Département de formation doctorale en informatique **École doctorale IAE + M**

UFR STMIA

Modèles de production de parole et reconnaissance à partir d'automates

THÈSE

présentée et soutenue publiquement le 3 décembre 1999

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1

(spécialité informatique)

par

Bruno Mathieu

Composition du jury

Président :	Jean-Paul HATON	Professeur, Université Henri Poincaré
Rapporteurs :	Shinji MAEDA	Directeur de recherche CNRS, ENST
	Jean CAELEN	Directeur de recherche CNRS, CLIPS
Examinateurs :	Didier GALMICHE	Maître de conférences, Université Henri Poincaré
	Jean-Marie PIERREL	Professeur, Université Henri Poincaré
	Yves LAPRIE	Chargé de recherche CNRS, LORIA

Centre de Recherche en Informatique de Nancy

Mis en page avec la classe TheseCRIN.

Résumé

L'objet de cette thèse est l'étude de la reconnaissance automatique de parole. Ce document débute avec une description des traitements acoustiques les plus répandus en vue de reconnaître la parole. Nous décrivons ensuite les diverses architectures qui ont été utilisées : comparaison dynamique de formes acoustiques, systèmes experts, réseaux neuro-mimétiques et modèles de Markov. Puis ce document se divise en deux parties.

Dans une première expérience, nous nous sommes intéressés à la reconnaissance de mots. Pour cela, nous utilisons des automates qui modélisent le vocabulaire. Celui-ci comporte les dix chiffres anglo-saxons, dont deux prononciations différentes pour le zéro. Le corpus de parole TiDigits a été utilisé par d'autres laboratoires ce qui nous permet de comparer nos résultats. La première étape est consacrée à la reconnaissance de mots isolés. Puis nous présentons une méthode de segmentation de séquences de chiffres. La fin de ce chapitre est consacrée à la reconnaissance de mots enchaînés et à une discussion sur les mérites et les faiblesses de notre approche.

La deuxième partie traite de l'utilisation d'un modèle de production qui pourrait être utilisé pour le reconnaissance de la parole. Nous commençons par présenter les équations acoustiques régissant l'écoulement de l'air dans le conduit vocal et divers modèles articulatoires. Ensuite nous justifions le choix du modèle articulatoire de Maeda. Nous décrivons comment nous avons adapté le modèle à un locuteur masculin. Puis nous présentons la méthode variationnelle utilisée pour retrouver les trajectoires des articulateurs en fonction de la parole prononcée. Une dernière section présente les logiciels réalisés.

En conclusion, nous résumons les résultats obtenus et donnons quelques perspectives en vue de reconnaître la parole continue quel que soit le locuteur.

Mots-clés: reconnaissance de parole, modèles de production de parole, modèle de Maeda, inversion acoustique-articulatoire.

Abstract

The framework of this thesis is speaker-independent automatic speech recognition. This document begins with a short survey of speech signal processing applied to speech recognition. Then, we describe several common architectures: dynamic time warping of acoustic shapes, artificial intelligence, neural networks and hidden Markov models. This document is made of two main parts.

The first part is devoted to the recognition of words. We are using finite state automata for modeling the eleven American spoken digits. As the speech database TiDigits has been used by other teams we can compare our results against thoose obtained with other approaches. The first step is concerned with isolated word recognition. Then, we describe how sentences of the database have been segmented. Last sections of this part describe continuous speech recognition of word sequences, as well as a discussion about strong and weak points of our approach.

The second part treats of the use of production models for speech recognition. We begin with a description of the acoustic equations which drive the air flow inside the vocal tract and we present several articulatory models. Then, we justify the choice of Maeda's model. We describe the adaptation of this model to a male speaker. Next, we describe the variational method used for recovering articulatory trajectories from the speech. Finally, the software we built, is described.

In the conclusion, we give the results obtained and we exhibit some perspectives for future work towards a better speaker indepedent continuous speech recognition system.

Remerciements

Je tiens à remercier les personnes du laboratoire LORIA pour leur accueil.

Je remercie en particulier Joseph Di Martino, Yves Laprie et Jean-Marie Pierrel qui m'encadrèrent et m'accueillirent au sein de leurs équipes.

Je m'incline aussi devant la compétence et la disponibilité de l'équipe système qui nous offrit d'excellentes conditions de travail.

Mes remerciement s'adressent ensuite à Shinji Maeda, Jean Caelen, Didier Galmiche et Jean-Paul Haton qui me font l'honneur d'être membres du jury.

Ma gratitude s'adresse aussi aux collègues croisés lors de conférences, en particulier Laurence Candille, Véronique Lecuit, Martine George, Alain Soquet et les participants à SPECOM-96 dont Юля Дроздова.

Je remercie enfin mes amis et ma famille qui ont su m'encourager lorsque cela était nécessaire.

À Carly A. Thrush.

Table des matières

Chapitre 1 Introduction **1**

Chapitre 2 Etat de l'art **5**

2.1 Les sons de la parole 6

2.2 Traitements acoustiques 7

2.2.1 Transformée de Fourier 7

2.2.2 Analyse par prédiction linéaire 10

2.2.3 Analyse par transformée homomorphique 11

2.2.4 Analyses dérivées de LPC 12

2.2.5 Conclusion 12

2.3 Architectures pour la reconnaissance 13

2.3.1 La comparaison de formes 13

2.3.2 Les modèles à base de connaissances 14

2.3.3 Les modèles neuromimétiques 16

2.3.4 Les modèles de Markov 18

Chapitre 3 Reconnaissance de mots à l'aide d'automates **23**

3.1 Reconnaissance de mots isolés 23

3.1.1 Introduction 23

3.1.2 Apprentissage 24

3.1.3 Reconnaissance 27

3.1.4 Conclusion 34

3.2 Reconnaissance de mots enchaînés 35

3.2.1 Modification de l'algorithme de reconnaissance 35

3.2.2 Nouvel apprentissage 37

3.2.3 Evaluation du taux de reconnaissance 37

3.2.4 Segmentation 40
3.2.5 Problèmes rencontrés lors de la segmentation 40
3.2.6 Résultats 41
3.2.7 Prise en compte du contexte droit 41
3.2.8 Conclusion 43

Chapitre 4 Utilisation de modèles de production 47

4.1 Introduction 47
4.2 La production de la parole 48
4.2.1 Le conduit vocal 49
4.2.2 Les équations acoustiques 50
4.2.3 Méthodes de résolution 54
4.2.4 Conclusion 56
4.3 Différents modèles articulatoires 56
4.3.1 Introduction : Insuffisance des modèles de la fonction d'aire pour la reconnaissance 57
4.3.2 Modèles physiques 58
4.3.3 Modèles géométriques 58
4.3.4 Modèles biomécaniques 59
4.3.5 Conclusion 59
4.4 Contrôle des paramètres 60
4.5 Inversion acoustique-articulatoire 61
4.5.1 Inversion par tabulation 61
4.5.2 Inversion par optimisation 61
4.6 Inversion acoustique-articulatoire dans le modèle de Maeda 62
4.6.1 Le modèle de Maeda 62
4.6.2 Acquisition des images et de la parole 63
4.6.3 Adaptation des échelles et estimation du contour extérieur . . . 65
4.6.4 Fidélité du modèle 65
4.6.5 Couverture de l'espace vocalique du locuteur 66
4.6.6 Adaptation de la paroi dure 69
4.6.7 Inversion 70
4.6.8 Conclusion 79
4.6.9 Réalisations logicielles 80

Chapitre 5 Conclusion et perspectives **83**

Annexes **87**

Bibliographie **93**

Table des figures

2.1 Signal temporel de parole 7
2.2 Zoom sur le signal 7
2.3 Fenêtre 8
2.4 Spectrogramme à bande large 9
2.5 Spectrogramme à bande étroite 9
2.6 Réseau 13
2.7 Construction de la base de règles. 14
2.8 Principe de la reconnaissance avec un système à base de connaissances. 15
2.9 Architecture neuronale. 16
2.10 Réseau récurrent de Jordan. 18
2.11 Modèle ergodique. 19
2.12 Modèle gauche-droite. 20

3.1 Construction d'un dictionnaire et codage d'une locution. 26
3.2 En haut deux locutions, en bas l'automate correspondant. 27
3.3 Recalage temporel. 29
3.4 Contraintes locales : asymétrique, symétrique, de Sakoe et Chiba symétrique. 30
3.5 Les contraintes utilisées et un exemple de chemin. 30
3.6 Principe de la reconnaissance de mots isolés 31
3.7 Modèle des séquences de chiffres 36
3.8 Principe de la reconnaissance de mots enchaînés 37
3.9 Contraintes pour la segmentation 40

4.1 Coupe sagittale du conduit vocal (voyelle /i/) 49

4.2 Elément de circuit électrique 55
4.3 Modèle articulatoire original sur sa grille semi-polaire 63
4.4 Espace acoustique F1-F2 après modification des facteurs d'échelle. Les losanges représentent les voyelles prononcées par notre sujet ; les points représentent l'échantillonnage aléatoire de l'espace articulatoire. 67
4.5 Espace acoustique F2-F3 après modification des facteurs d'échelle . . . 67
4.6 Espace acoustique F1-F2 après ré-estimation de la paroi dure 68
4.7 Espace acoustique F2-F3 après ré-estimation de la paroi dure 68
4.8 Espace acoustique F1-F2 finalement obtenu 70
4.9 Espace acoustique F2-F3 finalement obtenu 71
4.10 Ensemble de trajectoires possibles pour la position de la langue lors d'une transition /iu/ avec le premier dictionnaire 73
4.11 Ensemble de trajectoires possibles pour la position de la langue lors d'une transition /iu/ avec le deuxième dictionnaire 74
4.12 Trajectoires articulatoires pour /iai/ 77
4.13 Trajectoires formantiques pour /iai/ 78
4.14 Trajectoires articulatoires pour /iui/ 78
4.15 Trajectoires formantiques pour /iui/ 79
4.16 Snorri 80
4.17 Fenêtre principale 81
4.18 Coupe sagittale, spectre et formants 81

Chapitre 1

Introduction

Cette thèse est consacrée à la reconnaissance automatique de la parole. Elle invite le lecteur à un périple à travers ce vaste domaine de recherche. Une thèse ne suffit malheureusement pas à parcourir ce thème dans ses moindres détails. D'une part, la reconnaissance de la parole touche à de nombreux domaines comme le codage et la synthèse de parole mais aussi le traitement du signal, la phonétique, la phonologie, l'intelligence artificielle, l'apprentissage... Dans un premier temps nous délimiterons donc la région à parcourir. D'autre part ce document ne peut pas couvrir exhaustivement l'étendue des travaux entrepris dans ce domaine depuis plusieurs dizaines d'années. Ces recherches ont encore un bel avenir devant elles et cette thèse atteindra son but si elle suscite chez le lecteur l'envie de poursuivre l'étude de la reconnaissance de la parole.

