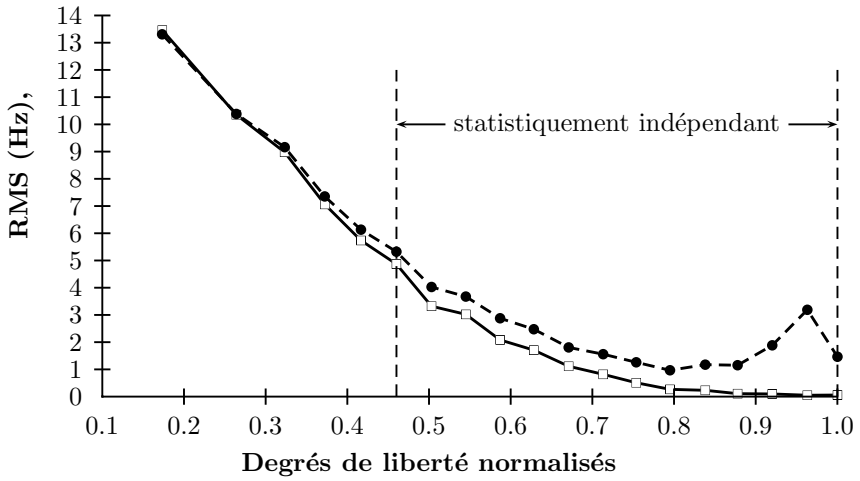


FIG. 5.5 – Évolution de l'erreur RMS moyenne en fonction du nombre normalisé de d.d.l. pour les B-splines (en trait plein) et les splines (en pointillé). L'axe des d.d.l. normalisés est divisé en 20 classes équilibrées. Ainsi chaque point des courbes représente un nombre identique de valeurs.

l'algorithme de recuit simulé ne converge pas vers une solution. Pour pallier ce problème, on pourrait envisager une structure différente pour décrire le vecteur de nœuds, par exemple en considérant un vecteur de décalage par rapport au premier nœud. Cette représentation permettrait notamment de diminuer le nombre de vecteurs non valides et ainsi d'accroître les performances du recuit simulé dans les cas limites. La courbe d'erreur des B-splines (courbe en trait plein) décroît de manière régulière jusqu'à 1. Le vecteur de nœuds des B-splines, qui permet la fusion de nœuds proches, est moins contraint que pour les splines. Dans ce cas, l'algorithme du recuit simulé n'a pas de difficulté à trouver une solution satisfaisante.

Sur la première partie de la courbe, on observe que les deux modèles sont équivalents. Le nombre de nœuds étant faible, il est normal que les B-splines et les splines aient des performances équivalentes. Avec un nombre faible de d.d.l. normalisé, la fusion de nœuds pour les B-splines n'apporte pas d'amélioration. D'une part, le faible nombre de nœuds implique que la fusion ne s'applique pas souvent, et d'autre part, lorsqu'elle s'applique l'amélioration n'est pas très importante. Lorsque le nombre de d.d.l. augmente, les B-splines prennent l'avantage et leur erreur RMS moyenne diminue plus rapidement que pour les splines.

Le tableau 5.1 page suivante présente le détail des intervalles de confiance à 99% de l'erreur RMS en fonction des d.d.l. normalisés. À partir de 0.5 d.d.l. normalisés, les deux modèles sont statistiquement indépendants et le modèle B-spline est significativement

TAB. 5.1 – Intervalles de confiance à 99% pour l'erreur RMS (Hz) et le nombre normalisé de d.d.l. pour chaque modèle.

d.d.l. norm.	splines		B-splines	
0.17	13.31	*±0.50*	13.46	*±0.46*
0.26	10.38	*±0.46*	10.35	*±0.43*
0.32	9.16	*±0.40*	8.98	*±0.43*
0.37	7.35	*±0.38*	7.06	*±0.41*
0.42	6.13	*±0.34*	5.74	*±0.36*
0.46	5.32	*±0.33*	4.87	*±0.33*
0.50	4.03	*±0.30*	3.32	*±0.26*
0.54	3.67	*±0.26*	3.02	*±0.23*
0.59	2.88	*±0.21*	2.08	*±0.21*
0.63	2.47	*±0.20*	1.71	*±0.19*
0.67	1.81	*±0.16*	1.11	*±0.15*
0.71	1.56	*±0.15*	0.82	*±0.12*
0.75	1.26	*±0.12*	0.51	*±0.09*
0.79	0.97	*±0.11*	0.27	*±0.06*
0.84	1.17	*±0.18*	0.23	*±0.05*
0.88	1.15	*±0.19*	0.11	*±0.03*
0.92	1.88	*±0.32*	0.10	*±0.03*
0.96	3.19	*±0.41*	0.05	*±0.01*
1.00	1.46	*±0.28*	0.06	*±0.02*

meilleur que le modèle spline. De plus, c'est à partir de ce même nombre de d.d.l. normalisé que les deux modèles atteignent une valeur d'erreur RMS de l'ordre de 4Hz. Pour cette valeur de 4Hz, les deux modèles autorisent un taux de compression moyen de 50%.

Les expérimentations menées sur un corpus du français montrent que le modèle B-spline donne de meilleurs résultats que le modèle spline. Néanmoins dans le contexte de la parole, les splines permettent d'atteindre un seuil minimal à partir duquel l'oreille est insensible aux variations de la mélodie (autour de 4 Hz (Klatt, 1973)) avec un facteur de compression de 50%. Quant aux courbes B-splines, plus générales que les splines, elles prennent en compte explicitement les discontinuités des contours mélodiques. Nous constatons que le recuit simulé est plus stable avec le modèle B-spline en raison du nombre plus limité de contraintes. Par la suite, nous retiendrons le modèle B-spline qui s'avère être le plus performant.

5.7 Optimisation du nombre de nœuds : résultats et discussion

Dans cette partie, nous présentons tout d'abord un exemple d'optimisation du nombre l de nœuds internes. Ensuite trois expériences sont présentées, elles permettent de répondre aux questions suivantes :

- Quelle est la relation entre l'erreur RMS et le nombre de d.d.l. ?
- Quelle est la sensibilité des critères MDL proposés par rapport à la précision ε ?
- Enfin, quel est le comportement des différents critères MDL proposés ?

Deux axes sont utilisés pour estimer la qualité des critères MDL : 1) l'erreur RMS commise avec le modèle sélectionné, 2) le gain en nombre de d.d.l.

5.7.1 Présentation d'un exemple

Dans ce paragraphe, nous présentons un exemple d'estimation de la structure d'un modèle pour un contour mélodique. Le contour mélodique considéré comporte $N = 43$ observations et est représenté sur la figure 5.6 page suivante.

La première étape consiste à estimer les différents modèles pour la courbe. Pour $l = 1$, on cherche la meilleure position du nœud interne par l'algorithme du recuit-simulé présenté précédemment. On effectue la même opération pour l allant de 2 à $\frac{N-m-1}{2}$ qui est le nombre maximal de nœuds internes possible, d'après l'équation (5.14). On obtient ainsi un ensemble de modèles, chacun correspondant à un nombre de d.d.l. possible.

Pour sélectionner le « meilleur » modèle, on applique ensuite le critère MDL. La figure 5.7 page 119 représente les valeurs des critères en fonction du nombre de nœuds internes. Le modèle choisi peut être différent d'un critère à l'autre suivant la pénalisation effectuée.

Intéressons nous, par exemple, au critère (a).3. Le modèle sélectionné par un critère est celui permettant d'atteindre le minimum de la longueur de description des données. Pour cet exemple, la figure 5.7 page 119 montre que le minimum pour ce critère est atteint avec 8 nœuds internes. Sur la figure 5.6 page suivante, les courbes observées et estimées sont représentées ainsi que l'emplacement et la multiplicité des nœuds. Les nœuds extrêmes ont une multiplicité égale à 4 imposée par le modèle, et l'on remarque que l'algorithme du recuit-simulé place un nœud interne de multiplicité 4 à l'abscisse 12 afin de reproduire une discontinuité du contour. Un nœud interne de multiplicité 3 est également placé à l'abscisse 40, impliquant une stylisation continue mais non dérivable à cet endroit.

Le modèle choisi par le critère (a).3 permet d'obtenir une bonne estimation du

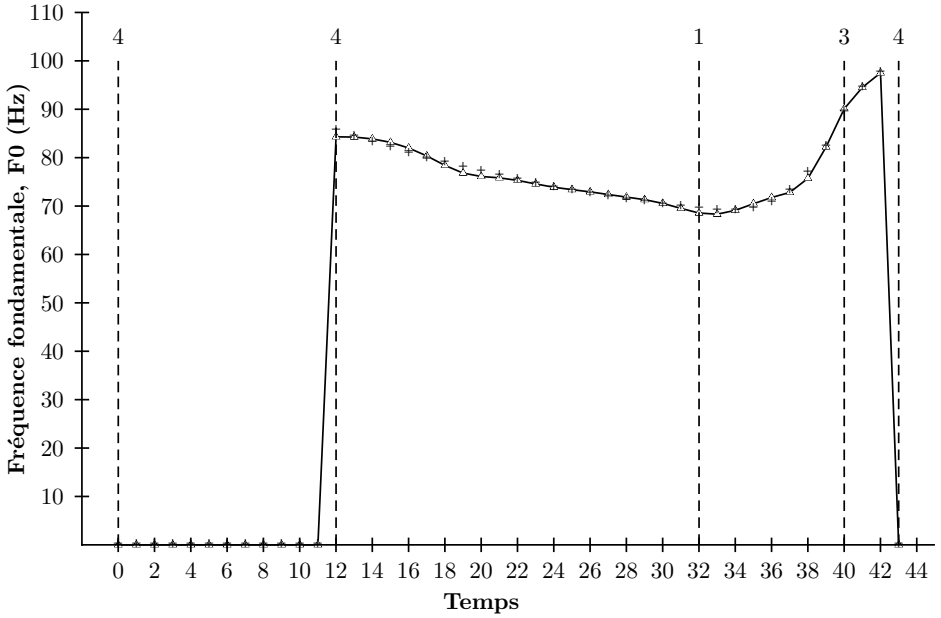


FIG. 5.6 – Estimation du modèle B-spline pour un contour mélodique. Le critère (a).3 est appliqué. Le modèle sélectionné (courbe pointillé) possède 8 nœuds internes. La courbe pleine représente le contour original. Les nœuds estimés ainsi que leur multiplicité sont positionnés sur la courbe.

contour mélodique avec une erreur RMS de 0.56Hz et un nombre normalisé de d.d.l. de 0.45.

5.7.2 Relation entre la RMS et les degrés de liberté du modèle

Ces expériences préliminaires sont réalisées afin d'observer le comportement de l'erreur RMS en fonction du nombre de d.d.l. des modèles. La courbe obtenue doit permettre de mesurer l'impact du nombre de paramètres sur la qualité d'estimation. Le nombre normalisé de d.d.l. présenté varie de 0 à 1.

La figure 5.8 page 120 montre la relation entre l'erreur RMS exprimée en Hz (unité de mesure des courbes observées) et les degrés de liberté du modèle. Toutes les courbes du corpus sont prises en compte sur cette figure.

Lorsque le nombre de d.d.l. augmente, l'erreur RMS diminue. Ce résultat, bien que prévisible, permet de justifier la recherche d'un compromis entre la précision du modèle et sa complexité. Notamment, on peut remarquer un changement de pente de la courbe avec un degré de liberté normalisé moyen voisin de 0.65. Cette valeur d'abscisse nous donne une erreur RMS moyenne correspondante proche de 1Hz. Pour tous les degrés de

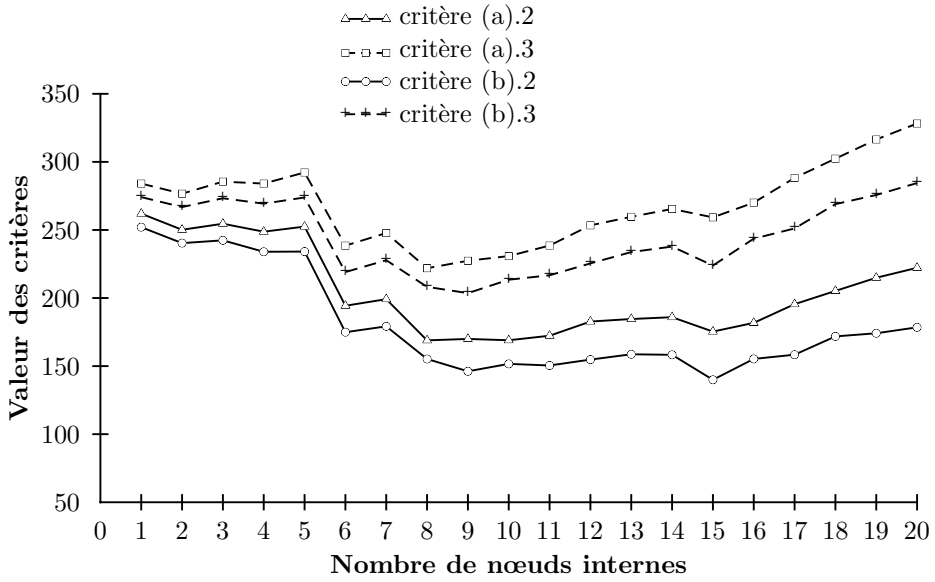


FIG. 5.7 – Estimation du modèle B-spline d'un contour mélodique. On applique les différents critères MDL (voir description complète 5.4.4 page 112).

liberté évalués, l'intervalle de confiance sur l'erreur RMS est assez fin.

Le but du critère MDL étant de déterminer un compromis entre erreur RMS et nombre de d.d.l., un critère satisfaisant devrait estimer une erreur RMS moyenne et un nombre de d.d.l. moyen aux environs de la zone de changement de pente de la courbe.

5.7.3 Sensibilité des critères MDL par rapport à ε

Dans ce paragraphe, on s'intéresse à la sensibilité des deux critères proposés relativement à la précision de la courbe, ε. Les figures 5.9 page 121 et 5.10 page 122 représentent respectivement l'évolution des intervalles de confiance de l'erreur RMS, et du nombre normalisé de d.d.l., en fonction des valeurs ε. Les valeurs de RMS et de ε sont représentées selon une échelle logarithmique. On peut noter, pour $\varepsilon = 1$, que l'on retrouve le cas où les points de contrôle sont considérés comme des valeurs entières.

On propose, section 5.4.4 page 112, un critère MDL fonction de la précision de reconstruction de la courbe. La pénalité MDL est adaptée à chaque courbe et à la qualité de sa modélisation. Il est important de situer les critères MDL de précision variable par rapport à ceux de précision fixe. Pour cela, on étudie le comportement de l'erreur RMS en fonction d'un échantillonnage de valeurs ε fixes possibles (critères (a).1 et (b).1). Ainsi, on évalue l'influence de la précision souhaitée des paramètres sur l'erreur RMS et les d.d.l.

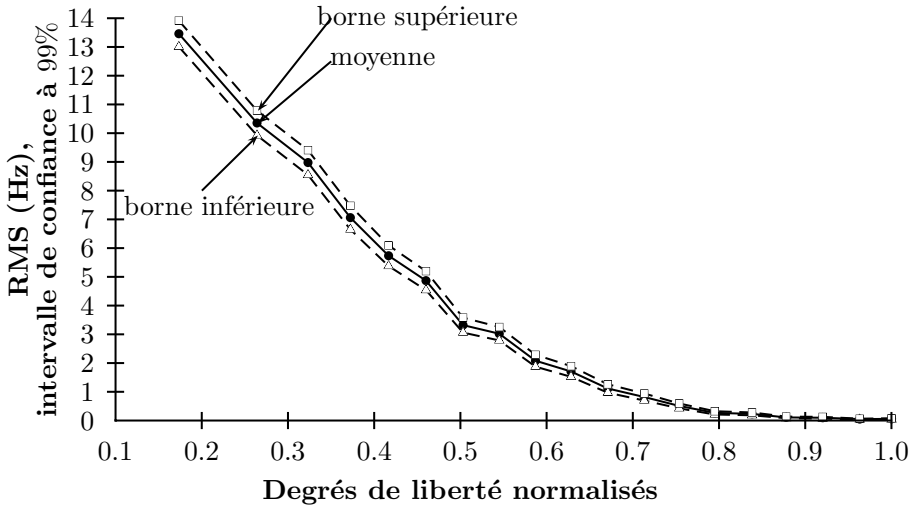


FIG. 5.8 – Évolution des intervalles de confiance à 99% pour l'erreur RMS en fonction du nombre normalisé de d.d.l.

5.7.3.1 Critère (a)

Moins le modèle reconstruit est précis (ε grand), plus le nombre de nœuds est important. En effet, lorsque ε augmente, le terme $-\log\varepsilon$ diminue, le critère sélectionne une valeur de l plus grande et l'erreur RMS diminue.

Sur la figure 5.9 page suivante, on peut noter une assez grande variation de l'erreur RMS. En effet, un facteur 10 sépare la valeur maximale de l'erreur RMS moyenne et son minimum. De même, on observe une grande variation du nombre moyen de d.d.l.

Ainsi, les variations de ce graphe soulignent l'influence de ε aussi bien sur l'erreur RMS que sur le nombre de d.d.l.

5.7.3.2 Critère (b)

Les observations précédentes sur le critère (a) restent valables pour (b). En effet, l'influence de ε est importante. Elle conditionne les performances du critère. Cependant, le critère (b) donne de moins bons résultats en terme d'erreur RMS moyenne. En effet, d'après la proposition 5.4.1 page 110, la longueur de description des points de contrôle est plus grande et pénalise donc plus le MDL, d'où une diminution du nombre de d.d.l. et une augmentation de l'erreur RMS. Cela reste cohérent avec l'évolution de l'erreur RMS moyenne en fonction du nombre de d.d.l. observée sur la figure 5.8.

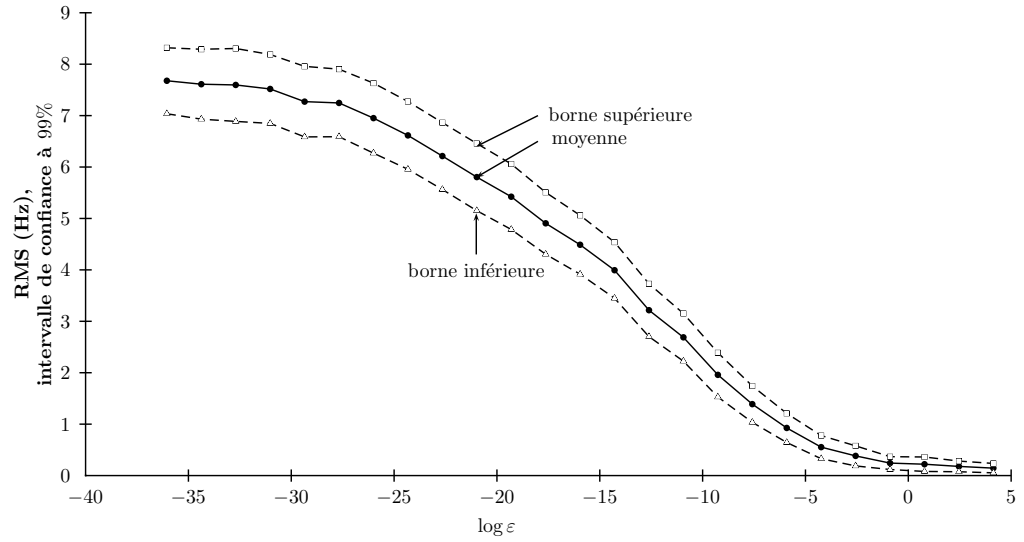


FIG. 5.9 – Évolution des intervalles de confiance à 99% pour l'erreur RMS moyenne en fonction de $\log \varepsilon$ selon le critère (a).

5.7.4 Analyse des critères MDL proposés

L'évaluation des critères proposés fait intervenir le calcul d'intervalles de confiance sur les valeurs moyennes de l'erreur RMS et du nombre de d.d.l. normalisé pour les modèles sélectionnés. En comparant ces résultats à ceux des expériences précédentes, on peut analyser la fiabilité des critères proposés. Contrairement aux expériences précédentes, on utilise ici un ε variable (critères (a).2, (a).3, (b).2 et (b).3).

Le tableau 5.2 page 123 résume les résultats des différents critères proposés en termes d'erreur RMS et de nombre normalisé de d.d.l. Les résultats présentés sont des intervalles de confiance à 99%.

Ces résultats permettent de distinguer deux types de compromis. Tout d'abord, le critère (a).2 sélectionne une erreur RMS moyenne de 0.66Hz avec un nombre normalisé de d.d.l. moyen de 0.603, tandis que le critère (b).2 donne respectivement 1.43Hz et 0.55. Ce dernier utilisant une longueur de description pour les points de contrôle plus grande, il est plus pénalisant et les valeurs de l sélectionnées sont plus petites. Il en résulte une augmentation de l'erreur RMS moyenne.

Deux modes variables ont été testés pour chaque critère. L'utilisation de $\|\mathbf{y} - \mathbf{B}\hat{\mathbf{c}}\|_\infty$ permet d'améliorer légèrement les résultats de RMS moyenne pour chacun des critères. En effet, d'après (5.13), $RMS \leq \|\mathbf{y} - \mathbf{B}\hat{\mathbf{c}}\|_\infty$. Les critères (a).3 et (b).3 sont donc moins pénalisant et permettent de sélectionner un nombre supérieur de d.d.l. que (a).2 et (b).2. Si l'on privilégie l'erreur RMS, les critères (a).3 et (b).3 peuvent être qualifiés de

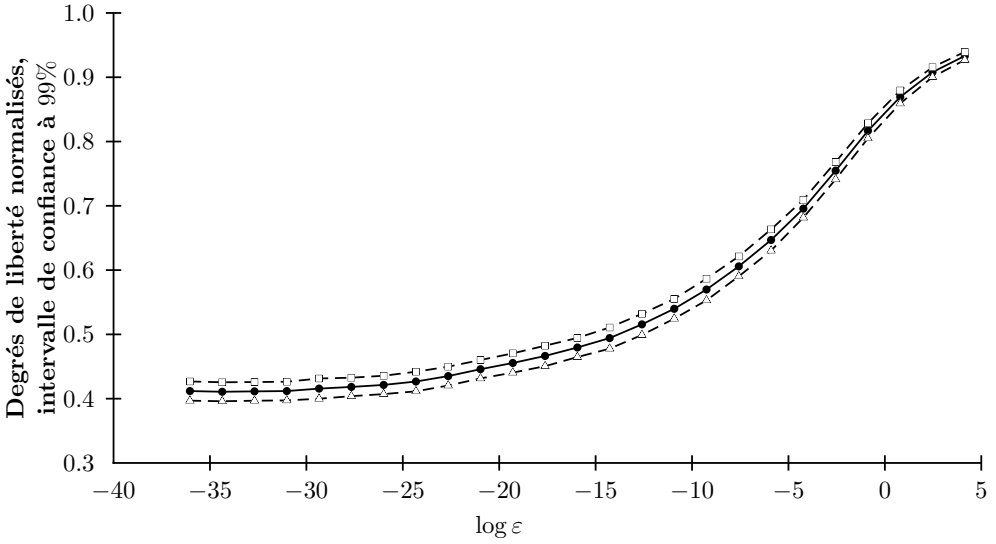


FIG. 5.10 – Évolution des intervalles de confiance à 99% pour le nombre normalisé de d.d.l. en fonction de $\log \varepsilon$ selon le critère (a).

meilleurs.

Par rapport au critère fixe, il est nécessaire de comparer, à erreurs RMS égales, les nombres de d.d.l. obtenus. De même une comparaison réciproque s'impose. Pour le critère (a).3, en choisissant la RMS moyenne égale à 0.42Hz, la figure 5.9 page précédente nous donne une valeur $\log \varepsilon$ de -3. En reportant cette valeur sur la figure 5.10, le nombre normalisé de d.d.l. est voisin de 0.75. Ainsi, pour une erreur RMS moyenne de 0.42Hz, le critère (a).3 estime un nombre normalisé de d.d.l. plus faible (autour de 0.63) que (a).1. Si l'on applique un raisonnement analogue à nombres de d.d.l. égaux, on peut conclure que le critère (a).3 est meilleur qu'un critère fixe. Ces conclusions sont également valables pour le critère (b).

Replaçons-nous par rapport à l'évolution générale de l'erreur RMS en fonction du nombre normalisé de d.d.l. (figure 5.8 page 120). Les critères (a).3 et (b).3 se positionnent dans la « zone d'inflexion » de la courbe aux points respectifs $(0.63, 0.42)$ et $(0.573, 1.02)$. Si on considère comme déterminant le critère de l'erreur RMS, on conclut alors que le critère (a) est meilleur que le critère (b). Cependant, sans cette hypothèse, les deux critères ne peuvent être différenciés, ils représentent deux formes de compromis.

TAB. 5.2 – Intervalles de confiance à 99% pour l'erreur RMS et le nombre normalisé de d.d.l.

Critères	d.d.l. norm.	RMS (Hz)
(a).2	0.603 ± 0.006	0.66 ± 0.09
(a).3	0.630 ± 0.006	0.42 ± 0.06
(b).2	0.550 ± 0.006	1.43 ± 0.15
(b).3	0.573 ± 0.007	1.02 ± 0.12

5.8 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle approche pour l'estimation de courbes ouvertes en utilisant un modèle B-spline. Nous l'avons traité dans un cadre général de positionnement libre des nœuds. Pour surmonter les principaux défauts d'une estimation au sens du maximum de vraisemblance, nous avons utilisé une stratégie de type recuit-simulé. La principale contribution concerne l'estimation du nombre de nœuds grâce à une méthodologie MDL qui tient compte de l'erreur de reconstruction. Les critères proposés reposent sur des bornes que nous avons établies à partir d'hypothèses sur les B-splines et qui limitent la précision des paramètres.

Nous avons montré expérimentalement sur des contours mélodiques du français qu'un modèle de régression B-spline donne de meilleurs résultats que le modèle spline : pour un taux de compression de 50%, l'erreur RMS est divisée par 4 pour les B-splines. Les courbes B-splines, plus générales que les splines, prennent en compte explicitement les irrégularités des contours mélodiques. Nous avons dans un deuxième temps cherché à valider expérimentalement les critères MDL proposés pour le modèle B-spline. Les expériences montrent que l'utilisation d'un critère variable permet d'améliorer de manière significative les résultats par rapport à un critère fixe. En outre, les critères variables permettent un taux de compression d'environ 40% tout en atteignant une erreur RMS de l'ordre de 1Hz. Les résultats obtenus sont encourageants. Nous envisageons de compléter cette étude par une comparaison avec des splines interpolantes tout en systématisant nos évaluations avec des modèles qui appliquent des stratégies déterministes lors de l'insertion des nœuds, (Cham et Cipolla, 1999).

La précision du modèle permet de regénérer des contours de très bonne qualité. Dans un contexte de synthèse de la parole, l'utilisation d'une stylisation telle que celle-ci peut-être envisagée de deux manière. La première est de conserver, pour chaque syllabe, le jeu de paramètres B-spline associé. Pour synthétiser une phrase, il est alors nécessaire de mettre en œuvre une stratégie de sélection d'unités prosodiques parmi l'ensemble

de tous les jeux de paramètres. La deuxième approche consiste à tenter de résumer la prosodie d'un locuteur en factorisant les jeux de paramètres ou les courbes similaires au sens d'un critère. Il s'agit donc d'effectuer une classification, un clustering des contours mélodiques du locuteur afin d'obtenir une représentation plus compacte de son espace prosodique.

Cette deuxième approche est celle que nous avons choisie puisqu'elle se situe dans la continuité de la stylisation. L'objectif est maintenant de modéliser des classes de contours mélodiques alors que le précédent était de modéliser des contours mélodiques. Sous cette hypothèse, le premier problème est que le modèle de stylisation choisi, conçu pour avoir une taille optimale pour chaque courbe, possède un nombre de paramètres variable. La comparaison de deux jeux de paramètres devient alors difficile. De plus, la durée des courbes est très variable ce qui empêche une comparaison directe des contours mélodiques pour évaluer leur proximité. Dans le chapitre suivant, nous proposons donc une méthode de classification qui repose sur l'utilisation de modèles de Markov apportant une solution possible à ce problème de vecteurs de longueur variable.

Chapitre 6

Apprentissage non supervisé de classes de contours mélodiques

6.1 Introduction

Un challenge important pour un système TTS serait de pouvoir offrir un large panel de modèles prosodiques de manière à diversifier les catalogues des voix. Actuellement, la majorité des systèmes de conversion de voix font appel à des corrections prosodiques globales (débit d'élocution et mélodie) (Gillett et King, 2003). Un enjeu important serait de pouvoir intégrer une transformation de modèles prosodiques, notamment de contours mélodiques, caractérisés de manière non supervisée à partir de quelques phrases enregistrées d'un locuteur cible.

Ce chapitre traite exclusivement du paramètre acoustique reconnu pour être un des principaux facteurs de perception de la prosodie, à savoir la fréquence fondamentale ou F_0. Comme nous l'avons vu au chapitre 3 page 53, une littérature importante traite de la stylisation de ces contours. Concernant la classification des contours mélodiques, on trouve peu de travaux portant sur une classification *non supervisée* du F_0. Il s'agirait de trouver un ensemble de formes mélodiques élémentaires à partir d'un ensemble de phrases pour lesquelles le signal de F_0 a été préalablement calculé. L'idée serait d'assembler en séquence les différentes classes de contours qui caractérisent l'espace mélodique du locuteur pour former une phrase mélodique complète (Mertens, 1989). Nous prenons pour hypothèse que l'élément atomique de notre espace de caractérisation mélodique est la syllabe. Il s'agit donc d'apprendre en aveugle un ensemble cohérent de classes de contours mélodiques à l'échelle de la syllabe.

Dans le chapitre 5 page 95, un modèle de stylisation des contours mélodiques est

présenté. Celui-ci repose sur le modèle B-spline qui permet de représenter de manière assez fine les contours mélodiques. Cependant, dans l'approche présentée, nous constatons que le nombre de paramètres du modèle B-spline varie d'un contour à un autre et une même forme mélodique peut se dérouler sur des supports temporels différents. Une classification de ces modèles pour un locuteur semble alors difficile. La difficulté majeure réside dans la prise en compte de la durée. Pour cette raison, nous avons choisi d'utiliser des chaînes de Markov cachées pour prendre en compte l'élasticité du support de représentation d'une forme élémentaire.

Dans ce chapitre, une méthodologie de classification non supervisée des contours mélodique est décrite. Cette méthodologie repose sur l'utilisation de modèles HMM en mode non supervisé appris dans (Lolive *et al.*, 2007b). Les détails mathématiques et algorithmiques pour l'estimation des paramètres d'un HMM ne sont pas présentés ici. Le lecteur intéressé pourra se reporter à l'article de Rabiner *et al.* (1989) pour une introduction aux modèles HMM. L'augmentation du nombre de classes est réalisé en appliquant une variante de la technique de Gaussian Splitting sur un ensemble de HMM. Plusieurs critères de sélection des classes à découper sont évalués (Lolive *et al.*, 2007a).

La structure des modèles HMM, ainsi que la technique mise en œuvre pour découper une classe de manière à en obtenir deux, sont présentées au paragraphe 6.2. Dans le paragraphe 6.3 page 128, l'algorithme d'apprentissage non supervisé permettant d'apprendre un ensemble de classes de contours mélodiques est présenté. La méthodologie expérimentale utilisée est présentée paragraphe 6.4 page 130 ainsi que les méthodes d'évaluation de la qualité des classes. Les résultats sont présentés et discutés au paragraphe 6.5 page 132.

6.2 Modélisation des classes par des HMM

6.2.1 Le modèle

L'objectif recherché est le partitionnement d'un ensemble de contours mélodiques de syllabes à l'aide de HMM (Rabiner, 1989; Bilmes, 1998). Dans notre approche, un HMM représente une classe et modélise des contours F_0, signal mono dimensionnel. La figure 6.1 page ci-contre présente la topologie des HMM utilisés. Leur construction repose sur la structure d'une syllabe. En effet, la linguistique nous apprend qu'une syllabe peut être décomposée en trois parties : l'attaque, le noyau et la coda. Cela nous conduit à utiliser un modèle possédant trois états émetteurs. De plus, comme l'attaque et la coda peuvent correspondre à des segments vides, le graphe de transition des HMM intègre des sauts évitant le premier ou dernier état émetteur.

La parole est un signal qui se déroule au fil du temps, il en est de même pour la structure d'une syllabe. Un HMM M_j est constitué de cinq états et ne possède donc pas de bouclage en arrière. Les états q_{0j} et q_{4j} sont respectivement les points d'entrée et de sortie du HMM. Ces deux états n'émettent pas de symbole et ont des temps de séjour nul. Quant aux états q_{ij}, pour i de 1 à 3, leurs émissions sont distribuées selon une loi gaussienne de moyenne μ_{ij} et de variance σ^2_{ij}.

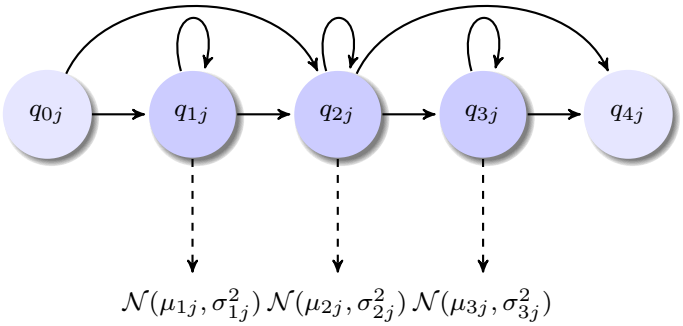


FIG. 6.1 – Structure d'un HMM M_j.

Pour chaque classe M_j de contours, on effectue l'apprentissage des paramètres du HMM associé par l'algorithme de Baum-Welsch en utilisant *HTK* (Woodland et Young, 1993). Les contours mélodiques sont étiquetés grâce à l'algorithme de Viterbi qui propose un décodage libre des modèles (mode non supervisé). La grammaire utilisée pour le décodage permet de respecter le caractère indivisible d'une syllabe. Elle ne possède pas de rebouclage, un seul HMM est choisi parmi l'ensemble des HMM $\mathcal{M}$. Ainsi lorsque l'on souhaite classer une nouvelle syllabe, on choisit le modèle le plus vraisemblable parmi tous les modèles possibles.

Ce travail se situe dans un cadre non supervisé, le nombre optimal de classes est inconnu a priori. Initialement, on considère que tous les contours mélodiques appartiennent à la même classe. Nous proposons une stratégie de classification hiérarchique descendante en augmentant le nombre de classes par division des classes existantes. La technique présentée au paragraphe 6.2.2 permet d'apporter une réponse à ce problème et permet également de fournir une initialisation des paramètres des HMM après division.

6.2.2 Gaussian Splitting

Après avoir décrit le modèle utilisé pour une classe, nous proposons maintenant une technique permettant de diviser une classe (un HMM) en deux classes distinctes.

Dans (Sankar, 1998; Rabiner *et al.*, 1989), on trouve deux cas d'application du

Gaussian Splitting. Le Gaussian Splitting est une technique qui permet d'augmenter le nombre de classes d'un système et de fournir une initialisation de l'algorithme d'apprentissage. Il consiste à perturber légèrement les paramètres des lois associées à chacun des états d'un HMM. Dans le cas présent, nous utilisons cette technique pour créer deux HMM à partir d'un seul.

Pour une classe du corpus d'apprentissage (un ensemble de syllabes), on note M_j le HMM dont les paramètres sont estimés selon un critère du maximum de vraisemblance. Le principe du splitting est que l'on peut améliorer la vraisemblance sur cet ensemble de données en utilisant deux HMM plutôt qu'un seul. Pour obtenir deux classes, on découpe alors le HMM M_j en perturbant uniquement les moyennes μ_{ij} des gaussiennes associées aux états q_{ij} (i, indice des états). Les moyennes sont perturbées suivant la direction de la déviation standard σ_{ij} de la gaussienne associée :

$$\mu_{ij}^{+} = \mu_{ij} + \epsilon * \sigma_{ij} \tag{6.1}$$

$$\mu_{ij}^{-} = \mu_{ij} - \epsilon * \sigma_{ij} \tag{6.2}$$

où ϵ est une constante fixée à 0.001 dans nos expériences. La spécialisation des deux nouveaux HMM est ensuite réalisée grâce à l'algorithme de Baum-Welch.

6.3 Apprentissage non supervisé des classes

L'apprentissage d'un ensemble de classes de contours mélodiques est réalisé dans un cadre non supervisé. Nous ne disposons pas de classes déjà établies à partir desquelles réaliser l'apprentissage des HMM. L'objectif consiste à regrouper des formes qui se ressemblent sous l'hypothèse de la modélisation proposée. La stratégie mise en place dans l'algorithme 1 page ci-contre permet d'obtenir un ensemble de classes à partir de trois éléments :

- l'ensemble des courbes à partitionner,
- une méthode de découpage des classes,
- et une mesure permettant de décider quelles classes doivent être divisées.

Ces trois éléments permettent de définir une méthode de classification que l'on peut qualifier de hiérarchique. En partant d'une classe globale, l'objectif est d'obtenir un partitionnement par divisions successives des classes existantes (ou un sous-ensemble) jusqu'à atteindre un état prédéfini. Cet état peut, par exemple, être la stabilisation d'un critère ou encore un nombre de classes prédéfini.

Entrées : *NbToSplit* le nombre de modèles HMM à diviser à chaque étape
Sorties : $\mathcal{M} = \{M_1, \ldots, M_p\}$

$\mathcal{M} = \{M_1\}$;
$e_{prev} = +Inf$;
$\epsilon = 1e^{-4}$;
converged = false;
répéter
    pour chaque *modèle HMM* $M_i \in \mathcal{M}$ **faire**
        apprendre M_i avec l'algorithme de Baum-Welch sur les données d'apprentissage
    fin
    Re-étiqueter toutes les syllabes du corpus de validation avec les nouveaux modèles $\mathcal{M}$ (Viterbi);
    Calculer l'erreur RMS moyenne e_{cur} entre chaque syllabe et le modèle de sa classe;
    si $e_{prev} - e_{cur} < \epsilon$ **alors**
        converged = true;
    sinon
        Diviser $\mathcal{M}$ en deux ensembles de HMM $\mathcal{M}_1$ et $\mathcal{M}_2$ avec card($\mathcal{M}_1$) = *NbToSplit*;
        Diviser chaque HMM de $\mathcal{M}_1$ dans $\mathcal{M}_1^{new}$;
        Rassembler $\mathcal{M}_1^{new}$ et $\mathcal{M}_2$ pour former un nouvel ensemble $\mathcal{M}^{new}$;
        Re-étiqueter toutes les syllabes avec le nouvel ensemble de HMM $\mathcal{M}^{new}$;
        $\mathcal{M} = \mathcal{M}^{new}$;
        $e_{prev} = e_{cur}$;
    fin
jusqu'à *converged* = *true* ;

Algorithme 1 : Algorithme non supervisé pour l'apprentissage des classes de contours mélodiques

L'algorithme 1, qui permet de remplir cet objectif, débute en ne considérant qu'une seule classe à laquelle est associé un HMM. À chaque itération, on découpe un sous-ensemble des classes existantes pour former de nouvelles classes. Considérons qu'un certain nombre d'itérations se soient déroulées, on possède alors un ensemble $\mathcal{M}$ de HMM. Après apprentissage des modèles de $\mathcal{M}$, l'erreur RMS moyenne globale est calculée sur un corpus de validation. Au niveau d'un contour de F_0 de longueur d, le calcul de l'erreur RMS est effectué de la manière suivante :

– on calcule la séquence d'états optimale $(T_t) \in \{q_{1j}, q_{2j}, q_{3j}\}^d$ du HMM M_j associé

à la syllabe avec l'algorithme de Viterbi
- à chaque état T_t, on associe $\mu_{T_t j}$ la valeur moyenne de la gaussienne associée à l'état T_t du HMM M_j.
- l'erreur RMS (Root Mean Square) est alors calculée entre les observations de F_0 et cette séquence de valeurs moyennes :

$$RMS^2 = \frac{1}{d}\sum_{t=1}^{d}(F_0(t) - \mu_{T_t j})^2$$

La convergence de l'algorithme est alors évaluée en fonction de l'erreur RMS sur l'ensemble du corpus de validation : on considère que l'algorithme a convergé si l'erreur RMS moyenne est stable ou si elle augmente. Si l'algorithme n'a pas convergé à cette étape, on construit le sous-ensemble $\mathcal{M}_1$ constitué des $NbToSplit$ HMM correspondant aux classes qui ont la plus forte valeur pour un critère donné. Quatre critères sont évalués dans la partie expérimentale (voir paragraphe 6.4.3 page ci-contre). Ce sont ces HMM qui sont découpés en deux, afin d'obtenir des classes plus fines en terme d'erreur RMS. Le nombre de HMM à diviser, $NbToSplit$, est un paramètre de l'algorithme.

Une fois le nouvel ensemble de classes $\mathcal{M}^{new}$ obtenu après découpage de $\mathcal{M}_1$, l'algorithme de Viterbi est utilisé pour modifier les étiquettes des contours de F_0 du corpus d'apprentissage et les faire correspondre aux nouvelles classes. On peut alors apprendre les nouveaux HMM sur l'ensemble d'apprentissage dont les labels ont été modifiés. Le processus de Gaussian Splitting est répété jusqu'à convergence de l'algorithme. Lors de la phase de découpage, il est possible qu'un HMM ne capte pas un nombre suffisant de courbes. Dans ce cas, l'algorithme se poursuit sans découper ce modèle.

Dans ce cas précis, le fait de conserver les mêmes classes peut permettre d'affiner l'estimation des modèles à l'étape suivante. Les courbes se situant aux « marges » de certaines classes peuvent ainsi changer de classe. Dans la plupart des cas, ce comportement rend possible la poursuite l'algorithme et le découpage d'autres classes.

6.4 Méthodologie

6.4.1 Corpus de F_0

Les expériences sont réalisées sur un ensemble de syllabes choisies aléatoirement parmi un corpus d'environ 7000 phrases enregistrées. L'enregistrement a été réalisé dans un studio professionnel. Le signal acoustique a été annoté puis segmenté en unités acoustiques. La fréquence laryngienne moyenne, F_0, a été analysée de manière automa-

tique à l'aide de la fonction d'auto-corrélation. Ensuite, un algorithme a été appliqué à la chaîne d'unités phonétiques de manière à repérer chaque syllabe. Le corpus de syllabes sélectionnées est divisé en un corpus d'apprentissage (8000 syllabes) et un corpus de validation (3000 syllabes).