Délimitons maintenant les frontières de la région à explorer. D'une part la parole est un canal de communication qui transmet beaucoup d'informations et pas uniquement une suite de phonèmes. Parmi elles, il y en a qui concernent le locuteur comme son sexe, son état émotionnel, son âge, son accent... La reconnaissance et l'identification du locuteur [24] utilisent ces informations entre autres. La parole transmet aussi une suite de mots dans une certaine langue. On peut donc créer des applications qui déterminent la langue du locuteur [72, 80, 43], ainsi que des logiciels de dictée automatique comme DragonDictate.... Mais elle est aussi le moyen de transmettre ses intentions, exprimées à l'aide de mots et de l'intonation donnée à la phrase. On touche là au domaine de la compréhension de la parole qui doit permettre un véritable dialogue oral entre l'homme et la machine. On peut considérer les applications de commande vocale où l'homme

donne des ordres à la machine souvent exprimés sous forme de verbes à l'impératif comme des cas dégénérés de ce problème.

Cette thèse abordera les domaines de la commande et de la dictée automatique, à travers les deux expériences décrites respectivement dans les chapitres 3 et 4. La première expérience porte sur la reconnaissance de mots. La deuxième s'apparente au codage articulatoire évoqué par Flanagan [29] « Consider a phonemic transmission of speech; that is, the written equivalent of the meaningfully distinctive sounds of speech. Take english for example... The written equivalent of the information is therefore less than 50 bits/s ». Dans cette expérience, nous essayons de retrouver la dynamique des articulateurs à partir du signal acoustique et de l'utilisation d'un modèle articulatoire qui exploite explicitement le fait que la parole est produite par un conduit vocal.

Les difficultés rencontrées lors de la reconnaissance de parole peuvent être classées selon les axes suivants :

- la taille du vocabulaire et sa complexité (mots phonétiquement proches, par exemple l'e-set),
- le niveau de bruit ambiant (SNR : Signal to Noise Ratio, rapport signal/bruit), l'existence d'un modèle du bruit,
- le nombre de locuteurs à reconnaître (mono, multi, omni), leurs caractéristiques (sexe, âge, accents régionaux...).

Dans les deux expériences, le milieu n'est pas bruité. La reconnaissance des mots utilise un petit vocabulaire et veut être indépendante du locuteur. La deuxième étude porte sur le décodage « acoustico-phonétique » en utilisant un modèle articulatoire. Celui-ci est propre à un locuteur. Aucun vocabulaire n'est utilisé : des problèmes de représentation du lexique en fonction de variables articulatoires sont en effet à prendre en compte. Ceci devrait permettre de reconnaître à terme un très grand vocabulaire. Aussi, bien que le modèle soit dédié à un locuteur particulier, il est possible d'adapter un modèle articulatoire à partir du signal acoustique comme Payan et Perrier [79]. On peut aussi envisager d'utiliser plusieurs modèles (pour les hommes, les femmes ou les enfants) simultanément.

Cette thèse comporte cinq chapitres. Après cette introduction nous tenterons de présenter les techniques les plus utilisées en reconnaissance de parole : il s'agit de dif-

férents traitements appliqués au signal acoustique ainsi que des multiples architectures parfois combinées permettant de reconnaître des mots ou de faire du décodage acoustico-phonétique. Là encore il faut remarquer que l'apparition de logiciels de reconnaissance sur le marché est due en partie aux avancées concernant la puissance du matériel informatique et aux progrès de la modélisation du langage [75]. Le chapitre suivant présentera une expérience de reconnaissance de mots isolés et enchaînés multi-locuteurs pour un vocabulaire de petite taille en utilisant des automates d'états finis. Inspirée des travaux de Joseph Di Martino [62], cette expérience a donné de bons résultats en reconnaissance de mots isolés. Nous tenterons d'expliquer pour quelles raisons la reconnaissance de mots enchaînés n'a pas conduit à des résultats comparables à ceux obtenus par l'utilisation de Modèles de Markov Cachés. Nous présenterons dans le chapitre suivant la seconde expérience qui tente de retrouver la dynamique du conduit vocal (c'est-à-dire les mouvements des articulateurs, ou gestes articulatoires) à partir du signal acoustique afin d'utiliser des variables articulatoires au lieu de trames acoustiques en entrée pour les architectures classiques (Modèles de Markov, Réseaux neuronaux, systèmes à base de connaissances). En effet nous pensons que les sons de la parole sont souvent mieux définis dans le domaine articulatoire que dans le domaine acoustique. Pour conclure nous tirerons un bilan des deux expériences et donnerons des pistes permettant d'améliorer les résultats obtenus.

Chapitre 2

Etat de l'art

Si des logiciels comme Dragon Dictate, ViaVoice, VoiceType, Locus... apparaissent sur le marché, cela est dû non seulement aux avancées en traitement du signal et dans les architectures utilisées en reconnaissance de parole, mais aussi à la croissance du rapport performances/coût du matériel informatique et à une meilleure modélisation du langage. En effet l'apprentissage de modèles de Markov ou de réseaux neuronaux nécessite une importante puissance de calcul ainsi qu'une grande capacité mémoire. Par ailleurs, un modèle de langage peut lever certaines ambiguïtés qui persistent même après un décodage acoustico-phonétique correct (par exemple des homophones comme /la/ peuvent représenter *la*, *là* ou encore *las*, et des homonymes comme *la* peuvent signifier l'article défini ou la note de musique). En outre il peut aussi corriger certaines erreurs du décodage en utilisant des contraintes lexicales, syntaxiques, sémantiques ou même pragmatiques.

Néanmoins, nous nous focaliserons dans cette partie sur les traitements apportés au signal acoustique puis sur les architectures les plus utilisées en reconnaissance de parole. Dans un premier temps nous allons décrire l'objet des traitements : c'est-à-dire la parole. Puis nous présenterons les différentes transformations effectuées sur celle-ci soit pour la représenter dans un domaine mieux approprié pour la reconnaissance, soit pour en extraire localement des indices permettant de classifier des segments de signal. Enfin nous présenterons les architectures utilisant ces représentations ou ces indices pour reconnaître la parole. En particulier nous nous intéresserons au principe de la programmation dynamique, aux Modèles de Markov, aux Réseaux Neuronaux, et aux

Systèmes à Base de Connaissances, sachant que ces architectures peuvent être couplées.

2.1 Les sons de la parole

Tout d'abord si le vocabulaire est de petite taille, comme par exemple dans le cas de la commande vocale, on peut envisager de modéliser indépendamment chaque mot. En revanche dès qu'il s'agit de reconnaître un grand vocabulaire—pour la dictée automatique par exemple—la taille du vocabulaire nécessite un découpage plus fin : syllabes, diphones, phones. L'alphabet phonétique international (IPA) donne une classification des différents sons. Chaque classe est appelée un phonème. Cette classification est utilisée pour donner une représentation acoustique de chaque entité lexicale. Ainsi un mot est transcrit par une suite de phonèmes adjacents. Cependant les mots sont prononcés différemment selon l'accent du locuteur, le contexte dans lequel se situe le mot... Pour pallier ces phénomènes on peut adopter une représentation sous forme de treillis phonétique pour chaque mot. Mais différents experts ne segmentent pas la parole de la même manière. Non seulement les frontières des mots, mais aussi la classe phonétique d'un son différent selon l'expert, en particulier lorsque celui-ci doit transcrire une phrase d'une langue différente de sa langue maternelle [22].

En fait le phonème n'a de réalité ni dans le domaine acoustique, ni dans le domaine articulatoire. C'est pourquoi on envisage de représenter les sons selon la manière dont ils sont produits (théories motrices). On définit ainsi le mode d'articulation lié aux sources (glotte, bruits de friction et d'explosion), l'organe articulant (langue, lèvres, luette), le lieu d'articulation (endroit où le son « se forme »). Celui-ci est intéressant puisqu'il permet par exemple de différencier les occlusives (/p/ /k/ et /t/, /b/ /g/ et /d/) qui sont plus difficiles à différencier dans le domaine acoustique. La variabilité interlocuteurs peut être prise en compte par l'utilisation de plusieurs modèles tout comme on le fait à partir d'une étude acoustique du signal. En revanche les causes de la variabilité intralocuteurs, dues au contexte et aux phénomènes de coarticulation (coarticulation anticipatoire : l'articulateur anticipe son déplacement vers le prochain lieu d'articulation, de rétention : le lieu d'articulation n'est pas atteint à cause de l'inertie de l'articulateur) sont certainement plus faciles à appréhender dans le domaine articulatoire.

Dans la section suivante nous décrivons les traitements du signal acoustique souvent utilisés pour la reconnaissance de la parole.