6.4.2 Préparation des données

La première étape est la conversion des valeurs de F_0 en cents. Le cent, représentant un centième de demi-ton, est une unité permettant d'établir un parallèle avec le fonctionnement de l'oreille. En effet, le cent suit une échelle logarithmique permettant de donner plus d'importance à certaines fréquences (ici celles proches de la fréquence de référence $F_0^{ref} = 110\text{Hz}$) par rapport aux fréquences beaucoup plus hautes. Le passage du Hertz au cent est exprimé dans l'équation (6.3).

$$F_0^{cent} = 1200 * \log_2 \left(\frac{F_0^{hertz}}{F_0^{ref}} \right) \tag{6.3}$$

La seconde étape est similaire au traitement réalisé dans (Yamashita *et al.*, 2003). Elle réalise une interpolation linéaire des parties non voisées de la courbe de F_0 au niveau de la phrase. Cette interpolation résulte de l'hypothèse selon laquelle il existe un geste mélodique continu. Les valeurs de la fréquence fondamentale seraient alors masquées lors des parties non voisées. De plus, on effectue une régression linéaire sur les contours de F_0 obtenus afin de les lisser et de supprimer les variations micro-prosodiques.

6.4.3 Évaluation

L'objectif est ici d'établir des classes de contours mélodiques dans un cadre non supervisé. Il nous est alors impossible d'utiliser les méthodes habituelles d'évaluation de la qualité des classes. En effet, dans un cadre supervisé, il est possible de juger la pertinence des classes sur un corpus étiqueté à la main en comptabilisant des erreurs de reconnaissance.

Dans notre situation, la qualité d'une classe sera évaluée par rapport à la similarité des courbes rassemblées en terme de forme indépendamment de leur durée. Pour cela, nous utilisons un calcul d'erreur RMS entre une syllabe et la trajectoire optimale du HMM associé. On peut alors obtenir une erreur RMS moyenne pour la classe entière que l'on veut la plus basse possible et notamment en dessous du seuil de JND habituel pour le F_0 (autour de 4Hz). De plus, pour être en mesure de calculer l'erreur RMS et comparer les résultats à un seuil de JND pour le F_0, il est nécessaire de convertir le

contour mélodique et la trajectoire du HMM du cent vers le Hertz.

Dans la partie suivante, trois expériences sont présentées. La première montre un exemple de contour mélodique ainsi que la trajectoire du HMM représentant sa classe. L'objectif de cette expérience est de montrer comment une courbe et sa durée sont captées par le HMM. La deuxième expérience présente l'évolution de l'erreur RMS pour le critère CMSE (Cumulative MSE) en fonction du nombre de classes pour trois valeurs de *NbToSplit*. Quant à la troisième expérience, elle compare les quatre critères de sélection suivants :

1. mRMSE : pour chaque classe, l'erreur RMS moyenne est calculée, le tri des classes est ensuite effectué suivant cette valeur,
2. RMSEv : la variance de l'erreur RMS est calculée pour chaque classe. Ainsi les classes possédant une variance de RMS faible sont conservées tandis que les classes possédant une forte variance d'erreur sont scindées,
3. CMSE : l'erreur globale d'une classe est calculée en faisant la somme des valeurs d'erreur RMS au carré (Cumulative MSE),
4. CMSE_n : pour chaque classe, on calcule la MSE cumulée divisée par le nombre de courbes dans la classe. Dans ce cas, l'erreur globale d'une classe est répartie de façon égale entre toutes les courbes.

Ces critères sont comparés en termes d'erreur RMS et de nombre de HMM à chaque itération de l'algorithme.

6.5 Résultats et discussion

6.5.1 Exemple de contour mélodique

La figure 6.2 page suivante montre un exemple de contour mélodique et la trajectoire du HMM associé à sa classe. On peut observer la succession des états du HMM dans le temps. Dans cet exemple, le HMM reste dans l'état q_1 pendant les quatre premières observations. La gaussienne associée à cet état possède une moyenne d'environ 107Hz. Dans cet exemple, l'erreur RMS entre la courbe de F_0 et la trajectoire du HMM est de 1Hz. L'analyse de cet exemple permet de se rendre compte que les états du HMM reflètent la forme générale de la courbe. Le déroulement temporel (et donc la longueur de la courbe) est capté par les rebouclages au niveau de chaque état du HMM. Chaque HMM reflète alors une forme particulière indépendante de la durée permettant de modéliser des contours mélodiques de longueurs différentes mais de formes similaires.

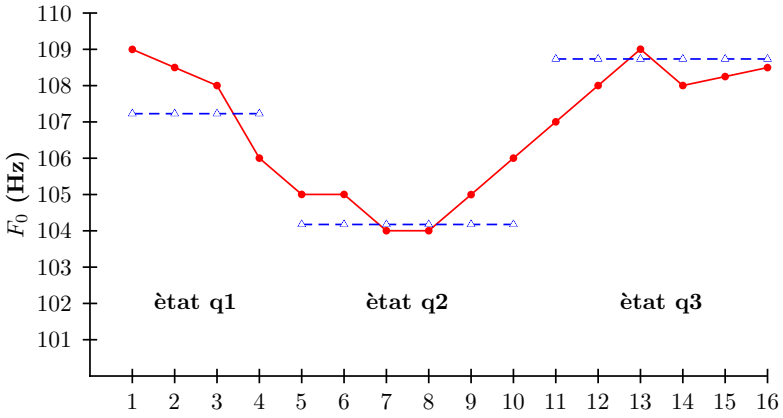


FIG. 6.2 – Exemple de contour de F_0 avec la trajectoire du HMM associé à sa classe. Le contour mélodique (ligne continue rouge) est superposé aux valeurs moyennes des gaussiennes associées aux états du HMM (tirets bleus). La séquence d'état du HMM pour cette syllabe est inscrite sous les courbes.

Un état, tel que représenté ici, ne permet pas de prendre en compte la pente d'un contour mélodique ce qui pourrait pourtant permettre de mieux suivre l'évolution du contour mélodique. En pratique, cela se traduirait par l'utilisation conjointe des valeurs de F_0 ainsi que des valeurs de sa dérivée. Cependant, la prise en compte des dérivées est plus complexe et conduit à des difficultés pour estimer la qualité des classes. Au lieu de tenir compte explicitement de la pente, on pourrait aussi augmenter le nombre d'états du modèle. Dans ce cas, on risque de se trouver face à des problèmes d'estimation des paramètres.

6.5.2 Résultats pour le critère CMSE

Dans les tableaux 6.1 page suivante et 6.2 page 135, les erreurs RMS moyennes en fonction du nombre de classes sont présentées. Cette expérience a été réalisée avec trois valeurs de *NbToSplit* différentes :

- *Split1* : la valeur de *NbToSplit* est 1, on divise au plus un seul HMM à chaque itération.
- *Split2* : la valeur de *NbToSplit* est 2, on divise au plus deux HMM à chaque itération.
- *SplitN* : tous les HMM sont divisés en deux à chaque itération. C'est une valeur particulière puisque variable à chaque itération.

Dans les trois colonnes du tableau 6.1 page suivante, on observe que l'erreur RMS moyenne (en Hertz), sur le corpus de validation, diminue avec l'augmentation du nombre

TAB. 6.1 – Erreur RMS moyenne (Hz) pour les trois variantes de *NbToSplit* sur le corpus de validation

No. de HMM	Split1	Split2	SplitN
1	11.44 ± 0.18	11.44 ± 0.18	11.44 ± 0.18
2	9.87 ± 0.16	9.87 ± 0.16	9.87 ± 0.16
4	9.23 ± 0.15	9.30 ± 0.15	9.30 ± 0.15
8	7.25 ± 0.15	7.87 ± 0.12	8.26 ± 0.14
16	5.48 ± 0.12	5.79 ± 0.11	6.74 ± 0.13
32	4.86 ± 0.11	4.82 ± 0.10	5.76 ± 0.12
64	4.56 ± 0.10	4.54 ± 0.11	5.15 ± 0.11
128	4.27 ± 0.10	4.25 ± 0.11	4.68 ± 0.11

de HMM. Cependant, l'erreur n'évolue pas de la même manière dans les trois cas. Pour *split1* et *split2*, le nombre de HMM que l'on divise à chaque itération est faible. Cela permet d'obtenir d'assez bons résultats (autour de 4Hz), par contre un plus grand nombre d'itérations est nécessaire pour obtenir 64 HMM que pour le cas *splitN*. Cela conduit à une convergence moins rapide. Par contre, une valeur plus élevée (cas *splitN*) permet une convergence plus rapide mais donne de moins bons résultats que les autres cas (erreur supérieure à 5Hz). De manière générale, on peut observer qu'un nombre relativement faible de classes (64 HMM) est nécessaire pour obtenir une erreur RMS moyenne voisine de 4Hz.

Diviser un nombre faible de HMM à chaque itération permet de concentrer l'effort de découpage sur les classes qui le nécessitent le plus au sens du critère choisi. La contrepartie est que pour obtenir un nombre prédéfini de classes, plus d'itérations sont nécessaires.

Dans le tableau 6.2 page suivante, on peut également observer les erreurs moyennes en fonction du nombre de classes exprimées en cent. L'évolution des erreurs est la même que dans le tableau 6.1. On peut noter que l'erreur commise à partir de 16 classes est inférieure au demi-ton (100 cents). De plus, pour les cas *split1* et *split2*, avec 128 classes, l'erreur est proche du quart de ton.

Les erreurs présentées dans ces deux tableaux permettent de voir que la distance entre une courbe et la trajectoire associée du HMM est faible. Cela signifie que les formes des courbes au sein d'une même classe sont assez proches. On en conclut qu'une classe reflète bien une forme de courbe particulière et que l'ensemble des classes forme une assez bonne partition de l'ensemble des contours mélodiques du corpus.

TAB. 6.2 – Erreur RMS moyenne (Cent) pour les trois variantes de *NbToSplit* sur le corpus de validation

No. de HMM	Split1	Split2	SplitN
1	165.50 ± 2.30	165.50 ± 2.30	165.50 ± 2.30
2	140.89 ± 2.01	140.89 ± 2.01	140.89 ± 2.01
4	130.96 ± 1.92	131.98 ± 1.90	131.98 ± 1.90
8	104.80 ± 2.05	113.86 ± 1.71	118.85 ± 1.92
16	79.81 ± 1.68	84.91 ± 1.58	98.26 ± 1.73
32	71.37 ± 1.56	70.53 ± 1.40	84.28 ± 1.69
64	66.97 ± 1.50	66.16 ± 1.53	75.63 ± 1.59
128	62.62 ± 1.48	62.03 ± 1.49	68.53 ± 1.50

6.5.3 Exemple de partitionnement avec 16 classes

Un partitionnement composé de 16 classes représente des contours assez variables comme le montre la figure 6.3 page suivante. Le partitionnement présenté possède une erreur RMS d'environ 5.6Hz ce qui est déjà assez faible. Le nombre de trajectoires de ce partitionnement peut être démultiplié en raison des noeuds facultatifs du HMM. En effet, même si dans l'exemple toutes les trajectoires sont représentées en prenant en compte le temps de séjour moyen dans chaque état, il est possible de « sauter » le premier et/ou le dernier état. Un HMM de la forme « montant/descendant » peut capter une courbe ayant une forme descendante en « sautant » le premier état.

Ce comportement peut se justifier par le fait que l'attaque ou la coda d'une syllabe peuvent être omises. La structure de la syllabe n'en est pas moins unique. Il s'agit simplement d'autoriser à un constituant d'une syllabe d'être non voisé alors que dans une autre syllabe il serait voisé. C'est par exemple le cas lorsque l'on remplace la consonne non voisée /p/ par la consonne voisée /b/.

L'étude du tableau 6.3 page 137 permet d'analyser plus finement le partitionnement obtenu. L'écart-type sur la durée des contours de F_0 au sein de chaque classe montre que les classes regroupent des contours de longueurs assez différentes. Par exemple, pour la classe (a), les syllabes ont une durée moyenne de 235.4 ms et les deux tiers d'entre elles ont une durée comprise entre 120.7 et 350.1 ms. Ces résultats témoignent d'une variabilité de longueur des contours importante au sein des classes. De ce point de vue, on peut penser que l'objectif du modèle HMM est rempli puisque c'est pour sa capacité à traiter les variations de durée qu'il a été choisi.

Les classes (d) et (e) présentent les erreurs RMS les plus faibles (resp. 3.70 et 3.35

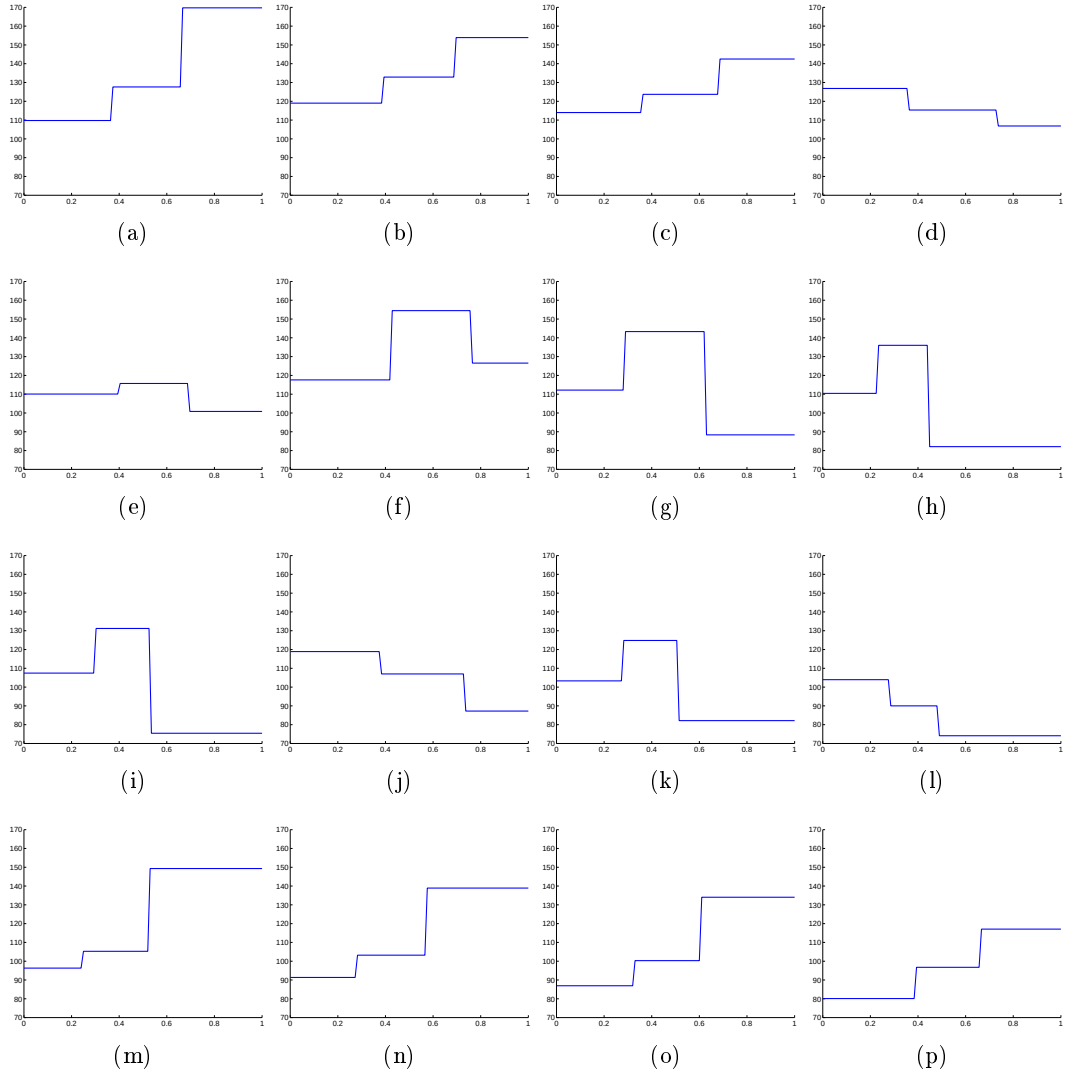


FIG. 6.3 – Trajectoires des HMM pour un partitionnement en 16 classes selon le critère CMSE. Toutes les courbes sont représentées sur une échelle commune en Hz. Pour chaque classe, la trajectoire du HMM est tracée en considérant le temps de séjour moyen dans chaque état du HMM et le support temporel est ensuite normalisé sur l'intervalle $[0, 1]$. On pourrait démultiplier le nombre de trajectoires en considérant différents temps de séjour dans chaque état. On peut noter la présence de trois grandes catégories de trajectoires : montant, descendant, montant/descendant. L'erreur RMS pour ce partitionnement est d'environ 5.6 Hz sur 3000 syllabes.

Hz). En observant les formes des trajectoires des HMM correspondant sur la figure 6.3, on se rend compte qu'il s'agit de courbes qui varient peu. À l'opposé, la classe (g) présente l'erreur la plus forte (12.01 Hz). La forme de cette classe laisse penser qu'elle

TAB. 6.3 – Exemple de partitionnement avec 16 classes et le critère CMSE. La deuxième colonne du tableau présente le nombre de syllabes captées par chaque HMM. On peut observer que ce nombre est assez variable. Les deux colonnes suivantes présentent la durée moyenne et l'écart-type de la durée des syllabes de chaque classe. Enfin, les deux dernières colonnes sont l'erreur RMS moyenne et l'écart-type de l'erreur RMS. L'erreur moyenne est très variable suivant les classes.

Classe	Nb. syl.	$\mu_{\text{durée}}$ (ms)	$\sigma_{\text{durée}}$	RMSE (Hz)	σ_{RMSE}
(a)	189	235.4	114.7	8.22	3.38
(b)	333	153.0	63.5	4.93	1.63
(c)	349	148.0	50.1	4.15	1.42
(d)	443	153.6	59.4	3.70	1.87
(e)	293	167.5	58.8	3.35	1.16
(f)	190	192.5	113.8	8.18	2.64
(g)	57	280.7	162.4	12.01	6.91
(h)	67	240.4	138.5	8.10	2.14
(i)	60	265.8	148.5	9.38	3.57
(j)	132	157.7	87.1	4.25	3.82
(k)	115	245.5	134.3	7.48	2.36
(l)	210	274.6	127.7	4.76	1.50
(m)	112	208.7	125.0	6.36	5.42
(n)	125	210.2	127.2	7.08	4.88
(o)	146	254.7	169.0	6.93	5.14
(p)	163	209.3	137.7	5.49	2.57

regroupe plutôt des courbes de F_0 dont les valeurs varient beaucoup.

De manière plus générale, il semble que plus les courbes varient et plus il est difficile de les modéliser avec le modèle proposé. On atteint ici une limite de ce modèle qui se trouve en difficulté du fait qu'un état du HMM ne représente pas un segment de droite quelconque. Intégrer la pente dans le modèle permettrait sans doute de lever ce problème.

À cette étape, en considérant que le critère de sélection est le CMSE comme dans cet exemple, la prochaine classe prioritaire pour être découpée est la classe (f). On peut noter ici que cette classe ne possède pas l'erreur RMS la plus forte. Avec un critère sur l'erreur RMS, la classe (g) aurait été sélectionnée à sa place. Le choix du critère de sélection des classes à découper est donc important afin de concentrer l'effort de classification sur les classes qui ont des chances de provoquer une amélioration globale plus importante.

6.5.4 Comportement du critère de sélection de classe

Évolution de l'erreur Concernant la sélection des classes à découper, nous avons testé quatre critères (voir paragraphe 5.5 page 112). La figure 6.4 montre l'évolution de l'erreur RMS pour chaque critère en considérant le cas *Split-1*. Nous pouvons tout d'abord remarquer que l'erreur diminue rapidement pendant les vingt premières itérations de l'algorithme. En effet, pendant ces premières itérations, le nombre de classes est faible et les données sont assez facilement séparables. En conséquence, l'ajout d'un nouveau HMM, c'est-à-dire incrémenter le nombre de classes de un, est très efficace tant que le nombre de HMM est faible. De plus, la différence entre les quatre critères n'est pas significative puisque les intervalles de confiance à 95% de l'erreur RMS moyenne ne sont pas séparables.

Concernant l'erreur RMS, le « meilleur » critère dans cette expérience est celui de la MSE cumulée (CMSE) qui donne une erreur proche de 4Hz.

Comme le nombre de classes est inconnu a priori, le nombre d'itérations est variable pour chaque critère. Dans le cas du critère mRMSE, il est à noter que le nombre d'itérations est relativement faible (inférieur à 60), tandis que dans les autres cas, il est supérieur à 150.

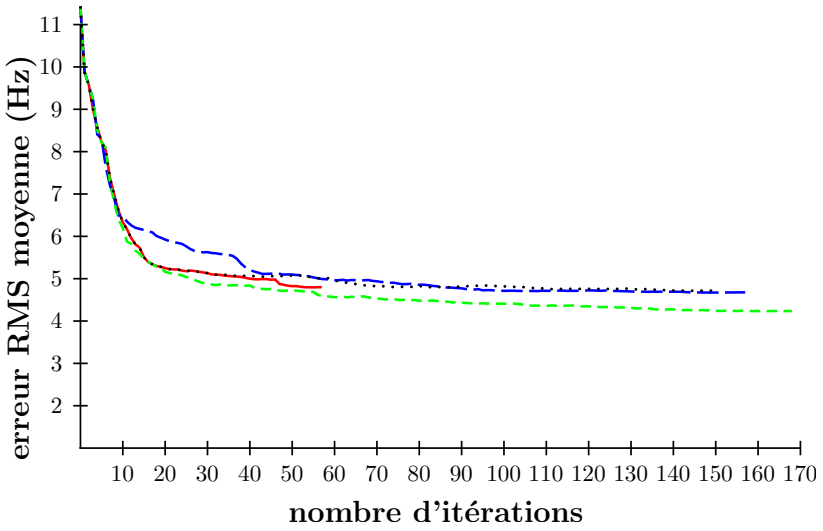


FIG. 6.4 – Évolution de l'erreur RMS pour les quatre critères de sélection dans le cas split-1 : mRMSE (ligne rouge), RMSEv (bleu, tirets long), CMSE (vert, tirets) and CMSE_n (noir, pointillés).

Évolution du nombre de classes La figure 6.5 page ci-contre représente l'évolution du nombre de HMM pour chaque critère. Comme dans la figure 6.4, le nombre d'ité-

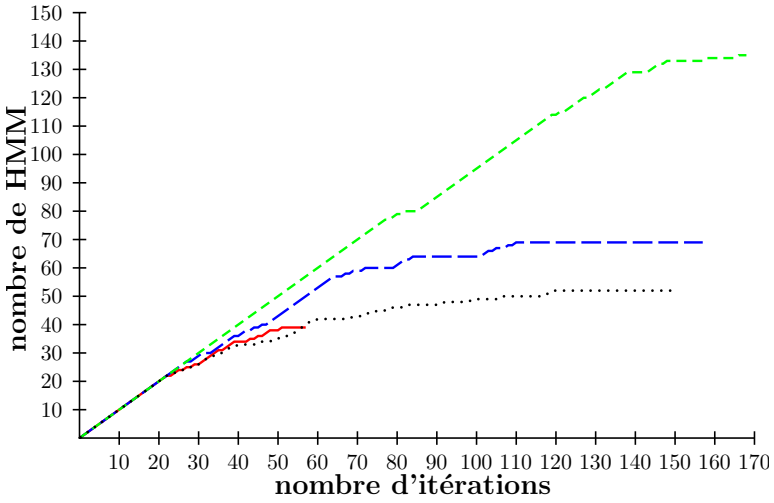


FIG. 6.5 – Évolution du nombre de HMM pour les quatre critères de sélection dans le cas split-1 : mRMSE (ligne rouge), RMSEv (bleu, tirets long), CMSE (vert, tirets) and CMSE_n (noir, pointillés).

rations est variable pour chaque critère. Concernant le nombre de HMM, l'évolution est assez différente pour les quatre critères. En effet, dans le cas CMSE, l'évolution du nombre de HMM est quasiment linéaire. Au contraire, dans le cas des critères RMSEv et CMSE_n, on peut observer des paliers au cours desquels le nombre de HMM est constant. Au cours de ces paliers, l'algorithme n'est pas en mesure de diviser un HMM. Cependant, quelques itérations avec un nombre constant de classes permettent à l'algorithme de recalculer les modèles et d'améliorer l'ensemble des classes. Ce traitement continue jusqu'à ce que l'erreur augmente ou bien se stabilise.

Lien entre erreur et nombre de classes La comparaison entre les figures 6.4 page précédente et 6.5 montre que lorsque le nombre de classes est élevé, les classes sont spécialisées et l'erreur RMS est faible. Dans ce cas, ajouter de nouvelles classes, permet d'obtenir des classes encore un peu plus spécialisées. La conséquence est que plus le nombre de classes est élevé, plus l'erreur est faible et stable.

Grâce à ces deux figures, nous pouvons réfléchir au critère le plus efficace. Tout d'abord, un critère peut être choisi en fonction de l'erreur RMS. Dans ce cas, le critère le plus pertinent des quatre proposés est le critère CMSE qui conduit à une erreur voisine de 4Hz. Ensuite, on peut s'intéresser au critère globalement le plus efficace, c'est-à-dire que l'on peut chercher le meilleur compromis entre l'erreur RMS et le nombre de classes. Ce point de vue fait du critère CMSE le moins efficace des quatre. En effet, le nombre de classes pour ce critère est le double des autres bien que l'erreur ne soit pas beaucoup

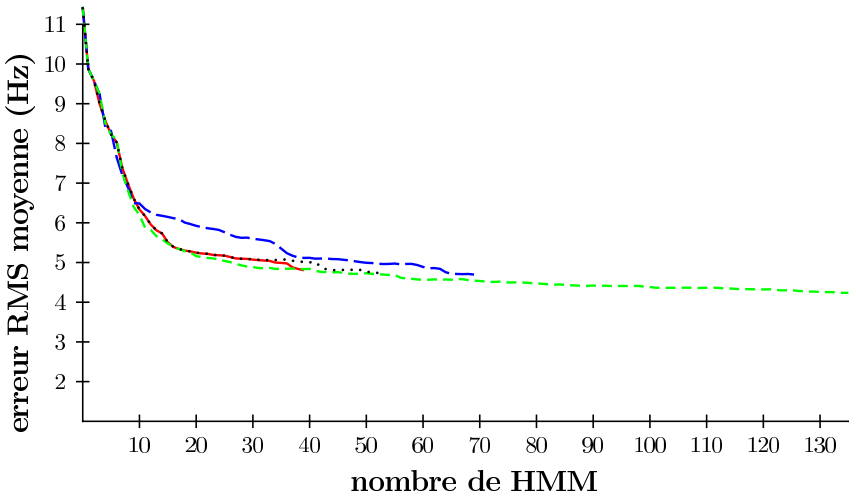


FIG. 6.6 – Évolution de l'erreur RMS moyenne en fonction du nombre de HMM pour les quatre critères de sélection dans le cas split-1 : mRMSE (ligne rouge), RMSEv (bleu, tirets long), CMSE (vert, tirets) and CMSE_n (noir, pointillés).

plus faible.

En dépit de ce fait, le critère CMSE possède des propriétés intéressantes que les autres ne possèdent pas :

- En effet, les critères mRMSE et CMSE_n sont des valeurs moyennes par rapport au nombre de contours de F_0 dans chaque classe. L'erreur est donc répartie de manière équitable sur les contours qui composent une classe. En conséquence, au sein d'une même classe, les courbes pour lesquelles la modélisation est de mauvaise qualité sont « masquées » par les courbes correctement modélisées.
- En ce qui concerne le critère RMSEv, une classe peut avoir une variance faible et pourtant une erreur RMS élevée. Cela rend ce critère inefficace par rapport au but recherché et l'erreur RMS ainsi que le nombre de HMM se stabilisent rapidement.
- Enfin, comme l'erreur globale pour le critère CMSE n'est pas divisée par le nombre de courbes captées par chaque classe, si une courbe possède une erreur élevée alors sa classe sera sûrement divisée. Également, ce critère est bien cohérent avec l'objectif d'une erreur RMS optimale par opposition au critère RMSEv.

Cette analyse est renforcée par les résultats de la figure 6.6. En effet, cette figure montre, pour le critère CMSE, une diminution très lente de l'erreur RMS moyenne en fonction du nombre de HMM. De plus, on observe que les résultats du critère RMSEv se situent au-dessus des trois autres. Enfin, les trois critères mRMSE, CMSE et CMSE_n se distinguent principalement par le nombre de HMM final lorsque l'algorithme converge.

6.6 Conclusion

Dans ce chapitre, une méthodologie d'apprentissage non supervisé de classes de contours mélodiques a été mise en œuvre en s'appuyant sur une modélisation par HMM. Les résultats montrent une assez bonne précision des classes. L'erreur RMS est voisine du seuil habituel de JND pour le F_0 de 4Hz. La modélisation par HMM permet en outre de rassembler des courbes de même forme au sein d'une classe et ceci indépendamment de la contrainte de durée des courbes.

Les expériences développées dans ce chapitre s'appuient sur des contours mélodiques à l'échelle des syllabes. On pourrait travailler sur d'autres supports que la syllabe, en utilisant des techniques de segmentation automatique du signal de parole. Le choix de la syllabe présente l'avantage d'offrir une unité présentant une structure bien définie ce qui facilite la construction d'un HMM.

Nous avons pu observer que le nombre de classes découpées à chaque itération influe sur le résultat ainsi que sur la rapidité du calcul et ceci de manière significative. Quatre critères de sélection des classes ont été évalués. Au regard des résultats, le critère le plus pertinent semble être le critère CMSE. Cependant, une meilleure stratégie serait peut-être de combiner plusieurs critères en utilisant un critère reposant sur l'erreur et un autre sur la variance de l'erreur.

Concernant le modèle choisi, intégrer d'autres informations au sein du HMM pourrait rendre plus efficace cette approche. En effet, prendre en compte la dérivée du F_0 dans le modèle permettrait de mieux capter les contours de F_0 qui ont une forte variation. Néanmoins, une telle extension engendre des problèmes d'évaluation de la classification.

Une application en conversion de voix est possible. En effet, disposant d'un ensemble de classes de contours mélodiques pour deux locuteurs, il faudrait alors trouver une fonction de conversion permettant de mettre en correspondance les classes d'un locuteur source avec celles d'un locuteur cible. S'inspirer des méthodes et outils mis en œuvre pour la transformation du timbre peut être bénéfique. À l'heure actuelle, les travaux concernant l'utilisation du partitionnement de l'espace prosodique par des HMM pour la transformation de la prosodie n'a pas encore abouti à des résultats probants.

Également, ce système fournit en sortie un jeu d'étiquettes correspondant à des motifs F_0. Ces étiquettes pourrait être utilisées dans un système de synthèse de parole pour l'enrichir et diversifier les voix de synthèse possibles. L'approche classique fondée sur les arbres de régression pourrait être mise en place pour prédire la classe du contour de F_0 à partir de données linguistiques. Le temps de séjour dans chaque état pourrait également être prédit par un modèle de durée.

Chapitre 7

Transformation de la prosodie par adaptation de GMM

7.1 Introduction

L'objectif d'un système de transformation de la parole est de modifier les phrases d'un locuteur source pour qu'elles soient perçues comme si elles avaient été prononcées par un autre locuteur, dont en particulier le locuteur cible. Ces dernières années, des domaines techniques tels que l'identification biométrique et la synthèse de parole à partir du texte, TTS, ont utilisé la méthodologie de transformation de la voix. En ce qui concerne l'identification biométrique, la transformation de la prosodie et la transformation de la voix de manière générale peuvent être utilisées pour mettre à l'épreuve les systèmes de vérification du locuteur ou d'identification du locuteur. Dans le champ de la synthèse de parole, la transformation de la voix peut avoir un impact important dans la mesure où un corpus d'unités de paroles, qui décrit une voix, associé à un ensemble de fonctions de transformation peut se substituer à l'approche classique selon laquelle chaque voix nécessite l'enregistrement d'un nouveau corpus d'unités acoustiques au complet.

Un système de transformation de la voix doit satisfaire deux principales exigences : la transformation des caractéristiques acoustiques segmentales et supra-segmentales. Dans ce chapitre, nous focalisons notre attention sur la transformation de la prosodie et plus particulièrement de la durée et de la fréquence fondamentale, F_0. De manière habituelle, un tel système de transformation peut être décomposé en trois étapes : stylisation, classification et transformation. Dans la littérature, nous avons vu au chapitre 4 page 73 que de nombreux travaux récents traitent de la transformation de la prosodie et plus

particulièrement du F_0 (Inanoglu, 2003; Gillett et King, 2003). Une approche classique consiste à modifier le F_0 en appliquant une transformation linéaire ou polynomiale qui repose sur des paramètres globaux des voix source et cible (Chappell et Hansen, 1998; Ceyssens *et al.*, 2002). D'autres approches décomposent ce problème complexe de transformation en sous-problèmes par un partitionnement de l'espace du F_0, comme cela est proposé par exemple par la solution *codebook* (Chappell et Hansen, 1998; Helander et Nurminen, 2007).

Avec les systèmes de conversion de voix classiques, il s'avère nécessaire de disposer, pour chaque phrase, d'un exemplaire prononcé par le locuteur source et d'un autre prononcé par le locuteur cible. En conséquence, deux corpus parallèles doivent être utilisés, ce qui constitue une hypothèse restrictive, pas toujours applicable suivant l'application visée. Relâcher cette contrainte permettrait de rendre plus souple la conception d'applications. Une réponse possible à ce problème, dans le cas de modèles paramétriques, peut être l'adaptation de modèles via une méthodologie d'adaptation au locuteur comme la MLLR, Maximum Likelihood Linear Regression, (Leggetter et Woodland, 1995). En synthèse de parole, Tamura *et al.* (2001) modélisent de manière conjointe les coefficients cepstraux sur une échelle mel, le F_0 et la durée en utilisant des MSD-HMM. Leur objectif est d'obtenir des modèles de phonèmes dépendant du locuteur par une adaptation MLLR de modèles indépendants du locuteur.

Dans ce chapitre, nous proposons une méthodologie pour la transformation de la prosodie applicable avec des corpus non parallèles (Lolive *et al.*, 2008a). L'information prosodique est considérée au niveau de la syllabe. L'idée sous-jacente est qu'une phrase mélodique peut être décomposée en unités de plus petite taille, qui, une fois mises bout à bout, permettent de reconstruire une phrase mélodique complète (Mertens, 1989). Considérant cette hypothèse, la conversion de prosodie peut être réalisée sur la base de séquences de syllabes transformées. Pour une syllabe, la durée et le F_0 sont représentés par un vecteur de taille fixe. L'espace mélodique d'un locuteur est représenté par un GMM, Gaussian Mixture Model. Ensuite, l'adaptation des paramètres du GMM source à l'aide des données cibles est mise en œuvre par l'application de la méthodologie MLLR. Une fonction de transformation des vecteurs prosodiques reposant sur les paramètres du GMM adapté est également proposée. Un vecteur transformé est calculé comme la somme pondérée des centroïdes du GMM adapté pour le locuteur source. Les coefficients de pondération sont les probabilités *a posteriori* des composantes du GMM en supposant connu le vecteur source observé. Cette approche a déjà été proposée pour la conversion du spectre et a montré des performances supérieures à l'approche par *mapping codebook* (Stylianou *et al.*, 1998).

Dans une première partie, nous présentons le modèle utilisé pour styliser la durée et le F_0. Ensuite, la modélisation GMM est décrite ainsi que la méthodologie d'adaptation et la fonction de transformation. La méthodologie expérimentale est ensuite présentée. Ce chapitre se termine par la présentation des résultats accompagnée d'une discussion de cette méthodologie.

7.2 Pré-traitement des données

7.2.1 Interpolation et lissage

Les contours de F_0 au niveau de la phrase sont pré-traités d'une manière semblable à celle proposée dans (Yamashita *et al.*, 2003). Tout d'abord, une interpolation est réalisée pour éliminer les parties non voisées du contour de F_0. Cette interpolation suit l'hypothèse selon laquelle un geste mélodique continu existe, la valeur de fréquence fondamentale serait alors masquée pendant les périodes non voisées. Les contours de F_0 obtenus après interpolation sont lissés à l'aide d'une spline cubique afin de supprimer les variations micro-mélodiques.

7.2.2 Représentation de la durée

La structure de la syllabe, déjà présentée au chapitre 6 page 125, est utilisée ici pour représenter la durée. Une syllabe peut être découpée selon trois parties : onset, noyau et coda. L'onset et la coda peuvent éventuellement être vides. À partir de cette structure, un vecteur $\mathbf{D} = (d_{\text{onset}}, d_{\text{nucleus}}, d_{\text{coda}})$ est construit pour caractériser la répartition de la durée au niveau de la syllabe. La durée de chaque partie est calculée comme un multiple de 10 ms.

7.2.3 Stylisation du F_0

Un contour de F_0, sur la base d'une syllabe, est représenté par un triplet $\mathbf{F} = (F_0^{10\%}, F_0^{50\%}, F_0^{90\%})$. Chaque coordonnée de ce vecteur correspond à la valeur de F_0 située, respectivement à 10%, 50% et 90% du support temporel de la syllabe. Ce processus revient à normaliser la durée des contours de F_0 en utilisant un support temporel unique pour tous les contours, comme cela est effectué dans (Reichel, 2007). La normalisation est une méthode simple permettant de supprimer les variations de longueur des contours et d'autoriser une comparaison directe de leur forme. Cependant, cette méthode présente l'inconvénient de considérer que le support temporel d'une syllabe est modifié de façon uniforme d'une syllabe à une autre, ce qui n'est pas forcément le cas.

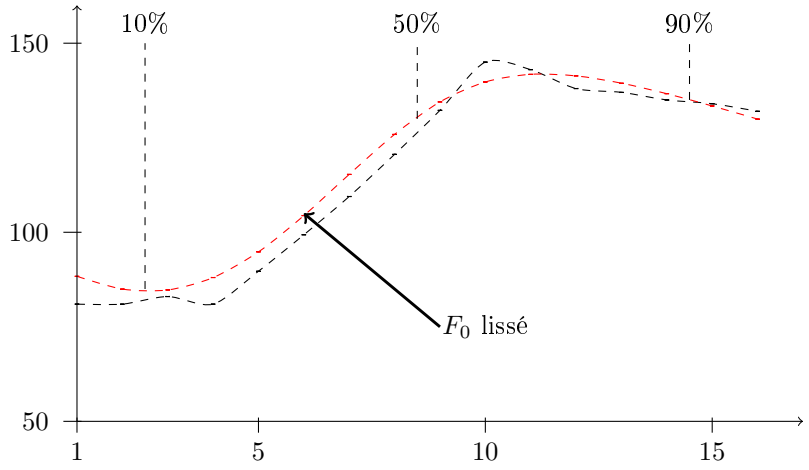


FIG. 7.1 – Stylisation de la prosodie d'une syllabe par un vecteur de dimension 6, $\mathbf{x} = (d_{\text{onset}}, d_{\text{nucleus}}, d_{\text{coda}}, F_0^{10\%}, F_0^{50\%}, F_0^{90\%})$. Le contour de F_0 est tout d'abord lissé et interpolé avant d'extraire les trois valeurs de F_0 située à 10%, 50% et 90% du support de la syllabe. Les valeurs de durée calculées pour le vecteur correspondent aux différentes parties de la syllabe.

7.2.4 Prosodie d'une syllabe

La prosodie d'une syllabe, ici F_0 et durée, est alors un vecteur de dimension six :

$$\mathbf{x} = (d_{\text{onset}}, d_{\text{nucleus}}, d_{\text{coda}}, F_0^{10\%}, F_0^{50\%}, F_0^{90\%}).$$

Ce vecteur, dont une illustration est présentée à la figure 7.1, permet la transformation conjointe du F_0 et de la durée et repose sur la structure d'une syllabe pour la durée. La représentation du F_0 est, quant à elle, arbitraire et en conséquence d'autres méthodes de stylisation peuvent être envisagées. Néanmoins, elles doivent respecter la contrainte de fournir pour chaque syllabe un jeu de paramètres de taille identique. En effet, ici, l'objectif de la stylisation est de pouvoir comparer les contours dans le domaine des paramètres à défaut de pouvoir effectuer une comparaison directe en raison des différences de longueurs des contours.

7.3 Transformation de la prosodie

L'approche que nous proposons offre une alternative à la méthodologie qui consiste à construire des corpus parallèles puis à aligner les phrases source et cible par DTW. L'objectif de cette approche est de transformer la prosodie entre un locuteur source et

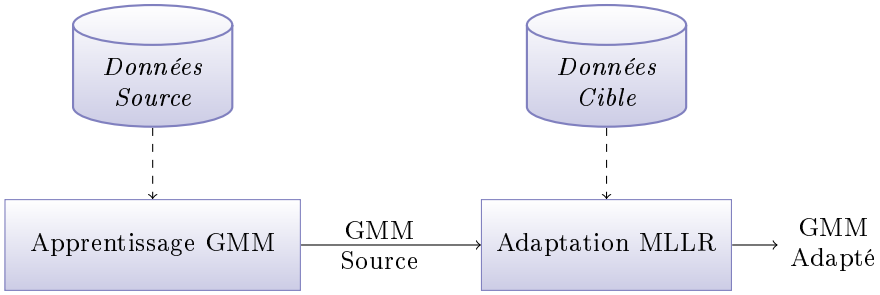


FIG. 7.2 – Architecture du système de transformation avec adaptation du GMM source aux données cible.

un locuteur cible sans utiliser de corpus parallèles, ce qui constitue un point fort de cette thèse. Des GMM sont utilisés pour modéliser les vecteurs source et cible, et la fonction de transformation repose sur les paramètres du GMM source adaptés aux données de la cible.

La figure 7.2 illustre cette méthodologie. La première étape consiste à apprendre un GMM sur les données issues du locuteur source et ensuite d'adapter les paramètres de ce GMM en utilisant les données du locuteur cible. La seconde étape du système de conversion est l'utilisation du modèle adapté pour transformer la durée et le F_0.

7.3.1 Modélisation par mélange de lois gaussiennes, GMM

Pour un locuteur, on considère l'ensemble $\mathbf{X}$ des vecteurs $\mathbf{x}$ qui représentent la prosodie de chaque syllabe. Un GMM $\mathcal{M}_{\mathbf{X}}$ à M gaussiennes est choisi pour modéliser l'ensemble $\mathbf{X}$ et sa distribution de probabilité est donnée par :

$$P(\mathbf{x}|\mathbf{\Theta}) \quad = \quad \sum_{m=1}^{M} \alpha_m P(\mathbf{x}|\theta_m)$$

où $\mathbf{\Theta} = (\alpha_1, \ldots, \alpha_m, \theta_1, \ldots, \theta_M)$ est le vecteur des paramètres, α_m est le coefficient de mélange associé à la m-ème gaussienne de paramètres $\theta_m = (\mu_m, \Sigma_m)$ et de distribution $P(\mathbf{x}|\theta_m)$.