2.2 Traitements acoustiques

Les différents traitements effectués sur le signal de parole ont pour but :

- d'en donner une représentation plus concise, en conservant les informations les plus pertinentes vis-à-vis de la tâche à accomplir. Les traitements sont différents dans le cas du codage, de l'étude de la prosodie, d'une application de commande vocale ou de dictée automatique,
- d'extraire des indices permettant de segmenter la parole et de classifier les segments acoustiques obtenus.

Nous verrons dans cette section les traitements les plus répandus : transformée de Fourier, analyse par prédiction linéaire, analyse cepstrale.

2.2.1 Transformée de Fourier

Tout d'abord, la figure 2.1 représente une vue d'un signal de parole. La figure 2.2 en représente un segment à une échelle plus fine. Le signal présente une certaine périodicité et stationnarité à court terme. D'où l'idée de le représenter dans le domaine fréquentiel.

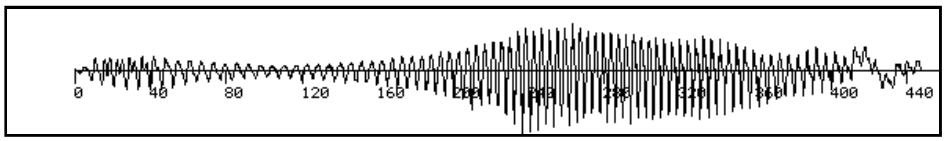


FIG. 2.1 – *Signal temporel de parole*

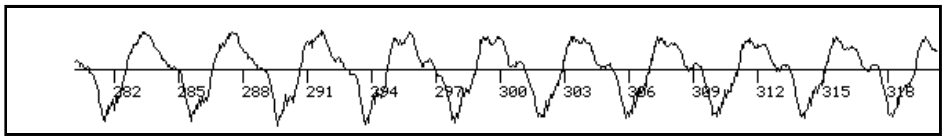


FIG. 2.2 – *Zoom sur le signal*

La transformée de Fourier d'un signal $s(t)$ est définie par $F(\omega) = [\Im(s)](\omega)] = \int_{\Re} s(t)e^{-j\omega t}\,\mathrm{d}t$ où ω est la pulsation complexe. Cependant, comme les sons ont une durée

limitée, il importe plus de connaître le spectre à court terme, c'est-à-dire les caractéristiques spectrales de la parole à un instant donné qui doivent permettre d'identifier les différents sons produits au cours du temps. En outre, il est impossible de connaître le signal de parole entre $-\infty$ et $+\infty$ (serait-ce intégrable?). Enfin, on ne peut considérer la parole comme étant un signal périodique stationnaire que sur une durée de quelques millisecondes. C'est pourquoi on a défini la transformée de Fourier à court terme du signal comme l'analyse d'une portion de signal vue au travers d'une fenêtre temporelle $h(t)$ comme celle représentée par la figure 2.3:

$$\Im(\omega, t) = \int_{-\infty}^{+\infty} s(\tau)h(t-\tau)e^{-j\omega\tau}\,\mathrm{d}\tau = \Im(s_t) \star \Im(h)$$

FIG. 2.3 – *Fenêtre*

Le prélèvement d'une partie du signal par une fenêtre induit donc une imprécision sur la valeur exacte de la transformée de Fourier du signal à une fréquence donnée. Ceci conduit à deux types de spectrogramme :

- le spectre bande large. Dans ce cas peu de points sont prélevés du signal, par conséquent la convolution intègre la « vraie » valeur du spectre sur des bandes de fréquence larges (cf. figure 2.4),
- le spectre bande étroite. Plus de points sont prélevés, le spectre de la fenêtre s'approche de la distribution de Dirac, donc la bande spectrale est de plus en plus fine (cf. figure 2.5).

En outre le signal est échantillonné et on ne le connaît qu'en certains points multiples de la période d'échantillonnage. Soit donc $\prod_\nu$ le peigne de Dirac pour la fréquence ν, $s_e = s.\prod_\nu$ est le signal échantillonné. La transformée de Fourier d'un produit étant égale au produit de convolution des facteurs, on a $\Im(s_e) = \Im(s) \star \Im(\prod_\nu)$. Par conséquent les valeurs des transformées de Fourier de s et de s_e sont égales au facteur v près pour

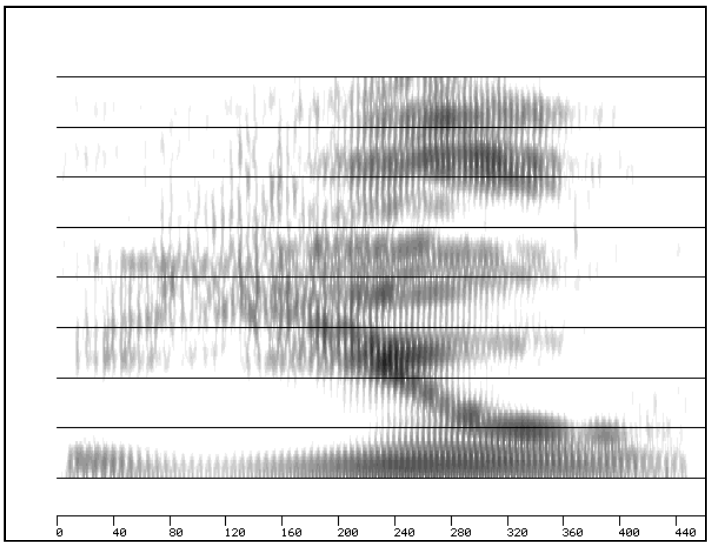


FIG. 2.4 – *Spectrogramme à bande large*

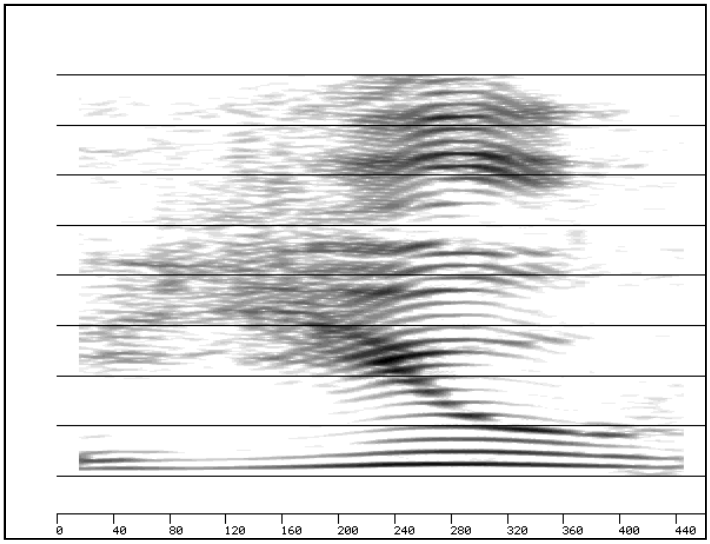


FIG. 2.5 – *Spectrogramme à bande étroite*

les valeurs de ω multiples de ν. L'échantillonnage induit donc une périodicité de la transformée de Fourier.

Ces deux représentations sont complémentaires. En effet, le spectre à bande large possède une meilleure définition temporelle et permet d'identifier les fréquences de résonance du conduit vocal (larges bandes foncées) plus précisément que sur le spectre à bande étroite. Cependant celui-ci, grâce à sa meilleure définition fréquentielle, fait apparaître les harmoniques (fines bandes foncées). Il est donc plus utile pour étudier l'influence de la source glottale.

2.2.2 Analyse par prédiction linéaire

Cette technique est fort utilisée dans le codage de parole. Pour la reconnaissance de la parole, elle se justifie en considérant le modèle de parole simplifié source/conduit. Dans ce modèle, le signal de parole y vaut $y_n = \sum_k a_k.y_{n-k} + Ge_n$. La transformée en z conduit à $Y(z)\left(1 - \sum_k a_k.z^{-k}\right) = GE(z)$. La fonction de transfert du conduit est donc égale à

$$H(z) = \frac{Y(z)}{E(z)} = \frac{G}{1 - \sum_k a_k.z^{-k}} \tag{2.1}$$

L'analyse par prédiction linéaire conduit à estimer les coefficients de la fonction de transfert du conduit vocal. Elle estime l'échantillon courant $\widetilde{y_n}$ en fonction des K échantillons précédents :

$$\widetilde{y}_n = \sum_{k=1}^{K} a_k.y_{n-k}$$

Soit $e_{n=}y_n - \widetilde{y_n}$ l'erreur commise en estimant y_n par $\widetilde{y_n}$. Le but est de trouver les coefficients a_k minimisant $E = \sum_n e_n^2$.

En dérivant l'erreur de prédiction E par rapport aux a_k on obtient $\frac{\partial E}{\partial a_l} = -2\sum_n y_n.y_{n-l} + 2\sum_n(\sum_{k=1}^{K} a_k.y_{n-k}).y_{n-l}$. Après réorganisation des termes, et en posant $x_{k,l} = \sum_n y_{n-k}.y_{n-l}$, on obtient les équations de Yule-Walker :

$$x_{l,0} = \sum_{k=1}^{K} a_k.x_{k,l}$$

Pour résoudre ces équations, on prend une fenêtre dans laquelle n varie. Ensuite deux manières permettent de construire les $x_{k,l}$:

- Soit on n'utilise que des y_k à l'intérieur de la fenêtre en fixant $y_k = 0$ en dehors.

On obtient alors le calcul par la méthode d'autocorrélation. Dans ce cas, les coefficients d'autocorrélation $x_{k,l} = x_{l,k}$ peuvent être obtenus soit directement à partir du signal temporel, soit comme transformée de Fourier inverse du spectre de puissance obtenu, par exemple, à partir d'une simulation dans le domaine fréquentiel. La résolution des équations de Yule-Walker s'effectue en utilisant l'algorithme de Durbin [83].

– Soit on conserve les vraies valeurs de y_k à l'extérieur de la fenêtre où n varie, et on résout le système par la méthode de covariance. Pour cela il faut inverser la matrice symétrique des $x_{k,l}$.