L'algorithme EM, mis en œuvre pour estimer les paramètres du GMM, est un algorithme itératif dont l'objectif est de maximiser la vraisemblance conjointe des données et du modèle, (Bilmes, 1998).

$$LL(\Theta; \mathbf{X}) \; = \; \sum_{n=1}^{N} \log \left(\sum_{m=1}^{M} P(\mathbf{x}_n|C_n = m, \Theta) P(C_n = m|\Theta) \right) \tag{7.1}$$

où C_n est une variable aléatoire discrète cachée représentant la classe de $\mathbf{x}_n$ qui prend ses valeurs dans l'intervalle $1, \ldots, M$.

À partir de l'équation de la log-vraisemblance (7.1), on dérive l'algorithme d'estimation des paramètres d'un GMM présenté dans le cadre algorithme 2. Cet algorithme repose sur les deux étapes d'estimation et de maximisation. La première consiste à estimer les statistiques nécessaires à la ré-estimation des paramètres dans la deuxième. L'algorithme s'arrête lorsqu'il a convergé vers un maximum local de la fonction de log-vraisemblance. L'initialisation des moyennes des gaussiennes du GMM est réalisée en appliquant une quantification vectorielle sur les données.

Entrées : $\mathbf{X} = \{\mathbf{x}_n, n = 1, \ldots, N\}$
Sorties : $\mathcal{M}_{\mathbf{X}}$

`/* Initialisation */`

Initialisation SVQ de μ_m, pour $m = 1, \ldots, M$
Initialisation de Σ_m, pour $m = 1, \ldots, M$ avec covariance globale de $\mathbf{X}$
$\alpha_m = 1/M$, pour $m = 1, \ldots, M$
répéter
`/* Etape E */`
$\gamma_{nm} = \alpha_m P(\mathbf{x}_n | C_n = m, \theta) \; / \; \left(\sum_{l=1}^{M} \alpha_l P(\mathbf{x}_n | C_n = l, \theta) \right)$
`/* Etape M */`
$\mu_m = \left(\sum_{n=1}^{N} \gamma_{nm} \mathbf{x}_n \right) \; / \; \sum_{n=1}^{N} \gamma_{nm}$
$\Sigma_m = \left(\sum_{n=1}^{N} \gamma_{nm} (\mathbf{x}_n - \mu_m)(\mathbf{x}_n - \mu_m)^T \right) \; / \; \sum_{n=1}^{N} \gamma_{nm}$
$\alpha_m = \frac{1}{N} \sum_n \gamma_{nm}$
jusqu'à *converged* = *true* ;

Algorithme 2 : Algorithme EM pour l'apprentissage d'un GMM

7.3.2 Adaptation du GMM source

La méthode d'adaptation MLLR (Maximum Likelihood Linear Regression) est proposée dans (Leggetter et Woodland, 1995; Gales et Woodland, 1996). Considérons un GMM $\mathcal{M}_{\mathbf{X}}$ avec comme vecteur de paramètres $\mathbf{\Theta} = (\alpha, \mu, \mathbf{\Sigma})$ et appris sur l'ensemble de données $\mathbf{X}$ constitué de vecteurs de dimension d. L'objectif est d'adapter les paramètres du GMM $\mathcal{M}_{\mathbf{X}}$ à l'ensemble de données $\mathbf{Y}$ en calculant pour chaque composante

de $\mathcal{M}_{\mathbf{X}}$ une transformation de ses paramètres μ_m et $\mathbf{\Sigma}_m$ de manière à maximiser la vraisemblance du GMM adapté sur l'ensemble $\mathbf{Y}$:

$$\hat{\mu}_m = \widehat{\mathbf{W}}_m \xi_m \tag{7.2}$$

$$\hat{\mathbf{\Sigma}}_m = \mathbf{B}_m^T \hat{\mathbf{H}}_m \mathbf{B}_m \tag{7.3}$$

où $\widehat{\mathbf{H}}_m$ est la matrice de transformation de la variance et $\mathbf{B}_m$ est l'inverse du facteur de Choleski de $\mathbf{\Sigma}_m^{-1}$. $\widehat{\mathbf{W}}_m$ est la matrice d'adaptation de la moyenne de taille $d \times (d+1)$ et $\xi_m = [1\, \mu_{m1} \ \ldots \ \mu_{md}]$ est le vecteur des moyennes étendu. On a alors :

$$\widehat{\mathbf{W}}_m = [\hat{\mathbf{b}}_m \, \hat{\mathbf{A}}_m]$$

L'approche MLLR consiste à trouver un ensemble de matrices de transformation qui, lorsqu'elles sont appliquées aux moyennes et variances des gaussiennes, permettent de maximiser la vraisemblance des données d'adaptation. L'estimation de $\widehat{\mathbf{W}}_m$ et de $\widehat{\mathbf{H}}_m$ est réalisée en appliquant l'algorithme EM avec les données d'adaptation. Une version simplifiée de cet algorithme consiste à estimer une transformation uniquement sur les moyennes. Le facteur $\widehat{\mathbf{H}}_m$ dans l'équation (7.3) devient alors constant et égal à l'identité ce qui donne :

$$\hat{\mathbf{\Sigma}}_m = \mathbf{\Sigma}_m \tag{7.4}$$

L'algorithme utilisé pour réaliser l'adaptation est décrit dans le cadre algorithme 3 page suivante. L'initialisation est réalisée en utilisant le GMM $\mathcal{M}_{\mathbf{X}}$ tandis que l'adaptation est effectuée grâce à l'ensemble de données $\mathbf{Y}$ de la voix cible. À l'issue de l'algorithme, on dispose d'un ensemble de matrices de transformation approchant une transformation linéaire entre les moyennes et variances source et celles de la cible.

7.3.3 Transformation de la durée et du pitch

Une analogie avec la définition de la conversion de voix donnée par Stylianou (Stylianou *et al.*, 1998) est possible. Ici, la prosodie de la voix source doit être modifiée pour qu'elle ressemble à celle de la voix cible. En s'appuyant sur les travaux effectués dans le domaine de la conversion de voix au niveau segmental, et en particulier sur le travail de Stylianou, une fonction de transformation peut être définie de la manière suivante :

$$\mathbf{x}' = \mathcal{F}(\mathbf{x}, \mathcal{M}_{\mathbf{X}}) = \sum_m P(m|\mathbf{x}) \mu_m \tag{7.5}$$

Entrées : $\mathbf{Y} = \{\mathbf{y}_n, n = 1, \ldots, N\}$, $\mathcal{M}_{\mathbf{X}}$
Sorties : $\mathcal{M}_{\text{adapté}}$

répéter
  /* Etape E */
  $\gamma_{nm} = \alpha_m P(\mathbf{y}_n | C_n = m, \theta) \; / \; \left(\sum_{l=1}^{M} \alpha_l P(\mathbf{y}_n | C_n = l, \theta) \right)$
  /* Etape M */
  $A = \left(\sum_n \gamma_{nm} \mathbf{x}_n \right)^T \xi_m^T$
  $B = (\xi_m^T \xi_m)^{-}1$
  $\widehat{\mathbf{W}}_m = AB \; / \; \sum_n \gamma_{nm}$
  $\hat{\mu}_m = \widehat{\mathbf{W}}_m \xi_m$
  $C_m = cholesky(\mathbf{\Sigma}_m^{-1})$
  $B_m = C_m^{-1}$
  $\widehat{\mathbf{H}}_m = C_m^T \left[\sum_n \gamma_{nm} (\mathbf{x}_n - \hat{\mu}_m)(\mathbf{x}_n - \hat{\mu}_m)^T \right] C_m \; / \; \left(\sum_n \gamma_{nm} \right)$
  $\hat{\mathbf{\Sigma}}_m = B_m^T \widehat{\mathbf{H}}_m B_m$
jusqu'à $converged = true$;

Algorithme 3 : Algorithme EM pour l'adaptation d'un GMM.
Le GMM initial $\mathcal{M}_{\mathbf{X}}$ est utilisé à la première itération pour le calcul γ_{nm} à travers la probabilité $P(\mathbf{y}_n | C_n = m, \theta)$. Les coefficients de mélange du GMM initial ne sont pas mis à jour et adaptés. Garder ces derniers intacts est nécessaire pour la fonction de transformation proposée à l'équation (7.6).

où $P(m|\mathbf{x})$ est la probabilité que $\mathbf{x}$ appartienne à la classe m et μ_m est la moyenne de la gaussienne m du GMM $\mathcal{M}_{\mathbf{X}}$.

La forme de cette fonction, selon Stylianou, s'apparente à une transformation de type Quantification Vectorielle dans la mesure où seuls les centroïdes μ_m de chaque classe sont utilisés. Dans le cas présent, étant donné que les données sont supposées non parallèles, il nous est impossible d'introduire la covariance croisée des données source et cible.

Pour transformer un vecteur source de l'espace prosodique source vers l'espace prosodique cible, le GMM adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$ est utilisé. La fonction de transformation peut alors être écrite en utilisant la matrice de transformation calculée à l'aide de la MLLR :

$$\begin{aligned} \mathcal{F}(x, \widehat{\mathcal{M}_{\mathbf{X}}}) &= \sum_m P(m|\mathbf{x}) \hat{\mu}_m \\ &= \sum_m P(m|\mathbf{x}) (\widehat{\mathbf{W}}_m \xi_m) \end{aligned} \tag{7.6}$$

où ξ_m est le vecteur moyenne étendu de la classe m du GMM source $\mathcal{M}_{\mathbf{X}}$, $\widehat{\mathbf{W}}_m$ est la matrice d'adaptation pour cette gaussienne.

7.4 Protocole expérimental

7.4.1 Données

Deux séries d'expériences sont menées en utilisant pour chacune d'entre-elles un couple de voix :

- **Série d'expériences 1** : la voix source est un corpus de parole lue utilisé dans un système de synthèse de la parole à partir du texte. La voix cible est extraite du corpus de français Ester et correspond à un style journalistique. Le locuteur choisi est « Simon Tivolle ».
- **Série d'expériences 2** : deux corpus de données issus de BREF120 (Larnel *et al.*, 1991) sont utilisés. BREF120 est un corpus de français multi-locuteur. La sélection d'un couple de voix parmi toutes les combinaisons possibles a été réalisée grâce à un test subjectif de dissemblance de la prosodie dont l'objectif est de sélectionner deux voix ayant une prosodie très différente. Ce test a été effectué sur les 40 couples de voix du même genre possédant le plus de phrases communes. 9 auditeurs ont répondu à la question « Le style d'élocution (intonation, débit, etc.) de ces deux voix vous semble-t-il : dissemblables, ..., identiques » (sur une échelle de 0 à 4). Les réponses ont permis de retenir les voix de deux femmes avec un score de $0,43$: *JMF* et *JNF*, resp. la source et la cible.

Pour chaque corpus, une segmentation automatique en phones est réalisée (Charonnat *et al.*, 2008). La fréquence fondamentale moyenne, F_0 a été obtenue grâce à la méthode YIN (de Cheveigne et Kawahara, 2002). Chaque séquence phonétique est segmentée en syllabes. Les contours de F_0 sont interpolés, puis lissés avant de représenter la prosodie comme il est décrit au paragraphe 7.2 page 145.

Pour chaque voix, 5000 syllabes sont utilisées pour l'apprentissage et 1500 syllabes pour la validation. La transformation de la prosodie entre les deux voix sélectionnées n'est pas aisée. En effet, la voix source est de la parole lue avec une prosodie relativement régulière tandis que la voix cible contient une prosodie beaucoup plus diversifiée, très particulière au style journalistique.

7.4.2 Expériences

Les GMM utilisés pour cette expérience comportent 32 gaussiennes avec des matrices de covariance diagonales. L'adaptation et la transformation du F_0 et de la durée sont difficiles à évaluer. En effet, dans le cadre de la transformation de la prosodie avec des corpus non parallèles, une évaluation subjective est clairement difficile à moins de disposer d'un sous-corpus parallèle dédié exclusivement à l'évaluation. Dans notre cas, ne disposant pas de sous-corpus parallèles, nous proposons d'évaluer la méthodologie de manière objective par validation croisée en s'appuyant sur la log-vraisemblance entre données et modèles. Nous considérons les trois GMM source $\mathcal{M}_{\mathbf{X}}$, cible $\mathcal{M}_{\mathbf{Y}}$ et adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$, appris respectivement :

- avec des données sources,
- avec des données cibles,
- à partir de $\mathcal{M}_{\mathbf{X}}$ adapté grâce aux données cibles.

On considère également la transformation des données sources à l'aide du modèle adapté, décrite par (7.6). Afin d'évaluer la qualité de ces nouvelles données par rapport aux observations relatives au locuteur cible, il faut dans un premier temps étudier l'effet de la fonction de transformation (7.5) sur les données initiales.

Les deux séries d'expériences concernent l'étude de l'adaptation des paramètres d'un GMM et de son utilisation pour la transformation de la prosodie et diffèrent par le processus d'adaptation utilisé :

- **Série d'expériences 1** : adaptation de la moyenne uniquement (Lolive *et al.*, 2008a). Dans ce cas, les coefficients de mélange et les variances des gaussiennes du GMM source sont conservés.
- **Série d'expériences 2** : adaptation de la moyenne et de la variance (Lolive *et al.*, 2008b). Dans ce cas, seuls les coefficients de mélange restent intacts par rapport au GMM source.

7.5 Résultats et discussion

7.5.1 Séries d'expériences 1 : adaptation de la moyenne

L'architecture du système de transformation proposé repose tout d'abord sur une étape de stylisation et également une étape de modélisation de l'espace prosodique du locuteur. Pour se rendre compte de la pertinence et des limites de la stylisation par un vecteur qui possède six coordonnées ainsi que de la classification de ces vecteurs par un GMM, on peut s'intéresser à la figure 7.3 page suivante qui présente un exemple de

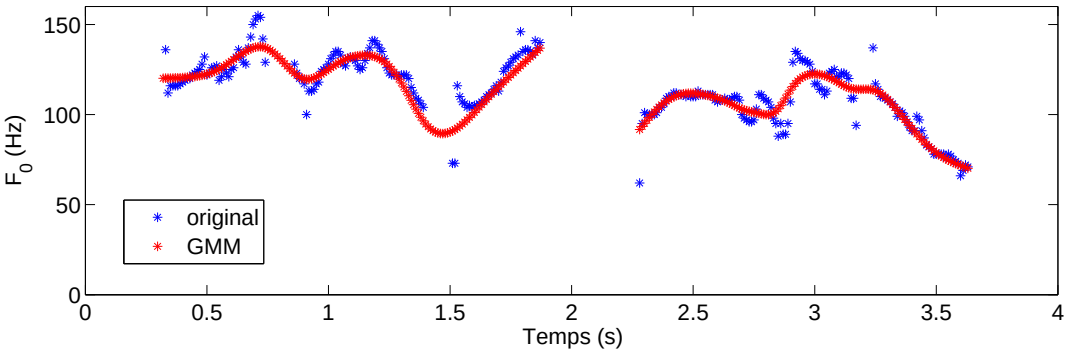


FIG. 7.3 – Génération du contour modélisé à partir des vecteurs d'observations au niveau de la syllabe et du GMM source. Cet exemple est réalisé sur une phrase issue du locuteur source, « Malgré ses nombreux appels, nous ne l'avons pas entendu. », dont le contour de F_0 est présenté en bleu. Le contour généré avec le GMM est présenté en rouge. L'erreur commise avec le processus de stylisation et classification complet sur cet exemple est de $6,2$Hz. L'équation (7.5) est utilisée pour obtenir le contour GMM.

génération d'un contour mélodique pour une phrase à partir des vecteurs obtenus par stylisation. Pour évaluer l'efficacité du modèle GMM, les vecteurs au niveau de la syllabe sont appliqués à l'équation (7.5). Le contour mélodique obtenu semble représenter l'évolution du contour original et on peut noter que l'erreur RMS sur cet exemple est de 6.2Hz. Néanmoins, il semble qu'un vecteur constitué de seulement trois points pour le F_0 ne permette pas de restituer finement le contour mélodique. Un vecteur de taille plus importante devrait permettre une meilleure stylisation des contours de F_0 et également restitution de ces contours.

Une alternative à cette méthode de stylisation en utilisant des courbes B-splines permettrait également de tenir compte de la durée et de la forme des contours mélodiques. Les courbes B-splines, que nous avons présentées au chapitre 5 page 95, sont caractérisées par un vecteur de nœuds et un vecteur de points de contrôle qui peuvent être considérés comme homogènes respectivement à la durée et aux valeurs de F_0 des contours mélodiques. En choisissant un nombre de paramètres fixe pour toutes les courbes, l'utilisation d'un tel modèle peut être envisagée pour styliser de manière plus fine les contours mélodiques.

Le tableau 7.1 page suivante permet d'observer les valeurs de vraisemblance pour l'apprentissage des GMM et l'adaptation du GMM source avec les données cible. On peut noter que pour l'apprentissage du GMM source sur les données de la voix source, les vraisemblances sur les ensembles d'apprentissage et de validation sont très proches. On peut faire le même constat pour le GMM cible et le GMM adapté sur les données

TAB. 7.1 – Log-vraisemblance pour $\mathcal{M}_{\mathbf{X}}$ et $\mathcal{M}_{\mathbf{Y}}$ ainsi que les résultats de l'adaptation de $\mathcal{M}_{\mathbf{X}}$ en utilisant $\mathbf{Y}$, intervalles de confiance à 95% sur les ensembles d'apprentissage.

	Apprentissage	Validation
$\mathcal{M}_{\mathbf{X}}$	$-18.00_{\pm 0.09}$	$-18.05_{\pm 0.19}$
$\mathcal{M}_{\mathbf{Y}}$	$-17.84_{\pm 0.10}$	$-17.97_{\pm 0.21}$
$\widehat{\mathcal{M}_{\mathbf{X}}}$	$-18.14_{\pm 0.10}$	$-18.28_{\pm 0.21}$

TAB. 7.2 – Log-vraisemblance pour les GMM source $\mathcal{M}_{\mathbf{X}}$, cible $\mathcal{M}_{\mathbf{Y}}$ et adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$ sur les données de validation source $\mathbf{X}$ et cible $\mathbf{Y}$.

	X	Y
$\mathcal{M}_{\mathbf{X}}$	$-18.05_{\pm 0.19}$	$-19.73_{\pm 0.23}$
$\mathcal{M}_{\mathbf{Y}}$	$-19.07_{\pm 0.19}$	$-17.97_{\pm 0.21}$
$\widehat{\mathcal{M}_{\mathbf{X}}}$	$-18.88_{\pm 0.22}$	$-18.28_{\pm 0.21}$

TAB. 7.3 – Comparaison des valeurs de vraisemblance pour les GMM $\mathcal{M}_{\mathbf{X}'}$, $\mathcal{M}_{\mathbf{Y}'}$ et $\mathcal{M}_{\mathbf{Z}'}$ les ensembles de données initiaux et transformés.

	$\mathbf{X}$	$\mathbf{Y}$	$\mathbf{X}'$	$\mathbf{Y}'$	$\mathbf{Z}'$
$\mathcal{M}_{\mathbf{X}'}$	$-28.04_{\pm 0.89}$	$-31.58_{\pm 0.99}$	$-5.12_{\pm 0.55}$	$-23.92_{\pm 0.38}$	$-23.18_{\pm 0.39}$
$\mathcal{M}_{\mathbf{Y}'}$	$-30.21_{\pm 1.04}$	$-29.02_{\pm 0.73}$	$-20.70_{\pm 0.59}$	$-8.89_{\pm 0.43}$	$-14.77_{\pm 0.33}$
$\mathcal{M}_{\mathbf{Z}'}$	$-27.43_{\pm 0.65}$	$-27.34_{\pm 0.38}$	$-20.05_{\pm 0.44}$	$-17.96_{\pm 0.31}$	$-4.93_{\pm 0.59}$

cible. Cela signifie en particulier que les modèles sont assez représentatifs des données et qu'il n'y a pas d'effet d'overfitting.

Le tableau 7.2, montre une comparaison des valeurs de log-vraisemblance pour les trois GMM source $\mathcal{M}_{\mathbf{X}}$, cible $\mathcal{M}_{\mathbf{Y}}$ et adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$ appris respectivement sur les ensembles de données source $\mathbf{X}$ et cible $\mathbf{Y}$. En analysant ce tableau ligne par ligne, on remarque que, pour $\mathcal{M}_{\mathbf{X}}$, la valeur de vraisemblance est meilleure sur l'ensemble $\mathbf{X}$ que sur l'ensemble $\mathbf{Y}$. Pour $\mathcal{M}_{\mathbf{Y}}$, le phénomène inverse se produit. Cela met en avant le fait que les ensembles $\mathbf{X}$ et $\mathbf{Y}$ possèdent des distributions différentes. De plus, le GMM $\widehat{\mathcal{M}_{\mathbf{X}}}$ possède une log-vraisemblance plus élevée sur l'ensemble de données cible $\mathbf{Y}$ que sur l'ensemble de données source $\mathbf{X}$. Ces résultats traduisent le fait que le GMM adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$ est plus proche de la distribution de $\mathbf{Y}$ que de celle de $\mathbf{X}$. L'adaptation a donc permis le déplacement de la distribution du GMM $\mathcal{M}_{\mathbf{X}}$ vers celle du GMM cible $\mathcal{M}_{\mathbf{Y}}$. On peut observer ce déplacement sur la figure 7.4 page suivante. Cette figure montre clairement que le GMM adapté est plus proche du GMM cible que du GMM source.

Soit $\mathbf{X}'$, $\mathbf{Y}'$, $\mathbf{Z}'$ trois nouveaux ensembles de données obtenus par transformation

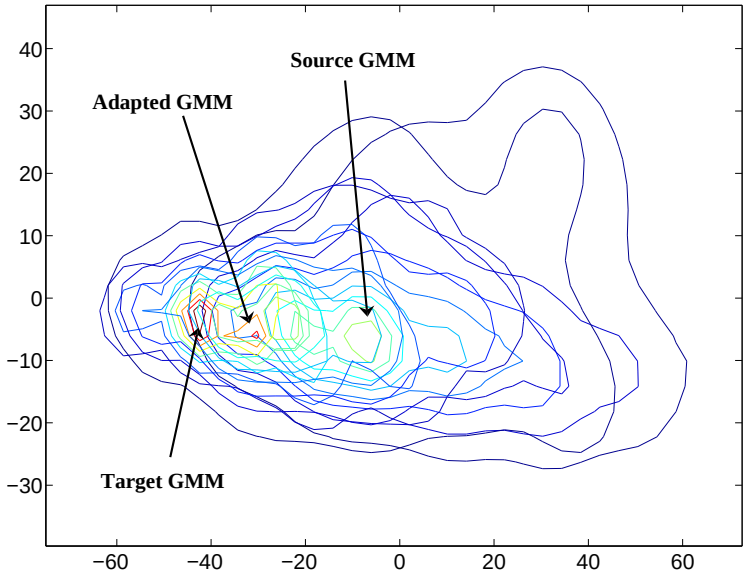


FIG. 7.4 – Projection des densités de probabilité des GMM source, cible et adapté. Cette projection est réalisée sur l'espace moyen des GMM source et cible en conservant les deux axes principaux à l'aide d'une PCA.

respective

- des données de la source par le GMM source $\mathcal{M}_{\mathbf{X}}$, en appliquant (7.5)
- des données de la cible par le GMM cible $\mathcal{M}_{\mathbf{Y}}$, par un calcul similaire
- des données de la source par le GMM adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$, en appliquant (7.6).

À partir de ces nouveaux ensembles de données, on apprend trois nouveaux GMM : $\mathcal{M}_{\mathbf{X}'}$, $\mathcal{M}_{\mathbf{Y}'}$, $\mathcal{M}_{\mathbf{Z}'}$, respectivement à partir de $\mathbf{X}'$, $\mathbf{Y}'$ et $\mathbf{Z}'$.

Intéressons-nous au comportement de la fonction de transformation grâce aux résultats présentés dans le tableau 7.3 page ci-contre. Les valeurs rapportées dans ce tableau concernent l'évaluation des trois GMM appris sur les données transformées $\mathbf{X}'$, $\mathbf{Y}'$, $\mathbf{Z}'$. Dans la première partie de ce tableau, on peut noter que les GMM $\mathcal{M}_{\mathbf{X}'}$, $\mathcal{M}_{\mathbf{Y}'}$, $\mathcal{M}_{\mathbf{Z}'}$ donnent de mauvais résultats sur les ensembles de données $\mathbf{X}$ et $\mathbf{Y}$. L'explication de ce phénomène est que la fonction de transformation a tendance à projeter les données autour des moyennes des gaussiennes. En effet, l'équation (7.5) montre que la valeur transformée est égale à la somme des moyennes des gaussiennes du GMM pondérées par la probabilité d'appartenance à la classe. Plus précisément, les données transformées se trouvent dans l'enveloppe convexe des moyennes des gaussiennes du GMM. La variance des données transformées est alors apportée uniquement par la probabilité d'appartenance à une classe ou une autre. De ce fait, les données transformées possèdent une

TAB. 7.4 – Comparaison des valeurs de log-vraisemblance pour les GMM $\mathcal{M}_{\mathbf{X}}$, $\mathcal{M}_{\mathbf{Y}}$ et $\widehat{\mathcal{M}_{\mathbf{X}}}$ sur les données de validation transformées.

	$\mathbf{X}'$	$\mathbf{Y}'$	$\mathbf{Z}'$
$\mathcal{M}_{\mathbf{X}}$	$-15.52_{\pm 0.15}$	$-16.87_{\pm 0.18}$	$-16.47_{\pm 0.17}$
$\mathcal{M}_{\mathbf{Y}}$	$-17.41_{\pm 0.19}$	$-15.46_{\pm 0.17}$	$-15.84_{\pm 0.16}$
$\widehat{\mathcal{M}_{\mathbf{X}}}$	$-16.85_{\pm 0.20}$	$-15.88_{\pm 0.16}$	$-15.49_{\pm 0.14}$

variance plus faible que les données originales.

La partie droite du tableau 7.3 page 154 confirme cette remarque en montrant des écarts très forts pour un GMM fixé sur les trois ensembles de données $\mathbf{X}'$, $\mathbf{Y}'$, $\mathbf{Z}'$. On peut noter que sur chaque ligne on retrouve le même ordre de classement des modèles que celui établi dans le tableau 7.4.

Pour pallier le manque de variabilité des données transformées, il serait intéressant d'adapter non seulement la moyenne des gaussiennes mais aussi leur variance et d'intégrer cette variance dans la fonction de transformation.

À l'aide de ces trois ensembles de données transformées et homogènes (car transformées par une fonction de même forme), nous allons pouvoir préciser le comportement de l'adaptation. Dans le tableau 7.4, on peut noter que les résultats obtenus sur les trois ensembles transformés sont ordonnés de la même manière pour les GMM $\mathcal{M}_{\mathbf{Y}}$ et $\widehat{\mathcal{M}_{\mathbf{X}}}$. Ces deux GMM sont plus performants sur $\mathbf{Y}'$ et $\mathbf{Z}'$ que sur les données $\mathbf{X}'$. Le GMM $\mathcal{M}_{\mathbf{Y}}$ est meilleur sur les données adaptées $\mathbf{Z}'$ que sur les données source $\mathbf{X}'$. Cela montre que les données adaptées (données de la source transformées par le GMM adapté) ont une distribution plus proche de celle de $\mathbf{Y}'$ que de $\mathbf{X}'$. La même observation est valable pour le modèle $\widehat{\mathcal{M}_{\mathbf{X}}}$. Cela montre bien que l'adaptation du GMM source aux données cibles a effectivement permis de se rapprocher de la distribution des données cibles.

7.5.2 Séries d'expériences 2 : adaptation moyenne/variance

Comme dans le cas de l'expérience précédente, la restitution du contour de F_0 en utilisant l'équation (7.5) permet d'obtenir un contour assez proche du contour original. La figure 7.5 page ci-contre, qui illustre ce résultat, présente le contour mélodique pour une phrase issue du locuteur *JMF* du corpus BREF120. L'erreur RMS entre le contour généré par le GMM et le contour original est dans ce cas de $10,2$Hz. En comparaison, on peut observer la figure 7.6 page suivante qui présente les contours de F_0 pour le locuteur cible *JNF* pour la même phrase que précédemment. Les contours réalisés par les deux locuteurs sont différents à la fois en termes de durée et de forme.

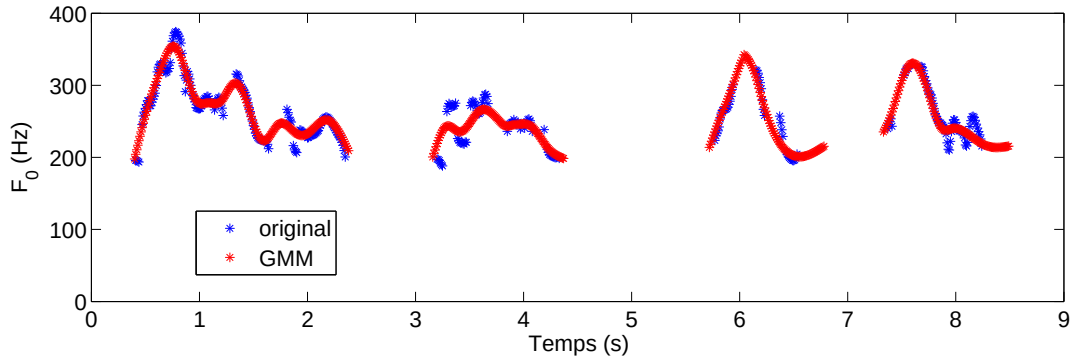


FIG. 7.5 – Génération du contour modélisé à partir des vecteurs d'observations au niveau de la syllabe et du GMM source. Cet exemple est réalisé sur une phrase issue du locuteur source *JMF* du corpus BREF120, « Des millions d'Américains ont fredonné God Bless America, White Christmas ou Puttin' on the Ritz. », dont le contour de F_0 est présenté en bleu. Le contour généré avec le GMM est présenté en rouge. L'erreur commise avec le processus de stylisation et classification complet sur cet exemple est de $10,2$Hz. L'équation (7.5) est utilisée pour obtenir le contour GMM.

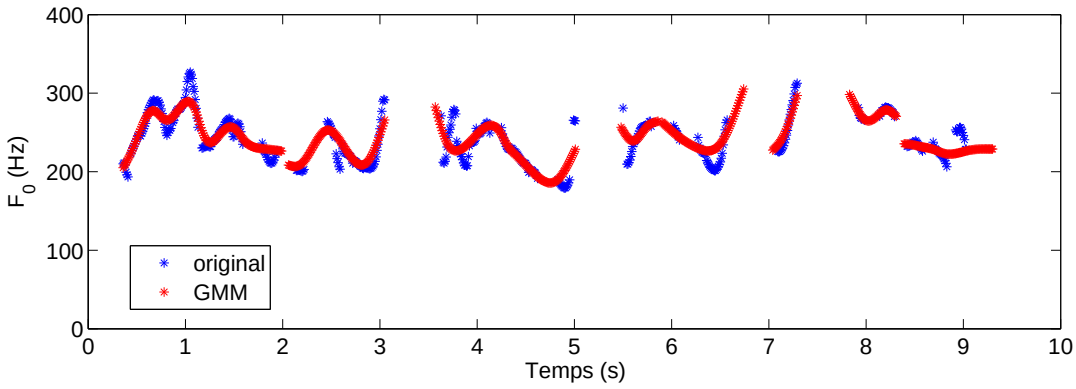


FIG. 7.6 – Génération du contour modélisé à partir des vecteurs d'observations au niveau de la syllabe et du GMM cible. Cet exemple est réalisé sur une phrase issue du locuteur cible *JNF* du corpus BREF120, « Des millions d'Américains ont fredonné God Bless America, White Christmas ou Puttin' on the Ritz. », dont le contour de F_0 est présenté en bleu. Le contour généré avec le GMM est présenté en rouge. L'erreur commise avec le processus de stylisation et classification complet sur cet exemple est de $12,2$Hz. L'équation (7.5) est utilisée pour obtenir le contour GMM.

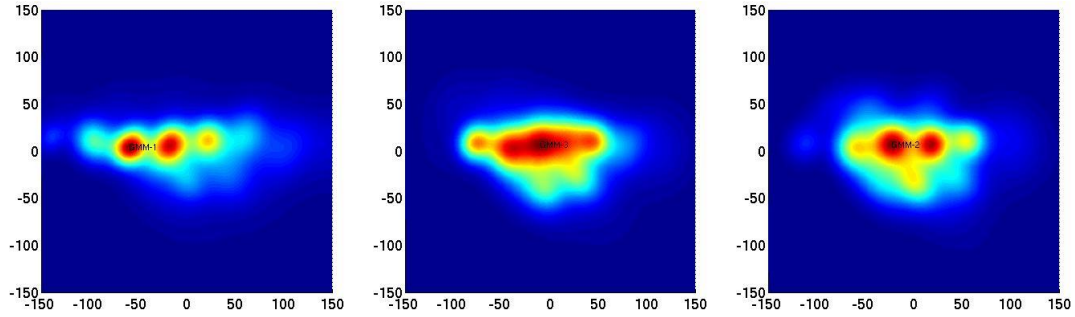


FIG. 7.7 – Projection des densités de probabilité du GMM source 7.7(a), du GMM adapté 7.7(b), et du GMM cible 7.7(c). La projection est réalisée selon les deux principales composantes obtenues par PCA de l'espace des données source et cible. On peut noter le déplacement des densités du GMM source vers celles du GMM cible.

TAB. 7.5 – Log-vraisemblances pour $\mathcal{M}_{\mathbf{X}}$ et $\mathcal{M}_{\mathbf{Y}}$ ainsi que les résultats de l'adaptation de $\mathcal{M}_{\mathbf{X}}$ en utilisant $\mathbf{Y}$ avec un intervalle de confiance à 95%.

	Appr.	Valid.
$\mathcal{M}_{\mathbf{X}}$	-21.11 ± 0.10	-21.25 ± 0.19
$\mathcal{M}_{\mathbf{Y}}$	-20.74 ± 0.10	-20.86 ± 0.20
$\widehat{\mathcal{M}_{\mathbf{X}}}$	-21.08 ± 0.10	-21.22 ± 0.18

Dans le tableau 7.5, nous pouvons observer les valeurs de log-vraisemblance pour l'apprentissage des GMM source et cible ainsi que pour l'adaptation du GMM source aux données cibles. Dans tous les cas, les valeurs de log-vraisemblance pour l'apprentissage sont proches des valeurs pour la validation. Ces résultats montrent que les GMM sont bien adaptés aux données et aucun effet d'overfitting n'apparaît.

Le tableau 7.6 page suivante montre une comparaison des valeurs de vraisemblance pour les GMM source $\mathcal{M}_{\mathbf{X}}$, cible $\mathcal{M}_{\mathbf{Y}}$ et adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$ par rapport respectivement aux données source $\mathbf{X}$ et cible $\mathbf{Y}$. En analysant ce tableau ligne par ligne, on peut noter que les GMM $\mathcal{M}_{\mathbf{X}}$ et $\mathcal{M}_{\mathbf{Y}}$ ont de meilleurs résultats respectivement sur les voix source et cible. De plus, le GMM $\widehat{\mathcal{M}_{\mathbf{X}}}$ est mieux adapté aux données de la cible, $\mathbf{Y}$, qu'à celles de la source, $\mathbf{X}$. Ces résultats montrent que le processus d'adaptation MLLR a déplacé les distributions des gaussiennes de $\mathcal{M}_{\mathbf{X}}$ vers celles de $\mathcal{M}_{\mathbf{Y}}$. Ce mouvement peut être observé sur la figure 7.7. On peut observer que la distribution adaptée 7.7(b) est plus proche de la distribution des données cibles 7.7(c). En effet, l'adaptation de la variance a permis d'élargir certaines gaussiennes tandis que l'adaptation de la moyenne a permis de déplacer la distribution source 7.7(a) vers la droite.

De la même manière que précédemment, afin d'évaluer le comportement de la fonc-

TAB. 7.6 – Log-vraisemblances pour les GMM source $\mathcal{M}_{\mathbf{X}}$, cible $\mathcal{M}_{\mathbf{Y}}$ et adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$ sur les données de validation source $\mathbf{X}$ et cible $\mathbf{Y}$.

	X	Y
$\mathcal{M}_{\mathbf{X}}$	-21.25 ± 0.19	-21.65 ± 0.19
$\mathcal{M}_{\mathbf{Y}}$	-21.49 ± 0.22	-20.86 ± 0.20
$\widehat{\mathcal{M}_{\mathbf{X}}}$	-21.68 ± 0.20	-21.22 ± 0.18

tion de transformation, considérons $\mathbf{X}'$, $\mathbf{Y}'$, $\mathbf{Z}'$ trois nouveaux ensembles de données obtenus par transformation respectivement des données :

- source par le GMM source $\mathcal{M}_{\mathbf{X}}$, en appliquant (7.5),
- cible par le GMM cible $\mathcal{M}_{\mathbf{Y}}$, en appliquant (7.5),
- source par le GMM adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$, en appliquant (7.6).

Les résultats obtenus dans le tableau 7.7 page suivante sont meilleurs qu'avec les ensembles de données originaux $\mathcal{M}_{\mathbf{X}}$ et $\mathcal{M}_{\mathbf{Y}}$. L'explication de ce phénomène est que la fonction de transformation a tendance à projeter les données vers les moyennes des gaussiennes. En effet, l'équation (7.5) indique que la valeur transformée est égale à la somme des moyennes des gaussiennes pondérées par la probabilité d'appartenance à une classe. Plus précisément, les données transformées sont situées dans l'enveloppe convexe formées par les moyennes des gaussiennes du GMM. De ce fait, la variance des données transformées est uniquement liée à la probabilité d'appartenance à une classe et elles possèdent donc une variance plus faible que les données originales. Des expériences complémentaires, non présentées ici par manque de place, ont permis de confirmer ce comportement de la fonction de transformation. Pour remédier à ce manque de variabilité, il serait utile d'introduire la variance des gaussiennes dans la fonction de tranformation.

Les résultats obtenus pour le GMM $\mathcal{M}_{\mathbf{Y}}$ montrent que les données adaptées ressemblent plus aux données cibles transformées $\mathbf{Y}'$ qu'aux données sources transformées $\mathbf{X}'$. Pour le GMM $\widehat{\mathcal{M}_{\mathbf{X}}}$, les résultats sont plus mitigés. En effet, il semble que les données $\mathbf{X}'$ soient plus vraisemblables que les données $\mathbf{Y}'$. Cela a tendance à montrer que les données source et cible, sont difficilement séparables. Il serait alors intéressant d'enrichir le modèle de prosodie, par exemple en introduisant la dérivée du F_0 au niveau des trois points sélectionnés. Ces informations supplémentaires permettraient de mieux distinguer la voix source de la voix cible.

Un exemple de transformation d'un contour mélodique pour une phrase est présenté sur la figure 7.8 page suivante. Pour réaliser cet exemple, les vecteurs au niveau syl-

TAB. 7.7 – Comparaison des valeurs de log-vraisemblance pour les GMM $\mathcal{M}_{\mathbf{X}}$, $\mathcal{M}_{\mathbf{Y}}$ et $\widehat{\mathcal{M}_{\mathbf{X}}}$ sur les ensembles de données de validation transformés.

	$\mathbf{X}'$	$\mathbf{Y}'$	$\mathbf{Z}'$
$\mathcal{M}_{\mathbf{X}}$	-18.60 ± 0.15	-19.80 ± 0.17	-19.34 ± 0.15
$\mathcal{M}_{\mathbf{Y}}$	-19.00 ± 0.16	-18.80 ± 0.17	-18.82 ± 0.16
$\widehat{\mathcal{M}_{\mathbf{X}}}$	-19.13 ± 0.16	-19.27 ± 0.16	-18.66 ± 0.15

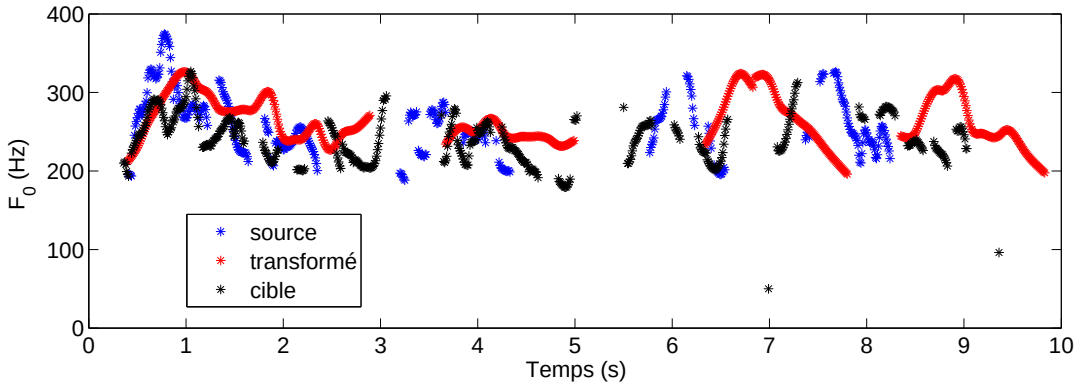


FIG. 7.8 – Exemple de transformation d'un contour mélodique. Les contours source (en bleu), cible (en noir) et transformé (en rouge) sont présentés sur cette figure. La transformation réalisée utilise le processus d'adaptation afin d'estimer les paramètres du GMM adapté et de construire la fonction de transformation définie par l'équation (7.6). La phrase transformée est la même que celle présentée à la figure 7.5 page 157 et dont la phrase correspond à « Des millions d'Américains ont fredonné God Bless America, White Christmas ou Puttin' on the Ritz. ». On peut observer sur cette figure une modification de la durée qui reflète le débit plus lent du locuteur cible. Une modification du contour mélodique au niveau de sa forme est également perceptible. Même si le contour transformé n'est pas très proche de celui du locuteur cible, on peut remarquer qu'il se démarque de celui réalisé par le locuteur source ce qui apporte un atout par rapport à une normalisation gaussienne.

labique du locuteur source sont transformés un par un en appliquant l'équation (7.6). Cette transformation permet d'obtenir le contour rouge en générant, pour chaque vecteur, un contour de F_0 puis en concaténant les contours au niveau de la syllabe. Le résultat semble assez différent du contour réalisé par le locuteur source ce qui peut être considéré comme un atout indéniable par rapport à la méthode par normalisation gaussienne. La durée globale du contour mélodique transformé est plus importante que celle du contour source. Cette observation est bien cohérente avec le débit plus lent du locuteur cible par rapport au locuteur source.