À partir des coefficients ainsi calculés, il est possible de rechercher les racines du dénominateur de l'équation 2.1 qui correspondent aux modes de résonance du conduit vocal (formants). Or les trois premières fréquences de résonance sont suffisantes pour caractériser les voyelles du français. Cette analyse est donc utilisée pour paramétrer le signal acoustique afin d'alimenter une architecture de reconnaissance, mais aussi pour l'estimation précise des valeurs des fréquences des formants.

2.2.3 Analyse par transformée homomorphique

Cette méthode a pour but de séparer les contributions du conduit vocal et des sources. D'après le modèle source/conduit, la parole est le résultat d'une convolution de la source par le filtre constitué du conduit vocal et de la radiation aux lèvres : $y(t) = s(t) \star c(t)$. Le but est de déconvoluer ce produit de convolution de manière à séparer les contributions respectives de la source et du conduit.

Pour cela on utilise la propriété de la transformée de Fourier : la transformée de Fourier d'un produit de convolution est égale au produit des transformées de Fourier des facteurs. Puis on utilise la propriété du logarithme : le logarithme d'un produit est égal à la somme des logarithmes facteurs. On retourne dans un domaine pseudo-temporel par une transformée de Fourier inverse, la transformée de Fourier inverse d'une somme étant égale à la somme des transformées de Fourier inverse des termes, il résulte alors de ces transformations une contribution additive de la source et du conduit. La fonction obtenue est appelée cepstre et le domaine dans laquelle elle prend ses valeurs est appelé domaine quéfrentiel. Pour les sons voisés, le cepstre présente un pic à une distance égale

à la période fondamentale F_0. En revanche, la contribution de la source est localisée principalement dans les premiers coefficients du cepstre. On peut alors appliquer un filtre pour séparer les premiers coefficients des derniers : on appelle ce filtrage liftrage. La réponse impulsionnelle du conduit vocal peut s'obtenir en appliquant un filtre qui supprime les derniers coefficients, puis en appliquant successivement une transformée de Fourier, une exponentiation et une transformée de Fourier inverse. Les coefficients cepstraux calculés sur une échelle Mel ou Bark peuvent servir directement à alimenter une architecture de reconnaissance de la parole.

2.2.4 Analyses dérivées de LPC

Des considérations sur la perception conduisent à des variantes des analyses purement mathématiques précédentes. Par exemple, l'analyse par prédiction linéaire perceptivement fondée est basée sur l'analyse par prédiction linéaire dans laquelle on utilise des critères psychoacoustiques afin d'obtenir un spectre auditif [38, 2].

L'analyse RASTA-PLP a pour but d'améliorer la robustesse du système de reconnaissance en milieu bruité. Elle est basée sur l'analyse PLP [39, 2].

2.2.5 Conclusion

Différentes méthodes permettent d'analyser le signal acoustique. Cette analyse est différente selon l'architecture de reconnaissance qu'elle alimente. Pour les systèmes à base de connaissances, on essaie d'extraire du signal acoustique les indices qu'utilisent les experts humains pour classifier les sons. Pour les architectures stochastiques, neuromimétiques basées sur le principe de la comparaison dynamique de formes acoustiques, on cherche une représentation du signal acoustique pertinente (c'est-à-dire qui conserve les caractéristiques discriminantes entres les différents sons) et concise (de manière à pouvoir effectuer les calculs en temps réel).

Nous allons maintenant présenter les architectures les plus couramment utilisées en reconnaissance automatique de la parole.

2.3 Architectures pour la reconnaissance

Cette section présentera brièvement les différentes architectures utilisées en reconnaissance de parole, depuis les débuts lorsque l'on comparait des formes acoustiques de référence à la parole à reconnaître en utilisant un algorithme de recalage temporel pour pallier les distorsions dues en particulier au débit d'élocution, jusqu'aux réseaux neuromimétiques réputés excellents pour leur pouvoir de discrimination, en passant par les systèmes à base de connaissances qui ont abouti à une meilleure compréhension de la production et de la perception de la parole et aux modèles de Markov qui donnent actuellement les meilleurs taux de reconnaissance.

2.3.1 La comparaison de formes

Il s'agit de comparer la parole à des formes acoustiques de référence. Cette technique a surtout été utilisée pour des applications de reconnaissance monolocuteur de mots. La principale source de variabilité dans ce contexte est constituée par les distorsions dues aux différences dans les débits d'élocution. Cette difficulté a été résolue en utilisant l'algorithme de recalage temporel (*Dynamic Time Warping*).

Le problème du recalage temporel est résolu en utilisant le principe de programmation dynamique. Ce principe permet de trouver le meilleur chemin dans un réseau. Pour l'appliquer à la reconnaissance de parole, les nœuds du graphe sont obtenus en faisant le produit cartésien du modèle et de la forme acoustique à reconnaître comme l'illustre la figure 2.6.

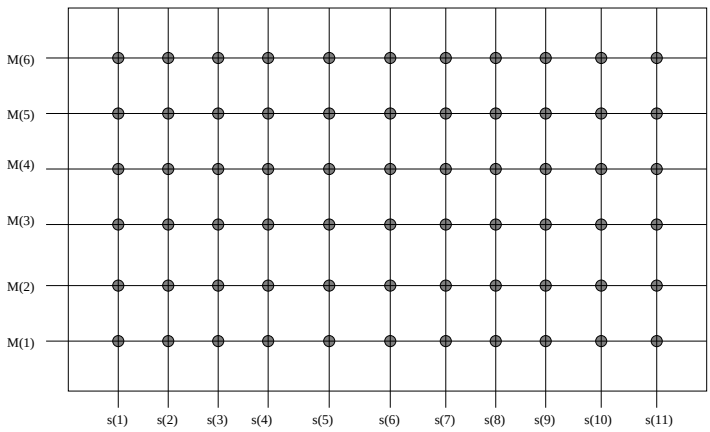


FIG. 2.6 – *Réseau*

Les arcs sont orientés pour prendre en compte l'écoulement du temps. Un coût est associé aux transitions. Le meilleur chemin est celui de coût le plus faible. En construisant n réseaux, pour autant de modèles, on obtient n chemins dont un de plus faible coût. La forme acoustique reconnue est celle qui correspond à la projection de ce chemin sur l'axe des ordonnées.

Pour reconnaître des mots enchaînés, on considère toutes les concaténations possibles des modèles. Cette méthode a été utilisée dans la première expérience de cette thèse (cf. chapitre 3).

2.3.2 Les modèles à base de connaissances

Utilisés principalement pour des applications devant traiter un grand vocabulaire, lorsqu'il n'est pas envisageable de prendre le mot comme forme acoustique de référence, ces méthodes ont beaucoup apporté en terme de compréhension de la perception et de la production de la parole. Il s'agit de construire une base de connaissances à partir de l'expertise des phonéticiens et phonologues sous forme de règles[30]. Les indices acoustiques extraits du signal sont ensuite utilisés pour construire une base de faits. Les règles pour lesquelles les prémisses sont satisfaits peuvent alors êtres appliquées pour enrichir cette base. Le principe de l'extraction des connaissances est illustré par la figure 2.7.

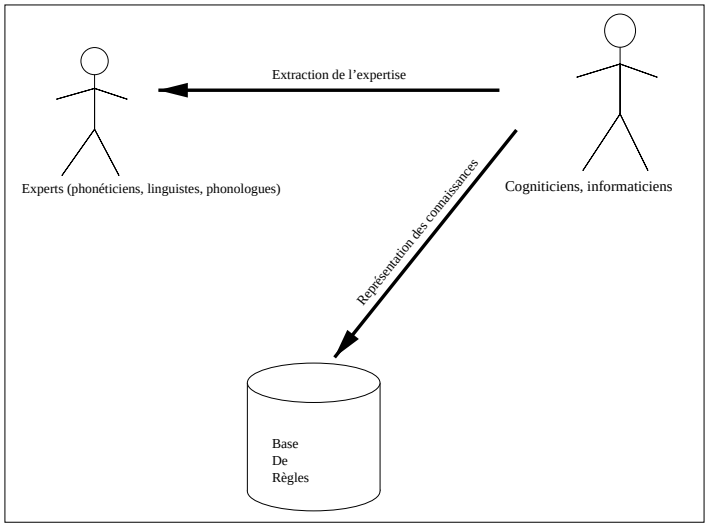


FIG. 2.7 – *Construction de la base de règles.*

Le processus de reconnaissance est schématisé sur la figure 2.8. Le traitement du signal produit des faits acoustiques. Ceux-ci servent de prémisses aux règles modélisant l'expertise humaine. L'application de ces règles produit de nouveaux faits qui enrichissent à leur tour la base. Ainsi à partir d'indices acoustiques (existence d'une barre d'explosion ayant telles propriétés, suivie par un noyau vocalique...) des faits de niveau supérieur (phonétiques, phonologiques, linguistiques) sont produits.

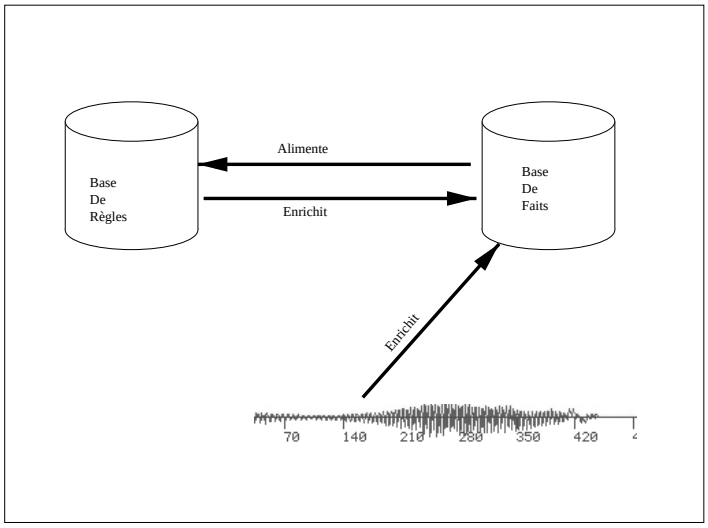


FIG. 2.8 – *Principe de la reconnaissance avec un système à base de connaissances.*

La difficulté majeure dans cette approche consiste à construire la base de règles. En effet les experts ne sont généralement pas unanimement d'accord. Par ailleurs, l'adjonction de nouvelles connaissances doit préserver la cohérence de la base, ce qui n'est pas toujours aisé à vérifier. En outre, les règles dépendent souvent de la langue ce qui réduit aussi la portée d'une telle architecture.

Néanmoins les aspects multidisciplinaires de cette approche sont intéressants et ont beaucoup apporté en terme de compréhension de ce qui caractérise la parole. En effet la comparaison dynamique de formes ne prend guère en compte les particularités du signal de parole, si ce n'est à travers les traitements acoustiques qui peuvent être fondés sur un modèle de production rudimentaire (modèle source/conduit).

2.3.3 Les modèles neuromimétiques

Les deux architectures précédentes ont un important point faible : l'apprentissage. Dans les systèmes basés sur la comparaison de formes acoustiques, le choix des références, la qualité de leur acquisition et les traitements appliqués conditionnent les résultats en reconnaissance. Dans le cas des systèmes à base de connaissances c'est la construction de la base de règles, avec les difficultés inhérentes à l'extraction du savoir-faire des phonéticiens, phonologues et linguistes [22] qui constitue l'apprentissage.

Le principe original du courant neuromimétique est de modéliser le système de perception de l'humain en construisant un réseau dont les états correspondent aux « neurones » tandis que les liens sont l'équivalent des « synapses » les reliant, par analogie avec le système nerveux humain. En effet, pour les tâches de perception, l'humain est très performant et robuste comparé aux technologies numériques actuelles. C'est pourquoi les réseaux de neurones formels sont utilisés en reconnaissance de formes en général. Cependant, ils peuvent aussi être vus comme une simple classe d'algorithmes parallèles, en oubliant les fondements biologiques qui leur ont donné naissance. La diffusion de la connaissance au travers du réseau conduit à une meilleure robustesse et l'architecture est intrinsèquement parallèle ce qui facilite les implantations efficaces. En outre, des algorithmes d'apprentissage permettent, à partir d'exemples, de distribuer la connaissance automatiquement au travers du réseau ce qui simplifie grandement le développement d'applications. Ainsi l'algorithme de rétropropagation de l'erreur [88] a suscité un regain d'intérêt pour ces architectures. Ce n'est cependant pas le seul algorithme d'apprentissage pour un réseau à couches non récurrent. Montana et Davis [69], par exemple, utilisent un algorithme génétique.

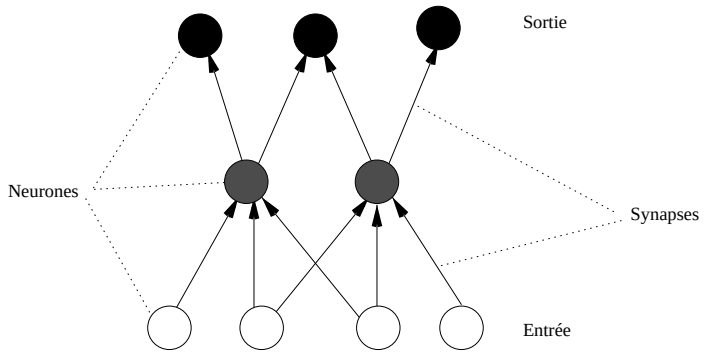


FIG. 2.9 – *Architecture neuronale.*

La figure 2.9 illustre une architecture neuronale classique[1]. L'apprentissage par rétropropagation consiste à présenter une entrée $e(i)$ au réseau et à calculer la sortie $o(k)$ correspondante. Cette sortie est comparée à la sortie attendue $a(k)$. La différence entre la sortie observée et celle calculée par le réseau est utilisée pour modifier les poids synaptiques :

- Phase de propagation : soit $e(i)$ le vecteur d'entrée, l'activation d'un neurone de la couche cachée est calculée par $c(j) = f(\sum_i \omega_{ij} e(i))$ où ω_{ij} sont les poids des synapses et f la fonction d'activation—qui calcule la sortie du neurone caché en fonction de la combinaison linéaire des entrées. De même $o(k) = g(\sum_j \omega_{jk} c(j))$ permet de calculer les composantes de la couche de sortie en fonction de la couche intermédiaire.
- Phase de rétropropagation : soit maintenant $a(k)$ les composantes du vecteur attendu en sortie. Le but est de modifier les ω_{ij} et ω_{jk} afin de minimiser $E = \sum_k \delta_k^2$ où $\delta_k = a(k) - o(k)$. En résolvant $\frac{\partial E}{\partial \omega_{jk}} = 0$ on obtient la formule $\Delta\omega_{jk} = \nu g'(o(k))\delta_k$ et $\Delta\omega_{ij} = \nu f'(c(j))\delta_j$ où l'on a posé $\delta_j = \sum_k \delta_k \omega_{jk}$. Ainsi on voit que l'erreur en sortie δ_k permet de calculer l'erreur δ_j de la couche cachée, d'où le nom de rétropropagation de l'erreur.

La connaissance est ainsi inférée à partir d'exemples. Les poids sont ajustés de manière itérative jusqu'à ce que la différence entre la sortie calculée et la sortie observée soit minime. Par exemple, en reconnaissance d'écriture, on présente en entrée une matrice de pixels, et en sortie un vecteur de dimension le nombre de lettres de l'alphabet, dont toutes les coordonnées sont nulles sauf celle correspondant à la lettre fournie en entrée qui vaut 1. Lors de la reconnaissance, on présente une matrice de pixels. Le neurone de sortie ayant la plus grande activation correspond à la lettre reconnue.

Deux types de réseau sont couramment utilisés en reconnaissance automatique de la parole : les réseaux récurrents (Jordan, Elman[27]) dans lesquels les sorties et/ou neurones de la couche cachée réactivent les entrées (cf. figure 2.10), ceci afin de prendre en compte l'écoulement temporel, et les réseaux à couches comme celui représenté sur

1. L'université de STUTGART produit *Stuttgart Neural Network Simulator*, avec une très bonne documentation pour les personnes intéressées par les réseaux de neurones formels.

la figure 2.9. Ces derniers sont généralement utilisés au sein d'autres architectures (à base de connaissances, modèles de Markov...) pour leur pouvoir de discrimination. Par exemple François [31] utilise des perceptrons multicouches pour améliorer la discrimination entre /p/, /t/ et /k/ au sein du système APHODEX. Des algorithmes d'apprentissage par l'exemple existent pour ces différentes architectures. Pour discriminer entre /no/ et /go/, Watrous et Shastri utilisent un réseau récurrent [105]. Wang et al. [104] utilisent un réseau de Hopfield comme post-processeur dans un réseau de croyance bayésien pour reconnaître les chiffres en Mandarin. Haton donne encore d'autres exemples d'utilisation des réseaux neuronaux isolés et couplés à d'autres architectures dans [35].

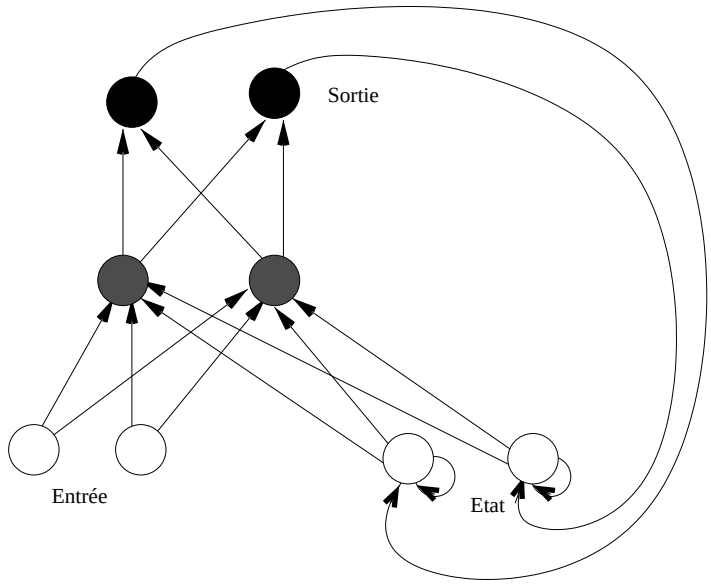


FIG. 2.10 – *Réseau récurrent de Jordan.*

2.3.4 Les modèles de Markov

Le principe de la modélisation par des modèles de Markov discrets est le suivant : on mesure une variable aléatoire discrète $X(t)$ et on cherche à modéliser le processus qui en est à l'origine par un modèle stochastique. Pour cela on se donne n états correspondant chacun à une observation potentielle. La probabilité d'une observation est supposée ne dépendre que de la suite d'observations déjà effectuée et non pas du temps. Selon le nombre d'observations passées servant à estimer la probabilité d'une nouvelle observation, on obtient des modèles de Markov discrets de différents ordres. Pour l'ordre 1, on

obtient donc $P(X(t) = o_i \, et \, X(t-1) = o_j) = P(X(t) = o_i \,|X(t-1) = o_j)\, P(X(t-1) = o_j)$. On peut représenter le modèle par un vecteur de probabilité a priori d'être dans un état o_i et une matrice de probabilité de transiter entre un état o_i et un état o_j. Un tel modèle est représenté par la figure 2.11. Entre deux états quelconques, il existe un arc qui les relie : ce type de modèle est appelé modèle *ergodique*.