Cette figure permet de se rendre compte visuellement d'une transformation d'un contour mélodique mais reste insuffisante afin de juger de la qualité du système. Ainsi, une évaluation en situation, en confrontant le système de transformation à un auditeur qui va juger de la similarité perceptive du contour transformé par rapport à celui qu'aurait réalisé le locuteur cible est importante.

7.6 Conclusion

Nous avons présenté une méthodologie permettant de répondre au problème de la transformation de la prosodie en présence de corpus non parallèles. La durée et le F_0 au niveau d'une syllabe sont représentés sous la forme d'un vecteur de taille fixe. Un modèle GMM est estimé sur les observations de durée et de F_0 pour un locuteur source. Nous proposons dans un premier temps l'application d'une approche MLLR pour l'adaptation des paramètres de la source en fonction de données de la voix cible. Dans un second temps, nous présentons un modèle de transformation linéaire d'un vecteur prosodique de la voix source. Ce modèle pondère les centroïdes adaptés en fonction de la distribution a posteriori des vecteurs de la voix source.

Le protocole expérimental repose essentiellement sur une validation croisée des modèles et des ensembles de données de validation. Une comparaison exhaustive entre modèles et données montre d'une part que le GMM adapté par MLLR modélise efficacement les données de la cible et d'autre part que la fonction de transformation produit des données aussi vraisemblables pour un modèle de cible que les données cibles elles-mêmes.

Deux variantes de l'adaptation sont utilisées. La première met en œuvre une adaptation de la moyenne uniquement tandis que la seconde réalise également l'adaptation des variances des gaussiennes. Les deux techniques permettent de déplacer la distribution du GMM source vers celle du GMM cible. L'ajout de l'adaptation de la variance permet d'augmenter légèrement les résultats mais pas de manière significative.

Le modèle proposé est statique, il ne tient pas compte du contexte syllabique des séquences de vecteurs prosodiques. Un prolongement à ce travail pourrait être la prise en compte de l'effet contextuel des observations en posant par exemple une modélisation markovienne. L'évaluation d'une méthode non parallèle est difficile et l'utilisation de corpus parallèles dans un objectif d'évaluation serait utile. D'autres expérimentations doivent être menées pour situer l'approche non parallèle par rapport à des techniques éprouvées qui reposent par exemple sur la DTW.

Une méthode d'évaluation doit alors être définie en tenant compte du fait que la transformation de la prosodie, ou de la voix en général, n'a pas pour objectif de faire *exactement comme* le locuteur cible. Doit-on alors demander à un auditeur de comparer la phrase transformée avec la phrase cible, ou plutôt rechercher une méthodologie d'évaluation différente à la manière de celle proposée par Helander et Nurminen (2007) ? Cette méthodologie, qui est décrite au paragraphe 4.2.1 page 86, permet de ne pas se contraindre à l'utilisation de corpus parallèles et semble une alternative intéressante au méthodologies d'évaluation classiques.

Conclusion de la deuxième partie

Dans cette deuxième partie, nous avons présenté des travaux concernant la stylisation, la classification et enfin la transformation de la prosodie.

Dans le chapitre 5 page 95, nous proposons un modèle génératif permettant de styliser la prosodie. Ce modèle repose sur un outil mathématique, les B-splines, qui permet de tenir compte explicitement des non-linéarités de la courbe de F_0, notamment au niveau des transitions voisées/non voisées. Un critère MDL est également mis en œuvre pour sélectionner le nombre de paramètres du modèle. Ce modèle permet d'obtenir une stylisation très fine des contours mélodiques puisque l'erreur RMS obtenue est égale à 0.42Hz pour un nombre de degrés de liberté du modèle égal à 63% du modèle complet. En contrepartie, le pouvoir explicatif d'un tel modèle est quasi-inexistant et l'étude de ce modèle nous a montré qu'il est difficile de l'utiliser afin de générer de la prosodie.

En particulier, les variations de longueur des contours mélodiques posent des problèmes pour les comparer entre-eux. Dans le chapitre 6 page 125, nous proposons une méthodologie qui repose sur l'utilisation de HMM afin d'absorber les variations de longueur des contours mélodiques. Le partitionnement des contours mélodiques est ensuite construit de manière hiérarchique descendante. Les résultats obtenus font état d'une erreur RMS voisine de 4Hz pour un nombre de classes égal à 64 sur les données utilisées. Cependant, il ne faut pas perdre de vue que ces résultats sont obtenus sur un corpus de parole lue, et que sur un corpus expressif, pour lequel les variations de F_0 sont plus importantes, les résultats seraient sans doute moins bons. Nous avons également montré expérimentalement que les variations de longueur sont convenablement modélisées par les HMM.

Une méthodologie de transformation reposant sur une régression linéaire est présentée dans le chapitre 7 page 143. L'apport de cette méthodologie consiste à transformer la prosodie entre deux locuteurs sans qu'il n'y ait nécessairement de parallélisme entre les données source et cible. Les résultats expérimentaux montrent que le processus d'adaptation permet de déplacer les distributions du modèle de la source vers celles du modèle

de la cible. Cependant, il faudrait également prendre en compte la dynamique à long terme de la prosodie et être en mesure d'incorporer des données d'ordre morphosyntaxique. En effet, même si le modèle proposé permet de supprimer la contrainte du parallélisme au niveau du texte, il apparaît nécessaire d'établir un parallélisme structurel entre les vecteurs source et cible que l'on souhaite faire correspondre.

Conclusion

Les travaux réalisés dans cette thèse ont pour objectif commun la transformation de la prosodie, et plus précisément les contours mélodiques, d'un locuteur source vers un locuteur cible. Le cheminement que nous avons adopté trouve son origine dans la stylisation des contours mélodiques. Cette étape est, comme nous l'avons noté, nécessaire et un préalable à la création d'une fonction de transformation.

Nous avons vu au chapitre 5 page 95 une méthode reposant sur les B-splines associée à un critère MDL permettant de styliser de manière fine les contours mélodiques d'un locuteur. L'efficacité de ce modèle tient au fait qu'il modélise de manière intrinsèque les discontinuités des contours mélodiques. Cependant, cette approche ne permet pas de résumer l'espace mélodique d'un locuteur à un ensemble restreint de modèles. Une étape de classification s'avère nécessaire. Nous savons également qu'un autre problème découle du précédent : pour établir un partitionnement, il faut être en mesure de comparer les contours mélodiques, ou leurs modèles, deux à deux.

Une telle fonction de comparaison, dans le cadre du modèle proposé n'est pas évidente. Aussi nous avons jugé plus utile de chercher un modèle qui permettrait de rendre compte de la forme générale d'un contour mélodique tout en autorisant une comparaison des contours à travers le modèle. Au chapitre 6 page 125, nous proposons une méthodologie utilisant un modèle HMM qui apporte une solution aux problèmes précédents. Les résultats en terme de classification grâce à ce modèle montrent que les classes obtenues contiennent des contours mélodiques de longueurs assez différentes et de formes similaires. On pourra cependant noter que pour des contours présentant des variations assez fortes, le modèle HMM proposé se trouve en difficulté. Une solution peut être apportée à ce problème en incorporant la dérivée première du F_0 dans le modèle.

Les résultats de ce modèle étant assez satisfaisants sur les corpus utilisés, nous avons ensuite tenté de l'utiliser afin de construire une fonction de transformation entre deux locuteurs. Nous nous sommes alors retrouvés confrontés au problème de l'appariement des contours du locuteur source à ceux du locuteur cible afin de créer une fonction

de transformation par « région » de l'espace prosodique. Ne trouvant pas de solution satisfaisante à ce problème, nous avons relâché la contrainte d'appariement, que l'on retrouve également dans le problème du parallélisme entre données sources et cibles dans les méthodologies de transformation de la voix.

Le chapitre 7 page 143 propose alors une méthodologie de transformation de la prosodie par adaptation de modèles. Un GMM est utilisé pour modéliser l'espace prosodique du locuteur source. Ce modèle est ensuite adapté aux données du locuteur cible en appliquant une régression linéaire MLLR. Cet algorithme d'adaptation, issu de travaux en reconnaissance de la parole, est utilisé pour modifier les paramètres du GMM source de façon à maximiser la vraisemblance par rapport aux données cibles. Il permet d'obtenir la matrice de transformation qui effectue le passage du modèle source au modèle adapté. Celle-ci peut alors être utilisée afin de construire une fonction de transformation.

L'avantage de cette méthode est de pouvoir apprendre une fonction de transformation en utilisant des données sources et cibles non parallèles. Une telle approche permettrait de créer plus facilement et à moindre coût de nouvelles voix. Cependant, il est important de noter que dans son état actuel, le modèle utilisé pour représenter l'espace prosodique d'un locuteur ne tient pas compte de la dynamique de la prosodie à travers plusieurs syllabes. De plus, nous avons également noté que l'appariement réalisé entre les contours sources et cibles devrait être piloté par des informations d'ordre morphosyntaxique.

Au cours de cette thèse, un certain nombre de difficultés ont été rencontrées et il a été nécessaire à chaque fois de trouver une solution, parfois pragmatique. Ainsi, nous avons montré que le modèle B-spline associé à un critère MDL permet une bonne stylisation des contours mélodiques. Cependant, comment peut-on comparer deux modèles qui possèdent un nombre de paramètres différents ? Nous n'avons pas trouvé de solution simple à ce problème à part utiliser un modèle possédant un nombre fixe de paramètres ce qui diminue fortement l'intérêt d'un tel modèle. Notamment, certains contours de F_0 devraient être écartés de la modélisation car le nombre de paramètres du modèle serait plus important que le nombre de points du contour. Ainsi, la solution que nous avons privilégiée est celle de traiter en premier lieu les différences de longueurs des contours en utilisant des HMM. Le modèle HMM permet de construire des classes de contours de F_0 assez homogènes par rapport à la forme de ceux-ci et de manière assez indépendante de la longueur des contours. Là encore, un problème de taille se pose lorsqu'il s'agit d'utiliser ces classes afin de construire une fonction de transformation. Plusieurs pistes, qui n'ont pas donné de résultats significatifs pour le moment, ont également été envisagées en

suivant trois axes : prédire la classe du contour suivant à partir du contour courant et des précédents, faire correspondre les classes du locuteur source à celles du locuteur cible, et construire une fonction de transformation par classe du locuteur source.

Finalement, ces difficultés ont permis de réorienter la thèse et d'aboutir à des propositions nouvelles pour styliser, classifier et transformer les contours mélodiques. Elles ont également contribué à mieux cerner le problème de transformation afin de proposer plusieurs axes de travail qui peuvent compléter cette étude :

- Utiliser l'information sur les classes de contours du locuteur source afin de construire une fonction de transformation pour chaque *région* de l'espace mélodique source.
- Intégrer la dynamique du F_0 dans la fonction de transformation : l'évolution de la mélodie possède une cohérence globale au niveau de la phrase, du groupe intonatif qu'il faut exploiter.
- Piloter la transformation par des informations morphosyntaxiques : ce point que nous avons déjà cité est important puisqu'il permettrait de retrouver un certain degré de parallélisme entre les syllabes prononcées par chaque locuteur et d'intégrer des facteurs contextuels sur la nature des syllabes précédentes et suivantes ainsi que sur la structure de la phrase.
- Proposer une méthodologie d'évaluation des systèmes de transformation de la voix : dans ce domaine, les méthodologies utilisées ne sont pas forcément adaptées à l'objectif de transformation et il n'existe pas de consensus qui permettait une comparaison efficace des différents systèmes. Une telle proposition pourrait éventuellement passer par la construction de corpus dédiés à la transformation du timbre ou de la prosodie.

Bibliographie

M. Abe, S. Nakamura, K. Shikano et H. Kuwabara : Voice conversion through vector quantization. *In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, p. 655–658, 1988.

A. J. Abrantes, J. S. Marques et I. M. Trancoso : Hybrid sinusoidal modeling of speech without voicing decision. *In Proceedings of the Second European Conference on Speech Communication and Technology (Eurospeech)*, p. 231–234, 1991.

P. D. Agüero, K. Wimmer et A. Bonafonte : Automatic analysis and synthesis of fujisaki's intonation model for tts. *In Proceedings of Speech Prosody 2004*, p. 427–430, Japan, 2004.

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

B. S. Atal : Automatic speaker recognition based on pitch contours. *Journal of the Acoustical Society of America*, 52(6):1687–1697, 1972.

B. Atal : Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, 64(4):460–475, 1976.

P. Barbosa et G. Bailly : Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication*, 15(1-2):127–137, 1994.

N. Barbot, O. Boeffard et D. Lolive : F0 stylisation with a free-knot b-spline model and simulated-annealing optimization. *In Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech)*, p. 325–328, Lisboa, Portugal, 2005.

M. E. Beckman et G. A. Elam : *Guidelines for ToBI labeling, Version 3.* Ohio State University, 1997.

M. E. Beckman et J. Hirschberg : *The ToBI Annotation Conventions*. URL `http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html`.

G. Beliakov : Least squares splines with free knots : global optimization approach. *Applied Mathematics and Computation*, 149:783–798, 2004.

G. Beller, D. Schwarz, T. Hueber et X. Rodet : Speech rates in french expressive speech. *In Speech Prosody 2006*, 2006.

J. A. Bilmes : A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Rap. tech., International Computer Science Institute, 1998.

F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz et D. A. Reynolds : A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 1:430 – 451, 2004.

A. Black et A. Hunt : Generating f0 contours from tobi labels using linearregression. *In Proceedings of the Fourth International Conference on Spoken Language (ICSLP'96)*, vol. 3, p. 1385–1388, Philadelphia, PA, USA, 1996.

O. Boeffard : Contributions à la synthèse de la parole. Habilitation à Diriger les Recherches, Université de Rennes 1, 2004.

O. Boeffard et F. Emerard : Application-dependent prosodic models for text-to-speech synthesis and automatic design of learning database corpus using genetic algorithm. *In Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)*, p. 2511–2514, September 1997.

P. Boersma et D. Weenink : Praat : doing phonetics by computer, 2008. URL `http://www.praat.org/`.

R. Boite, H. Bourlard, T. Dutoit, J. Hancq et H. Leich : *Traitement de la Parole*. Presses Polytechniques Universitaires Romandes, 2000. ISBN 2-88074-388-5.

J.-F. Bonastre, F. Bimbot, L.-J. Boe, J. P. Campbell, D. A. Reynolds et I. Magrin-Chagnolleau : Person authentication by voice : A need for caution. *In Proceedings of the 8th European Conference on Speech Communication and Technology*, p. 33–36, 2003.

J.-F. Bonastre, D. Matrouf et C. Fredouille : Augmentation du taux de fausse acceptation par transformation inaudible de la voix des imposteurs. *In Actes des XXVIèmes Journées d'Etudes sur la Parole*, p. 11–14, Dinard, France, 2006.

A. Breen : Speech synthesis models : a review. *Electronics & Communication Engineering Journal*, 4(1):19–31, 1992.

L. Breiman, J. H. Friedman, R. A. Olshen et C. J. Stone : *Classification and Regression Trees*. Wadsworth Inc, 1984.

H. G. Burchard : Splines (with optimal knots) are better. *Applicable Analysis*, 3:309–319, 1974.

Calliope : *La parole et son traitement automatique*. Editions Masson, 1989.

J. P. Campbell : Speaker recognition : A tutorial. *Proceedings of the IEEE*, 85 (9):1437–1462, 1997.

E. Campione et J. Véronis : Une évaluation de l'algorithme de stylisation mélodique momel. *Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence*, 19:27–44, 2000.

T. Ceyssens, W. Verhelst et P. Wambacq : On the construction of a pitch conversion system. *In Proceedings of the European Signal Processing Conference*, p. 1301–1304, 2002.

T.-J. Cham et R. Cipolla : Automated b-spline curve representation incorporating mdl and error-minimizing control point insertion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(1):49–53, 1999.

D. T. Chappell et J. H. L. Hansen : Speaker-specific pitch contour modeling and modification. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, p. 885–888, 1998.

L. Charonnat, G. Vidal et O. Boeffard : Automatic phone segmentation of expressive speech. *In Proceedings of the International Conference on Language Resources and Evaluation*, 2008.

C. Coker : A model of articulatory dynamics and control. *In Proceedings of the IEEE*, vol. 64, p. 452–460, 1976.

C. d'Alessandro, V. Darsinos et B. Yegnanarayana : Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. *IEEE Transactions on Speech and Audio Processing*, 6(1):12–23, 1998.

C. De Boor : Splines as linear combinations of b-splines. a survey. *Approximation Theory II*, p. 1–47, 1976.

A. de Cheveigne et H. Kawahara : Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111:1917–1930, 2002.

S. de Tournemire : *Identification et génération automatique de contours prosodiques pour la synthèse vocale à partir du texte en français.* Thèse, Ecole nationale supérieure des télécommunications, 1998.

P. Delattre : Les dix intonations de base du français. *The French Review*, 40(1):1–14, 1966.

A. P. Dempster, N. M. Laird et D. B. Rubin : Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 39 (1):1–38, 1977.

A. Di Cristo : *Intonation in French, in Intonation Systems : A Survey of Twenty Languages*, chap. 11, p. 195–218. Cambridge University Press, 1998. ISBN 052139550X.

A. Di Cristo : Interpréter la prosodie. *In XXIIèmes Journées d'Etudes sur la Parole*, Aussois, France, 2000.

V. Digalakis, J. R. Rohlicek et M. Ostendorf : Ml estimation of a stochastic linear system with the em algorithm and its application to speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1(4):431–42, October 1993.

G. Doddington : Speaker recognition - identifying people by their voices. *Proceedings of the IEEE*, 73(11):1651–1664, 1985.

K. Dusterhoff et A. W. Black : Generating f0 contours for speech synthesis using the tilt intonation theory. *In ESCA Workshop of Intonation*, p. pp 107–110, September 1997.

K. E. Dusterhoff, A. W. Black et P. Taylor : Using decision trees within the tilt intonation model to predict f0 contours. *In Proceeding of the Sixth European Conference on Speech Communication and Technology (Eurospeech)*, p. 1627–1630, 1999.

H. Duxans, A. Bonafonte, A. Kain et J. Van Santen : Including dynamic and phonetic information in voice conversion systems. *In Proceedings of the International Conference on Spoken Language Processing*, p. 1193–1196, 2004.

H. Duxans, D. Erro, J. Pérez, F. Diego, A. Bonafonte et A. Moreno : Voice conversion using exclusively unaligned training data. *In TC-STAR workshop on speech-to-speech translation*, p. 237–242, june 2006.

D. Erickson : Expressive speech : Production, perception and application to speech synthesis. *Acoustical Science and Technology*, 26(4):317–325, 2005.

M. A. T. Figueiredo, J. M. N. Leitão et A. K. Jain : Unsupervised contour representation and estimation using b-splines and a minimum description length criterion. *IEEE Transactions on Image processing*, 9(6):1075–1086, 2000.

I. Fónagy : Fonctions de l'intonation : Essai de synthèse. *Flambeau*, 29:1–20, 2003.

H. Fujisaki : Information, prosody, and modeling - with emphasis on tonal features of speech. *In Proceedings of Speech Prosody 2004*, p. 1–10, 2004.

H. Fujisaki et K. Hirose : Analysis of voice fundamental frequency contours for declarative sentence of japanese. *Journal of the Acoustical Society Japan*, 5(4):233–242, 1984.

M. J. Gales et P. C. Woodland : Mean and variance adaptation within the mllr framework. *Computer Speech & Language*, 10:249–264, 1996.

B. Gillett et S. King : Transforming f0 contours. *In Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, p. 1713–1716, Geneva, September 2003.

O. Goubanova et S. King : Bayesian networks for phone duration prediction. *Speech Communication*, 50(4):301–311, 2008.

V. Granville, M. Krivanek et J.-P. Rasson : Simulated annealing : a proof of convergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16 (6):652–656, June 1994.

D. Griffin : Multiband excitation vocoder. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(8):1223–1235, 1988.

M. H. Hansen et C. Kooperberg : Spline adaptation in extended linear models. *Statistical Science*, 17:2–51, 2002.

M. H. Hansen et B. Yu : Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.

Z. Hanzlicek et J. Matousek : F0 transformation within the voice conversion framework. *In Proceedings of the Interspeech Conference*, p. 1961–1964, Antwerp, Belgium, 27-31 August 2007.

E. E. Helander et J. Nurminen : A novel method for prosody prediction in voice conversion. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, p. 509–512, 2007.

D. Hirst, N. Ide et J. Véronis : Coding fundamental frequency patterns for multilingual synthesis with intsint in the multext project. *In Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, p. 77–80, Mohonk Mountain House, New Paltz, NY, USA, September 1994.

D. J. Hirst, A. Di Cristo et R. Espesser : Levels of representation and levels of analysis for the description of intonation systems. *In* M. Horne, éd. : *Prosody : Theory and Experiment, Kluwer Academic Pusblisher*, vol. 14, p. 51–87, 2000.

D. J. Hirst et R. Espesser : Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, 15:75–85, 1993.

H. Hollien, W. Majewski et E. T. Doherty : Perceptual identification of voices under normal, stress, and disguised speaking conditions. *Journal of Phonetics*, 10:139–148, 1982.