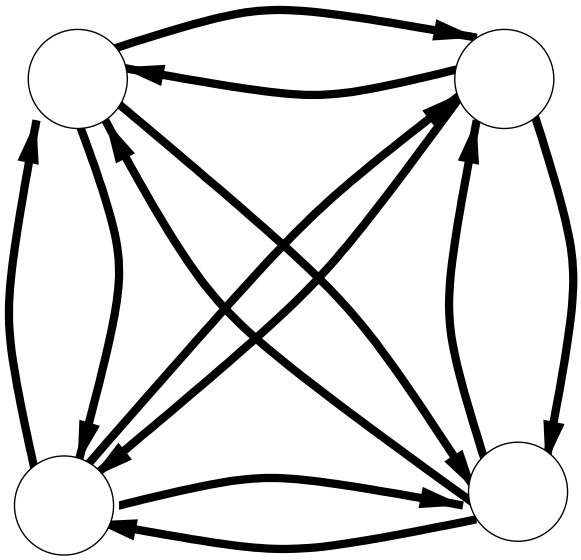


FIG. 2.11 – *Modèle ergodique.*

Dans les modèles de Markov cachés les observations ne sont pas directement observables. À chaque état correspond une distribution de probabilité d'observations. Ainsi un modèle de Markov caché discret est représenté par :

1. un vecteur de probabilité de se trouver a priori dans un état q_i,
2. une matrice de transition entre les états q_i et q_j
3. une distribution de probabilité des observations à l'intérieur de chaque état.

L'avantage par rapport aux modèles de Markov observables en parole est la possibilité d'utiliser un modèle *gauche-droite* (cf. figure 2.12), dans lequel la topologie du graphe représente naturellement l'axe du temps. Les différentes distributions de probabilité à l'intérieur des états peuvent être interprétées comme les caractéristiques acoustiques modélisées.

Une fois choisie la topologie du graphe, il faut estimer les paramètres, c'est-à-dire les coefficients de la matrice de transition, les coefficients du vecteur des probabilités

initiales et les distributions de probabilité dans les états. Pour cela, un algorithme célèbre et efficace a été trouvé. Il s'agit de l'estimation de Baum-Welch, appelée aussi *forward-backward algorithm*. En outre, des critères de convergence existent pour cet algorithme [5].

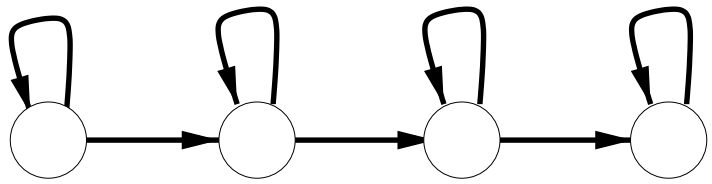


FIG. 2.12 – *Modèle gauche-droite.*

Après avoir entraîné autant de modèles que de formes acoustiques à reconnaître, on peut calculer la probabilité d'une séquence d'observations connaissant chacun des modèles. La forme acoustique reconnue est celle dont l'état final du modèle est le plus probable.

De la même manière que lors de la comparaison de formes acoustiques (cf. section 2.3.1) on peut construire un graphe en reliant les états initiaux et finaux des différents modèles afin de reconnaître des enchaînements de formes acoustiques. La reconnaissance est alors effectuée en utilisant l'algorithme de Viterbi (principe de la programmation dynamique).

L'existence d'algorithmes d'apprentissage par des exemples et la preuve de convergence ont contribué au succès de cette approche. La modélisation stochastique est d'ailleurs aussi utilisée pour modéliser le langage [70], ou pour d'autres tâches de perception comme la reconnaissance de l'écriture [91]. Comme pour les modèles neuromimétiques, l'apprentissage s'effectue à l'aide d'exemples et par conséquent les problèmes de modélisation de l'expertise humaine ne se posent pas immédiatement. Néanmoins plusieurs difficultés subsistent comme le choix de la topologie (nombre d'états principalement), la distribution de probabilité à l'intérieur des états (mesure discrète, mixtures de gaussiennes...) ou encore la modélisation de la durée [84].

À ses début, la modélisation markovienne modélisait séparément chaque forme acoustique en utilisant l'algorithme d'apprentissage MLE (Maximum Likelihood Estimation). Ceci conduisait à une discrimination assez pauvre pour les formes acoustiques proches et à un regain d'intérêt pour les réseaux neuromimétiques réputés pour leur

pouvoir discriminant [42]. Il existe maintenant des algorithmes qui permettent de pallier ce problème. Ainsi Bahl et al. [4] proposent un algorithme qui maximise non plus la vraisemblance de chaque modèle, mais le taux de reconnaissance sur l'espace d'apprentissage. Il utilise pour cela deux coefficients qui visent à inhiber les probabilités des transitions qui sont à l'origine des erreurs (d'où le nom de *Corrective Training*), et à maximiser les autres, comme la règle de Hebb dans les réseaux neuronaux. Même s'ils ne démontrent pas la convergence de l'algorithme, ce dernier a donné expérimentalement de bons résultats. Il a été adapté pour la reconnaissance de mots enchaînés par Lee [51, 52] dans le système SPHINX. Un autre algorithme a amélioré le pouvoir discriminant des HMM. Il s'agit de l'apprentissage MMIE (Maximum Mutual Information Estimation). Tandis que MLE maximise la probabilité a posteriori de l'espace d'apprentissage pour chaque modèle, MMIE maximise la probabilité a posteriori du modèle correspondant aux données, puisque ce critère est celui utilisé pour le décodage de la phrase. L'apprentissage des paramètres d'un modèle tient compte des autres modèles afin d'améliorer le pouvoir discriminant. Ceci conduit cependant à une beaucoup plus grande complexité de l'apprentissage. Cardin, Normandin et al. ont commencé à appliquer cet algorithme pour la reconnaissance de mots isolés [18]. Puis ils présentent dans [77] des approximations rendant possible l'utilisation d'un tel algorithme pour la reconnaissance de mots enchaînés. Dans [76], ils donnent une solution possible au choix du nombre de mixtures par état dans un modèle de Markov en utilisant MMIE. En effet, une bonne modélisation nécessite assez de paramètres pour modéliser précisément les caractéristiques acoustiques de l'espace d'apprentissage, mais d'un autre côté l'estimation d'un grand nombre de paramètres requiert un immense espace d'apprentissage.

Chapitre 3

Reconnaissance de mots à l'aide d'automates

Dans cette partie, nous allons présenter les travaux commencés lors de mon stage de DEA et poursuivis pendant ma première année de thèse sous l'encadrement de Joseph Di Martino. Dans une première étape, nous décrirons une technique qui a donné de bons résultats en reconnaissance de mots isolés. Puis nous montrerons comment elle a été étendue à la reconnaissance de mots enchaînés et les difficultés qui se sont présentées. Pour conclure, nous ferons le bilan des avantages et des points faibles de cette méthode.

3.1 Reconnaissance de mots isolés

3.1.1 Introduction

Les modèles de Markov (Hidden Markov Models, HMM) nécessitent un apprentissage long—ce qui en soit n'est pas un problème—mais demandent aussi un corpus représentatif pour chaque mot à reconnaître, ce qui n'est pas sans poser quelques difficultés. En outre, la topologie du modèle de Markov est fixée a priori et n'est pas modifiée lors de l'apprentissage. En effet, seules les probabilités de transition entre les différents états et les probabilités d'observation à l'intérieur de ces derniers sont ajustées à l'aide des algorithmes classiques d'apprentissage dans les modèles de Markov. Le choix de la topologie initiale est donc un problème délicat. Des tentatives ont vu le jour pour introduire des connaissances phonétiques ou/et phonologiques lors de la

construction de la topologie du modèle [25, 101].

La méthode que nous proposons est inspirée des premiers travaux en reconnaissance automatique de la parole, à l'époque où l'on utilisait un taux de dissemblance entre la locution à reconnaître et des locutions de référence, calculé grâce à la programmation dynamique. Les travaux ont pour origine la thèse de Joseph Di Martino [62] qui a utilisé les travaux de Sakoe et Chiba [89]. Cependant les coûts de cette technique sont élevés que ce soit en mémoire, pour sauvegarder les locutions de référence, ou en temps CPU à la reconnaissance. Par ailleurs l'article de Baum-Welch [5] à donné une preuve de convergence de l'algorithme du MLE (Maximum Likelihood Estimation) qui fut dès lors utilisé dans les modèles de Markov. Ceci a contribué au déclin de l'approche par programmation dynamique. Cependant en construisant un dictionnaire de vecteurs acoustiques (avec l'algorithme de quantification de Linde Buzo et Gray LBG [56]) puis en codant les locutions de référence tout en les factorisant à gauche, nous avons pu réduire l'espace mémoire et le temps de calcul requis ce qui permet d'obtenir de très bons résultats en reconnaissance de mots isolés comme nous allons le voir dans les sections suivantes.

3.1.2 Apprentissage

La base de donnée utilisée est TIDIGIT qui a servi à de nombreuses études. Le choix de cette base permet de comparer les résultats obtenus avec d'autres travaux. Le signal est échantillonné à $f_e = 20$kHz. La parole n'est pas bruitée. Le vocabulaire est formé des 11 chiffres anglo-saxons (*o* et *zero* pour le 0). Pour l'apprentissage, nous disposons de 55 locuteurs, des hommes adultes de divers états des USA. Chacun d'eux a prononcé deux fois l'ensemble des chiffres, ce qui fait 110 locutions de référence par mot. L'apprentissage s'effectue suivant les étapes suivantes :

- paramétrisation du signal,
- construction du dictionnaire,
- construction de l'automate.

Nous allons détailler ces trois étapes.

Paramétrisation du signal

Nous utilisons l'analyse cepstrale (cf. section 2.2.3) pour analyser le signal. Cette analyse suppose l'hypothèse couramment admise : le signal de parole $x(t)$ résulte de la convolution de sources (vibrations des cordes vocales, bruit de friction) avec la fonction de transfert du conduit vocal.

Tous les 256 échantillons (16 ms) nous extrayons du signal acoustique préaccentué

$$s(i) = x(i) - 0.96x(i-1)$$

des fenêtres de $N = 512$ échantillons ce qui correspond à une durée de 32ms—une puissance de deux étant idéale pour la FFT (*Fast Fourier Transform*)— Le fenêtrage utilisé est celui de Hamming

$$h(i) = 0.54 - 0.46\cos\left(\frac{2\pi i}{N-1}\right),\ i \in [0, N-1]$$

et $h(i) = 0$, $i \geq N$. La transformée de Fourier à court terme de s,

$$\widetilde{s}_\tau(f) = \sum_{i=-\infty}^{\tau} s(i)h(\tau - i)e^{-\frac{2\pi f i}{N}}$$

est calculée pour f compris entre 0 et la fréquence de Nyquist $\frac{f_e}{2}$ =10kHz. Puis nous calculons le logarithme $\log_{\mathcal{C}}(\widetilde{s}_\tau)$ défini par $\log_{\mathcal{C}}(z) = \log(|z|) + j\arg(z)$ où $z \in \mathcal{C}$. Enfin nous intégrons ce logarithme suivant 20 canaux répartis selon une échelle fréquentielle non linéaire mel :

$$1\,mel = 1125\log(0.0016f + 1)\,Hz$$

où f est la fréquence en Hz et m la fréquence en mel. Cette échelle a l'ambition de modéliser la sensibilité de l'oreille humaine. Une transformée de Fourier inverse fournit les coefficients. Les premiers d'entre eux, qui contiennent le plus d'information au sujet de la fonction de transfert du conduit vocal, sont conservés [23].

Ainsi nous obtenons un vecteur de paramètres MFCC (Mel Frequency scaled Cepstral Coefficient) $\overrightarrow{c}$ toutes les 16ms. L'étape suivante consiste à quantifier cet espace i.e. à trouver l'ensemble de *prototypes* $\overrightarrow{p}_k$ le plus représentatif de cet ensemble de vecteurs.

Construction du dictionnaire

Nous utilisons l'algorithme LBG (Linde Buzo Gray [56]). La construction est itérative : le premier prototype $\overrightarrow{p}_1$ est le centre de gravité de l'ensemble des vecteurs

cepstraux, la première partition est constituée d'une seule partie. Puis nous divisons chaque partie en deux, itérativement, jusqu'à l'obtention du nombre de classes voulu.

1. Calcul du centre de gravité $\overrightarrow{p}_1$ de l'espace, $N = 1$, $i = 1$
2. Une légère perturbation de $\overrightarrow{p}_i$ donne $\overrightarrow{p}_{2i} = 1.0001\overrightarrow{p}_i$ pour i allant de 1 à N, initialisation de la distorsion du codage $d = \infty$
3. Partition de l'espace : $E_i = \left\{ \overrightarrow{p} \left| \|\overrightarrow{p} - \overrightarrow{p}_i\| < \left\|\overrightarrow{p} - \overrightarrow{p}_j\right\| , \forall j \neq i \right. \right\}$ pour $i \in [1, N]$
4. Calcul de la nouvelle distorsion $d' = \sum_{i=0}^{N} \sum_{\overrightarrow{p} \in E_i} \|\overrightarrow{p} - \overrightarrow{p}_i\|^2$ et de leurs centres de gravité $\overrightarrow{p'}_i = \sum_{k \epsilon E_i} \overrightarrow{p}_k$
5. Si $d' < d$ alors on remplace p_i par p'_i, on remplace d par d' et on retourne en 3
6. Tant que N est différent du nombre de classes voulu, retourner en 2

À l'aide des *codebooks* obtenus, nous codons les locutions en remplaçant chaque vecteur par l'indice du prototype le plus proche. Nous obtenons ainsi un ensemble de suites de vecteurs prototypes comme modèle pour chacun des mots. Par exemple, sur la figure 3.1, l'ensemble des vecteurs cepstraux est partitionné en 4 classes. La locution représentée par la flèche—qui donne le sens du temps— sera codée 4, 3, 2, 2, 2, 2, 1, 1, 1, 4, 1.

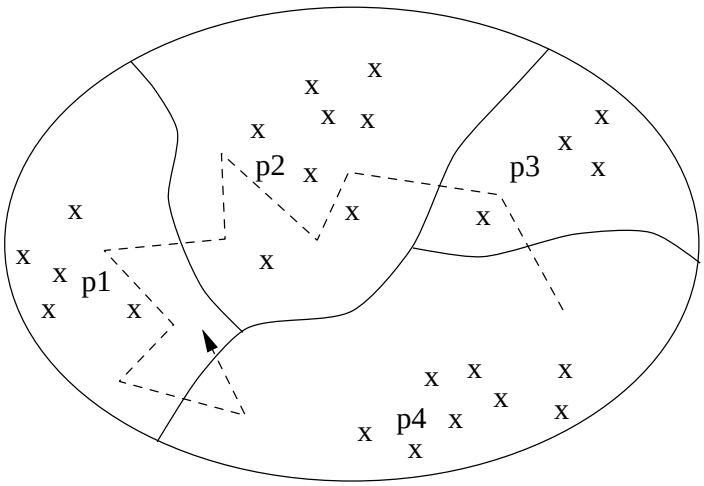


FIG. 3.1 – *Construction d'un dictionnaire et codage d'une locution.*

Construction de l'automate

Pour chaque mot, nous construisons un automate en factorisant à gauche les locutions codées [26]. Cependant, afin de réduire au maximum la place mémoire nécessaire—

et par conséquent le temps CPU requis pour reconnaître une locution—nous introduisons des boucles quand 4 prototypes identiques sont consécutifs. Ceci permet aussi de contraindre l'alignement temporel de telle sorte que les zones les plus stables du signal soient les plus sujettes aux dilatations et aux contractions. La figure 3.2 illustre la construction d'un tel automate. Elle représente le codage de deux locutions d'un mot et l'automate correspondant. On remarque que le premier prototype $\overrightarrow{p}_1$ est factorisé. Deux boucles ont été introduites dans les parties stables du signal : pour le prototype $\overrightarrow{p}_3$ dans la première locution et pour $\overrightarrow{p}_1$ dans la seconde.

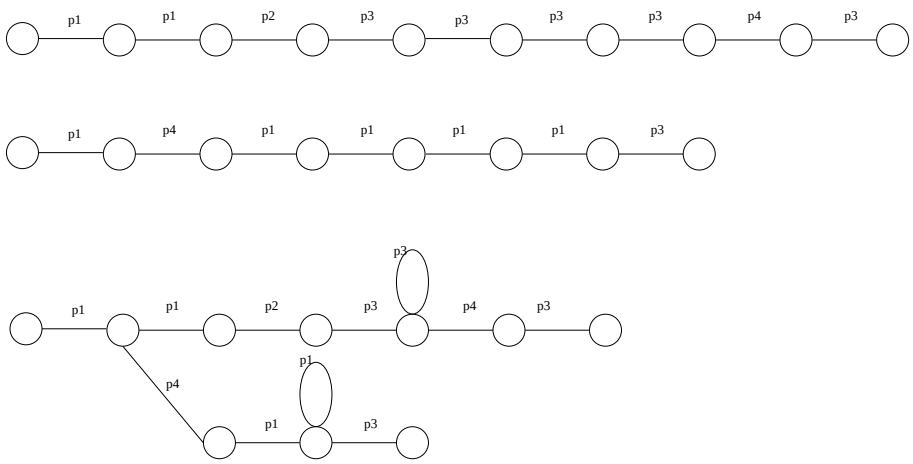


FIG. 3.2 – *En haut deux locutions, en bas l'automate correspondant.*

3.1.3 Reconnaissance

Dans la base TIDIGIT, nous disposons d'un ensemble de 56 locuteurs disjoint de celui qui a servi à l'apprentissage. Ces locuteurs ont prononcé deux fois chaque mot ce qui fait $56 \times 2 \times 11 = 1232$ mots isolés à reconnaître. Comme pour l'apprentissage, ces locutions sont paramétrisées en utilisant les MFCC. Puis elles sont codées suivant les prototypes de chacun des modèles. Enfin pour chaque locution, nous comparons son codage à l'automate du mot correspondant en utilisant la programmation dynamique.

Nous allons commencer cette section par la description du principe utilisé, puis nous décrirons les résultats obtenus :

- comparaison du taux de reconnaissance avec des résultats obtenus en utilisant un modèle de Markov discret, et incrémentalité de l'apprentissage,

- influence de la dérivée des vecteurs cepstraux, quantification des références et reconnaissance temps réel.

Enfin, avant de décrire l'extension de la méthode au problème de la reconnaissance de mots enchaînés, nous discuterons ces résultats.

Principe de la reconnaissance

Nous mesurons un taux de dissemblance entre la locution à reconnaître et chacun des automates. Le mot reconnu est celui dont le modèle donne le plus petit taux de dissemblance[2]. Celui-ci est défini comme étant le minimum des taux de dissemblance entre la locution à reconnaître et chacune des branches de l'automate. Le taux de dissemblance entre une locution paramétrisée $\overrightarrow{c}_i$, $i = 1, \ldots, N$ et une suite de prototypes $\overrightarrow{p}_j$, $j = 1, \ldots, M$ est défini par :

$$D(\overrightarrow{c}, \overrightarrow{p}) = \min_W \left(\sum_{k=1}^{K} \frac{d(\overrightarrow{c}_{W_i(k)}, \overrightarrow{p}_{W_j(k)}) P(W(k-1), W(k))}{N(W)}\right)$$

où $W = (W_i, W_j)$ est un *chemin de recalage,* c'est-à-dire une fonction de $[1, \ldots, K]$ dans $[1, \ldots, N] \times [1, \ldots, M]$ croissante au sens large (pour rendre compte de l'écoulement du temps) et telle que $W(1) = (1, 1)$ et $W(K) = (N, M)$. $d(.,.)$ est une distance dans l'espace des vecteurs cepstraux, nous avons utilisé une distance euclidienne. P est une pondération qui dépend de contraintes locales imposées à W et $N(W) = \sum_{k=1}^{K} P(W(k-1), W(k))$ est un facteur de normalisation rendant $D(\overrightarrow{c}, \overrightarrow{p})$ indépendant de la longueur du chemin.