J. Holmes : The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer. *IEEE Transactions on Audio and Electroacoustics*, 21 (3):298–305, 1973.

Z. Inanoglu : Transforming pitch in a voice conversion framework. Rap. tech., St. Edmond's College, University of Cambridge, july 2003.

L. Ingber : Adaptive simulated annealing (asa) : lessons learned. *Control and Cybernetics*, 25(1):33–54, 1996.

D. L. B. Jupp : Approximation to data by splines with free knots. *SIAM Journal of Numerical Analysis*, 15(2):328–343, 1978.

A. Kain et M. Macon : Spectral voice conversion for text-to-speech synthesis. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, p. 285–288, 1998.

A. B. Kain : *High Resolution Voice Transformation.* Thèse de doctorat, Faculty of the OGI School of Science and Engineering at Oregon Health and Science University, 2001.

Y. Kang, J. Tao et B. Xu : Applying pitch target model to convert f0 contour for expressive mandarin speech synthesis. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, p. 733–736, 2006.

S. Kirkpatrick, C. Gelatt et M. P. Vecchi : Optimization by simulated annealing. *Science*, 220(4598):671–680, 13 May 1983.

D. H. Klatt : Discrimination of fundamental frequency contours in synthetic speech : implications for models of pitch perception. *The Journal of the Acoustical Society of America*, 53(1):8–16, January 1973.

D. H. Klatt : Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, 67(3):971–995, 1980.

D. H. Klatt : Review of text-to-speech conversion for english. *The Journal of the Acoustical Society of America*, 82(3):737–793, September 1987.

H. Kuwabara et Y. Sagisaka : Acoustic characteristics of speaker individuality : Control and conversion. *Speech Communication*, 16(2):165–173, 1995.

L. F. Larnel, J.-L. Gauvain et M. Eskenazi : Bref, a large vocabulary spoken corpus for french. *In Proceedings of the Second European Conference on Speech Communication and Technology (Eurospeech)*, p. 505–508, 24-26 September 1991.

T. C. M. Lee : An introduction to coding theory and the two-part minimum description length principle. *International Statistical Review*, 69(2):169–183, 2001.

C. J. Leggetter et P. C. Woodland : Flexible speaker adaptation using maximum likelihood linear regression. *In Proceedings of the Fourth European Conference on Speech Communication and Technology (Eurospeech)*, p. 1155–1158, 1995.

Y. Linde, A. Buzo et R. M. Gray : An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980.

M. Lindstrom : Penalized estimation of free-knot splines. *Journal of Computational and Graphical Statistics*, 8:333–352, 1999.

D. Lolive, N. Barbot et O. Boeffard : Comparing b-spline and spline models for f0 modelling. *In* P. Sojka, I. Kopecek et K. Pala, éds : *Lecture Notes in Artificial Intelligence - Proceedings of the 9th International Conference on Text, Speech and Dialogue - Brno, Czech Republic*, vol. 4188, p. 423–430, Berlin, Heidelberg, 2006a. Springer Verlag.

D. Lolive, N. Barbot et O. Boeffard : Melodic contour estimation with b-spline models using a mdl criterion. *In Proceedings of the 11th International Conference on Speech and Computer (SPECOM)*, p. 333–338, Saint Petersburg, Russia, 2006b.

D. Lolive, N. Barbot et O. Boeffard : Modélisation b-spline de contours mélodiques avec estimation du nombre de paramètres libres par un critère mdl. *In Actes des XXVIèmes Journées d'Etudes sur la Parole*, p. 499–502, Dinard, France, 2006c.

D. Lolive, N. Barbot et O. Boeffard : Proposition d'un critère mdl pour l'estimation de courbes ouvertes modélisées par des b-splines. *In* L. Miclet, éd. : *Actes de la 8ème Conférence Francophone sur l'Apprentissage Automatique - Trégastel, France*, p. 219–234. Presses Universitaires de Grenoble, 2006d.

D. Lolive, N. Barbot et O. Boeffard : Clustering algorithm for f0 curves based on hidden markov models. *In* P. Wagner, J. Abresh et W. Hess, éds : *Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, p. 85–89, Bonn, Germany, 2007a.

D. Lolive, N. Barbot et O. Boeffard : Unsupervised hmm classification of f0 curves. *In Proceedings of Interspeech'2007*, p. 478–481, Antwerp, Belgium, 2007b.

D. Lolive, N. Barbot et O. Boeffard : Pitch and duration transformation with non-parallel data. *In* P. A. Barbosa, S. Madureira et C. Reis, éds : *Proceedings of Speech Prosody 2008*, p. 111–114, 2008a.

D. Lolive, N. Barbot et O. Boeffard : Transformation de la prosodie par adaptation mllr de gmm. *In Actes des XXVIIèmes Journées d'Etudes sur la Parole*, Avignon, France, 2008b.

A. Louw et E. Barnard : Automatic intonation modeling with intsint. *In Proceedings of the 15th Annual Symposium of the Pattern Recognition Association of South Africa*, p. 107–111, November 2004.

J. Makhoul : Linear Prediction : A Tutorial Review. *Proceedings of the IEEE*, 63 (4):561, avr. 1975.

F. Malfrère, T. Dutoit et P. Mertens : Automatic prosody generation using suprasegmental unit selection. *In Proceedings of the 3rd ESCA/COCSADA Workshop on Speech Synthesis*, p. 323–328, Jenolan Caves, Australia, 1998.

T. Masuko, K. Tokuda, T. Kobayashi et S. Imai : Speech synthesis using hmms with dynamic features. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, p. 389–392, 1996.

T. Masuko, K. Tokuda, N. Miyazaki et T. Kobayashi : Pitch pattern generation using multispace probability distribution hmm. *Systems and Computers in Japan*, 33 (6):62 – 72, 2002.

H. Matsumoto, S. Hiki, T. Sone et T. Nimura : Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Transactions on Audio and Electroacoustics*, 21(5):428–436, 1973.

P. Mertens : L'accentuation de syllabes contiguës. *I.T.L.*, v95:p145–165, 1992.

P. Mertens : Accentuation, intonation et morphosyntaxe. Rap. tech., Département de linguistique, K.U. Leuven, 1993.

P. Mertens : Automatic recognition of intonation in french and dutch. *In Proceedings of the First European Conference on Speech Communication and Technology (Eurospeech)*, p. 46–50, 1989.

L. Mesbahi, V. Barreaud et O. Boëffard : Comparing gmm-based speech transformation systems. *In Proceedings of Interspeech'2007*, p. 1989–1992, Antwerp, Belgium, 2007.

L. Mesbahi, V. Barreaud et O. Boëffard : Non-parallel hierarchical training for voice conversion. *In Proceedings of the 16th European Signal Processing Conference*, Lausanne, Switzerland, 2008.

R. Miller : *The structure of singing.* Schimer Books, Macmillan Inc, 1986.

H. Mixdorff : *Intonation Patterns of German - Model-based Quantitative Analysis and Synthesis of F0-Contours.* Thèse de doctorat, TU Dresden, 1998.

H. Mixdorff : A novel approach to the fully automatic extraction of fujisaki model parameters. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, p. 1281–1284, Istanbul, Turkey, 2000.

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

Y. Morlec, G. Bailly et V. Aubergé : Synthesis and evaluation of intonation with a superposition model. *In Proceedings of the Fourth European Conference on Speech Communication and Technology (Eurospeech)*, p. 2043–2046, 1995.

A. Mouchtaris, V. der Spiegel et P. Mueller : Non-parallel training for voice conversion by maximum likelihood constrained adaptation. *In ICASSP*, vol. 1, p. I–1,I–4, 2004.

S. Mouline, O. Boëffard et P. C. Bagshaw : Automatic adaptation of the momel f0 stylisation algorithm to new corpora. *In Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, p. 961–964, Jeju Island, Korea, 2004.

G. Möhler : Comparing two different principles of parametric f_0 modeling. *Acoustical Society of America Journal*, 105:1245–+, fév. 1999.

G. Möhler et A. Conkie : Parametric modeling of intonation using vector quantization. *In Proceedings of the 3rd ISCA Workshop on Speech Synthesis*, p. 311–316, November 1998.

M. Narendranath, H. A. Murthy, S. Rajendran et B. Yegnanarayana : Transformation of formants for voice conversion using artificial neural networks. *Speech Communication*, 16:207–216, 1995.

S. Narusawa, N. Minematsu, K. Hirose et H. Fujisaki : Automatic extraction of model parameters from fundamental frequency contours of english utterances. *In Proceedings of the 7th International Conference on Spoken Language Processing*, p. 1725–1728, Denver, Colorado, USA, 2002.

B. Necioglu, M. Clements, T. Barnwell et A. Schmidt-Nielsen : Perceptual relevance of objectively measured descriptors forspeaker characterization. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, p. 869–872, 1998.

L. R. Rabiner : A tutorial on hidden markov models and selected applications in speech recognition. *In Proceedings of the IEEE*, vol. 77, p. 257–286, 1989. ISSN 0018-9219.

L. R. Rabiner, C.-H. Lee, B.-H. Juang et J. G. Wilpon : HMM clustering for connected word recognition. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, p. 405–408, mai 1989.

U. D. Reichel : Data-driven extraction of intonation contour classes. *In Proceedings of the 6th ISCA Workshop on Speech Synthesis, Bonn.*, p. 240–245, 2007.

M. D. Riley : Tree-based modelling for speech synthesis. *In Proceedings of the ESCA Workshop on Speech Synthesis*, p. 229–232, Autrans, France, September 1990.

J. Rissanen : *Stochastic complexity in Statistical Inquiry*. World Scientific, Singapore, 1989.

A. Rosenberg : Automatic speaker verification : A review. *Proceedings of the IEEE*, 64(4):475–487, 1976.

K. N. Ross et M. Ostendorf : A dynamical system model for generating fundamental frequency for speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 7(3):295–309, May 1999.

R. A. Rutenbar : Simulated annealing algorithms : an overview. *IEEE Circuits and Design Magazine*, 5(1):19–26, 1989.

S. Sakai et J. Glass : Fundamental frequency modeling for corpus-based speech synthesis based on statistical learning techniques. *In Proceedings of the ASRU Conference*, p. 712–717, 2003.

A. Sakurai, K. Hirose et N. Minematsu : Data-driven generation of f0 contours using a superpositional model. *Speech Communication*, 40(4):535–549, June 2003.

A. Sankar : Experiments with a Gaussian Merging-Splitting Algorithm for HMM training for Speech Recognition. *In Proceedings of the DARPA Speech Recognition Workshop*, fév. 1998.

A. Schmidt-Nielsen et T. H. Crystal : Speaker verification by human listeners : Experiments comparing human and machine performance using the nist 1998 speaker evaluation data. *Digital Signal Processing*, 10:249–266, 2000.

H. Schwetlick et T. Schütze : Least squares approximation by splines with free knots. *BIT Numerical Mathematics*, 35:361–384, 1995.

C. H. Shadle et R. I. Damper : Prospects for articulatory synthesis : A position paper. *In Proceedings of the Fourth ISCA ITRW on Speech Synthesis*, 2001.

S. d. S. Silva et S. L. Netto : Closed-form estimation of the amplitude commands in the automatic extraction of the fujisaki's model. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, p. 621, 2004.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert et J. Hirschberg : Tobi : A standard for labeling english prosody. *In Proceedings of the Second International Conference on Spoken Language Processing (ICSLP'92)*, p. 867–870, Banff, Alberta, Canada, October 1992.

S. Singh et T. Murry : Multidimensional classification of normal voice qualities. *The Journal of the Acoustical Society of America*, 64(1):81–87, 1978.

Y. Stylianou, O. Cappe et E. Moulines : Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6(2):131–142, 1998.

Y. Stylianou, J. Laroche et E. Moulines : High-quality speech modification based on a harmonic + noise model. *In Proceedings of the Fourth European Conference on Speech Communication and Technology (Eurospeech)*, p. 451–454, 1995.

M. Tamura, T. Masuko, K. Tokuda et T. Kobayashi : Adaptation of pitch and spectrum for hmm-based speech synthesis using mllr. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, p. 805–808, 2001.

J. Tao, Y. Kang et A. Li : Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1145–1154, July 2006.

P. A. Taylor : Automatic recognition of intonation from F0 contours using the rise/fall/connection model. *In Proceedings of Third European Conference on Speech Communication and Technology (Eurospeech), Berlin*, 1993.

P. A. Taylor : The rise/fall/connection model of intonation. *Speech Communication*, v15:p169–186, 1995.

P. A. Taylor : The Tilt intonation model. *In Proceedings of the Fifth International Conference on Spoken Language Processing*, Sydney, 1998.

P. A. Taylor : Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, v107(3):p1697–1714, 2000.

K. Tokuda, H. Zen et A. Black : An hmm-based speech synthesis system applied to english. *In Proceedings of the IEEE Workshop on Speech Synthesis*, p. 227–230, 2002.

K. Tokuda, T. Masuko, N. Miyazaki et T. Kobayashi : Hidden markov models based on multi-space probability distribution for pitch pattern modeling. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, p. 229–232, 1999.

C. Traber : *Talking Machines : Theories, Models and Designs*, chap. Fo generation with a database of natural F0 patterns and with a neural network, p. 287–304. Elsevier B.V., 1992.

M. Unser : Splines, a perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6):22–38, 1999.

H. Valbret, E. Moulines et J. Tubach : Voice transformation using psola technique. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, p. 145–148, 1992.

J. P. van Santen, T. Mishra et E. Klabbers : Estimating phrase curves in the general superpositional intonation model. *In Proceedings of the 5th ISCA Speech Synthesis Workshop*, p. 61–66, Pittsburgh, June 2004.

G. Vannier : *Etude des contributions des structures textuelles et syntaxiques pour la prosodie : application à un système de synthèse vocale à partir du texte.* Thèse, Université de Caen, 1999.

W. Voiers : Toward the development of practical methods of evaluating speaker recognizability. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, p. 793–796, 1979.

C. W. Wightman : Tobi or not tobi ? *In Proceedings of Speech Prosody 2002*, p. 25–29, Aix-en-Provence, France, April 2002.

P. C. Woodland et S. J. Young : The htk continuous speech recognizer. *In Proceedings of the Third European Conference on Speech Communication and Technology (Eurospeech)*, p. 2207–2219, 1993.

Y. Xu et Q. E. Wang : Pitch targets and their realization : Evidence from mandarin chinese. *Speech Communication*, 33(4):319–337, March 2001.

Y. Yamashita, T. Ishida et K. Shimadera : A Stochastic F0 Contour Model Based on Clustering and a Probabilistic Measure. *In IEICE Transactions on Information and Systems*, vol. E86-D, p. 543–549, mars 2003.

B. Zellner : Caractérisation du débit de parole en français. *In Acte des XXIIèmes Journées d ?Etude sur la Parole (JEP 98)*, 1998.

Table des figures

1.1 Signal et spectrogramme de « un beau tir groupé » 10
1.2 Représentation spectrale d'un son pur et d'un son complexe 12
1.3 Coupe sagittale présentant les principaux acteurs de la phonation 14
1.4 Vue des cordes vocales 15
1.5 La vibration des cordes vocales 16
1.6 Les dix intonations de base de Delattre 20
1.7 Spectrogramme de la phrase : « Le public est ému par Debussy » 24

2.1 Phase d'apprentissage d'une fonction de transformation de la voix 32
2.2 Phase d'utilisation d'une fonction de transformation de la voix 33
2.3 Architecture générale d'un système de synthèse de la parole à partir du texte. 37
2.4 Phase d'apprentissage d'un système de vérification du locuteur 47
2.5 Phase de test d'un système de vérification du locuteur 47

3.1 Exemple de transcription réalisée à l'aide de ToBI. 56
3.2 Exemple de séquence de symboles INTSINT. 57
3.3 Fonction d'approximation d'un contour de F_0 avec le modèle PaIntE. . 58
3.4 Prototype d'un accent de sommet pour le modèle R/F/C. 60
3.5 Courbe de F_0 et courbe spline quadratique obtenue avec MoMel. . . . 63
3.6 Diagramme fonctionnel du modèle de Fujisaki. 64
3.7 Forme des commandes de groupe du modèle de Fujisaki. 65
3.8 Forme des commandes d'accent du modèle de Fujisaki. 66
3.9 Rôle du muscle cricothyroïdien dans les mouvements du cartilage thyroïdien. 67

4.1 Exemple de transformation du F_0 par normalisation gaussienne. 75
4.2 Exemple de scatterplot pour deux locuteurs. 76

4.3 Emplacement des points de mesure pour la transformation de Gillett et King. 78
4.4 Diagramme de flot pour la transformation par codebook. 80
4.5 Illustration des *pitch targets* et de leur réalisation de surface. 83
4.6 Système de transformation de la prosodie reposant sur CART. 84

5.1 Exemples de bases de fonctions B-splines cubiques. 101
5.2 Influence de l'ordre de multiplicité des nœuds sur les fonctions de la base de B-splines cubiques. 102
5.3 Exemple de courbe B-spline cubique avec ses points de contrôle 103
5.4 Comparaison entre spline et B-spline 114
5.5 Évolution de l'erreur RMS moyenne en fonction du nombre normalisé de d.d.l. 115
5.6 Estimation du modèle B-spline pour un contour mélodique 118
5.7 Estimation du modèle B-spline d'un contour mélodique 119
5.8 Évolution des intervalles de confiance à 99% pour l'erreur RMS en fonction du nombre normalisé de d.d.l. 120
5.9 Évolution des intervalles de confiance à 99% pour l'erreur RMS moyenne en fonction de $\log \varepsilon$ selon le critère (a). 121
5.10 Évolution des intervalles de confiance à 99% pour le nombre normalisé de d.d.l. en fonction de $\log \varepsilon$ selon le critère (a). 122

6.1 Structure d'un HMM M_j 127
6.2 Exemple de contour de F_0 avec la trajectoire du HMM associé à sa classe 133
6.3 Trajectoires des HMM pour un partitionnement en 16 classes. 136
6.4 Évolution de l'erreur RMS pour les quatre critères de sélection dans le cas split-1 138
6.5 Évolution du nombre de HMM pour les quatre critères de sélection dans le cas split-1 139
6.6 Évolution du nombre de HMM pour les quatre critères de sélection dans le cas split-1 140

7.1 Stylisation de la prosodie d'une syllabe par un vecteur de dimension 6 . 146
7.2 Architecture du système de transformation avec adaptation du GMM source aux données cible. 147
7.3 Génération du contour modélisé avec le GMM source 153
7.4 Projection des densités de probabilité des GMM source, cible et adapté. 155

7.5 Génération du contour modélisé avec le GMM source 157
7.6 Génération du contour modélisé avec le GMM cible 157
7.7 Projection des densités de probabilité des GMM source, adapté et cible . 158
7.8 Exemple de transformation d'un contour mélodique 160

Résumé

Les travaux de cette thèse se situent dans le cadre de la transformation de la prosodie en se focalisant sur la fréquence fondamentale, F_0, facteur jugé proéminent dans le traitement de la prosodie. En particulier, nous nous intéressons aux différentes étapes nécessaires à la construction d'un tel système. La première est de représenter les contours mélodiques d'un locuteur. La deuxième est d'avoir une vue de l'ensemble de l'espace mélodique de ce locuteur grâce à une procédure de classification. Enfin, la troisième est de poser une fonction de transformation de la prosodie du locuteur source vers celle du locuteur cible. Pour chaque étape, nous proposons une méthodologie qui tient compte des problèmes qui se sont posés à l'étape précédente. Un modèle B-spline est proposé pour la stylisation des contours mélodiques. Cette approche permet de modéliser les discontinuités des contours mélodiques et de sélectionner automatiquement le nombre de paramètres du modèle. Pour représenter l'espace mélodique du locuteur, une approche par modèles de Markov est introduite afin de regrouper les contours mélodiques de formes similaires indépendamment de leur durée. Enfin, une méthodologie de transformation de la prosodie à partir de corpus non parallèles par une technique d'adaptation au locuteur est présentée. Les résultats obtenus en terme de transformation du F_0 sont encourageants et tendent à montrer qu'il est nécessaire de traiter la dynamique du F_0 et de piloter la transformation par des informations d'ordre morphosyntaxique.

Mots-clés : Informatique, Synthèse automatique de la parole, Prosodie, Apprentissage automatique, Classification automatique, Processus de Markov, Théorie des Splines

Abstract

The work presented in this thesis lies within the scope of prosody conversion and more particularly the fundamental frequency conversion which is considered as a proeminent factor in prosody processing. This document deals with the different steps necessary to build such a conversion system. The first one is to represent the melodic contours of a speaker. The second one is to cluster his melodic contours to obtain a global view of the melodic space. The last one is to choose a conversion function that enables the transformation from the source prosody to the target one. For each step, we propose a methodology which takes into account the issues and difficulties encountered in the previous one. A B-spline model is considered to model the melodic contours. This approach enables to model the irregularities of the contours and to select automatically the model number of parameters. To represent the melodic space of a speaker, a HMM based approach is introduced to group melodic contours with similar shapes despite their different lengths. To finish, a prosody transformation methodology using non-parallel corpora based on a speaker adaptation technique is derived. The results we obtain are encouraging and tend to show that it is necessary to model the evolution of the melody and to drive the transformation system by using morpho-syntactic information.

Keywords : Computer science, Speech synthesis, Prosodic analysis, Machine learning, Automatic classification, Markov processes, Spline theory