Ces chemins de recalage permettent de tenir compte des différents débits d'élocution en comparant des locutions de longueurs différentes de manière judicieuse. En effet le changement du débit d'élocution n'affecte pas également tout les phonèmes et par conséquent le recalage linéaire $W_i = \frac{N}{M} W_j$ ne suffit pas. La figure 3.3 donne un exemple de recalage temporel.

Mais le calcul des taux de dissemblance pour chacun des chemins est excessivement coûteux. En effet, il existe M^N chemins satisfaisant ces contraintes. Heureusement,

2. Le taux de dissemblance n'est pas forcément une distance, en particulier la propriété de symétrie n'est pas forcément vérifiée : cela dépend des contraintes utilisées.

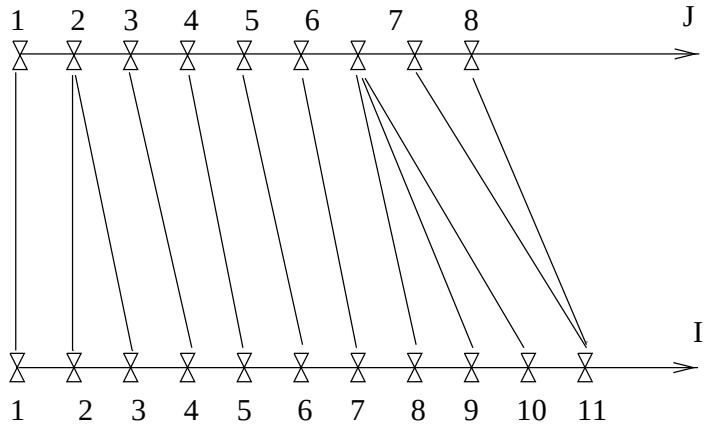


FIG. 3.3 – *Recalage temporel.*

la programmation dynamique permet le calcul du chemin optimal W (qui minimise $D(\overrightarrow{c}, \overrightarrow{p})$) avec un coût en $\mathcal{O}(N.M)$ opérations.

Soit W un chemin optimal qui conduit de (i_1, j_1) à (i_2, j_2). Alors si $(i, j) \in W$, le chemin menant de (i_1, j_1) à (i, j) est lui aussi optimal.

Maintenant, des contraintes assurent que chaque vecteur de la locution à reconnaître est mis en correspondance avec un prototype. Posons $W_o(i, j)$ le chemin optimal entre $(1, 1)$ et (i, j). Alors ce chemin est composé d'un chemin optimal $W_o(i-1, j_0)$ et d'un chemin entre $(i-1, j_0)$ et (i, j). On peut donc calculer pour chaque i tous les chemins optimaux tels que $W_i = i$. La recherche du chemin optimal est donc en $\mathcal{O}(N.M)$. En outre, il suffit maintenant de mémoriser $2N$ taux de dissemblance car les coûts des chemins arrivant en $(i-2, j)$ ne sont pas utilisés pour calculer les coûts de ceux qui aboutissent en (i, j).

La figure 3.4 présente trois exemples de contraintes locales, ainsi que les pondérations P. Deux types de pondérations sont habituellement choisis :

- les pondérations symétriques : $P(W(k-1), W(k)) = W_i(k) - W_i(k-1) + W_j(k) - W_j(k-1)$ donc $N(W) = W_i + W_j$
- les pondérations asymétriques : $P(W(k-1), W(k)) = W_i(k) - W_i(k-1)$ donc $N(W) = W_i$

Nous avons choisi des contraintes qui obligent les dilatations à s'effectuer dans les zones stables (où l'on a introduit des boucles). La figure 3.5 illustre un chemin optimal possible satisfaisant ces contraintes.

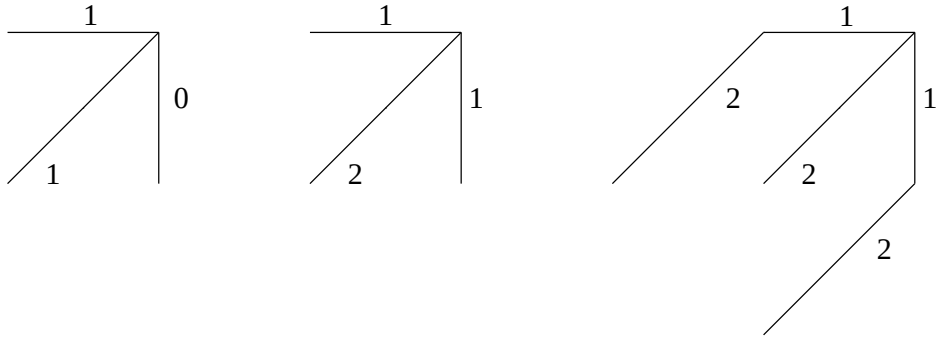


FIG. 3.4 – *Contraintes locales : asymétrique, symétrique, de Sakoe et Chiba symétrique.*

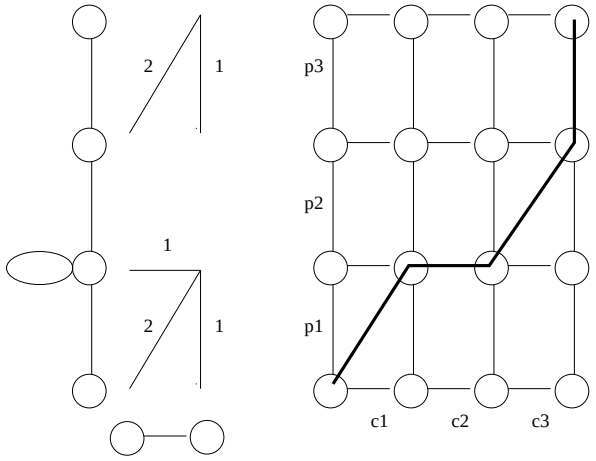


FIG. 3.5 – *Les contraintes utilisées et un exemple de chemin.*

Voici l'algorithme de calcul du taux de dissemblance entre une locution et un automate :

Initialiser C_{00} le cumul des distances le long du chemin optimal finissant à l'état j=0 de l'automate au temps i= 0 à 0

Pour i de 1à N

Initialiser la C_{i0} à l'infini : contraindre le chemin optimal à débuter en (0,0)

Pour j de 1 au nombre d'états de l'automate

Calculer $C_{ij} = \min(\min_{j'} C_{i-1,j'} + P((i-1,j'),(i,j)) * d(c_i,p_j), \min_{j' \neq j} C_{ij'} + P((i,j'),(i,j)) * d(c_i,p_j))$ le cumul des distances le long du chemin optimal finissant en (i,j) où j' est un état prédécesseur de j

Calculer L_{ij} la longueur de ce chemin en fonction des états prédécesseur de j

Le taux de dissemblance vaut $\min_{j'}(\frac{C_{N,j'}}{L_{Nj'}})$ où j' est un état final de l'automate.

La figure 3.6 illustre cet algorithme.

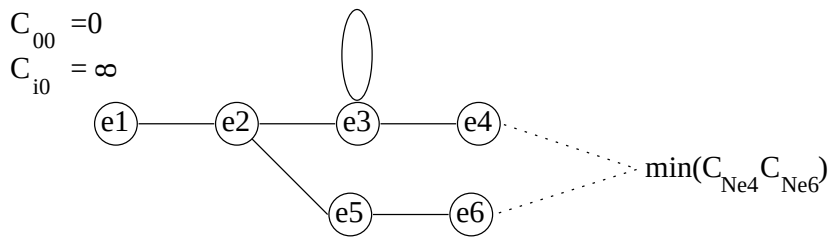


FIG. 3.6 – *Principe de la reconnaissance de mots isolés*

Enfin, le mot reconnu est celui dont le modèle donne le taux de dissemblance le plus petit.

Incrémentalité de l'apprentissage

Une des difficultés rencontrées lors de l'utilisation des modèles de Markov est la réestimation des paramètres (moyenne et matrice de covariance à l'intérieur des états

pour les modèles semi-continus et probabilités de transition). En effet, étant donnée la manière dont s'effectue l'apprentissage, les locuteurs mal représentés dans le corpus, ou les conditions différentes (milieux bruités, microphones différents lors de l'acquisition du corpus d'apprentissage et de la reconnaissance...) font chuter le pourcentage de mots reconnus. Nous avons donc testé la possibilité pour notre système d'apprendre à corriger ses erreurs. Pour cela, les locutions mal reconnues sont codées puis ajoutées dans le modèle correspondant.

En utilisant 23 coefficients cepstraux et 256 prototypes, nous avons obtenu un taux de reconnaissance de 98.5% soit 19 erreurs sur 1232 locutions à reconnaître. Ces résultats sont comparables aux 99.4% obtenus en utilisant des modèles de Markov discrets [82]. Après adjonction de ces 19 locutions dans les automates correspondants, seules 4 erreurs subsistaient, soit un taux de reconnaissance de 99.9%.

Reconnaissance temps réel

À chaque instant, notre algorithme effectue autant de calculs de distance qu'il y a de vecteurs prototypes, et autant de comparaisons qu'il y a d'états dans les automates. Le nombre de prototypes étant fixé, le coût croît de façon linéaire en fonction du nombre d'états.

Pour réduire la place mémoire ainsi que le temps CPU nécessaires à la reconnaissance d'un mot, nous devons diminuer le nombre d'états, par exemple en choisissant un sous-ensemble des locutions de la base d'apprentissage comme locutions de référence. Pour cela nous avons développé un algorithme de quantification des formes de référence. D'autre part, certains auteurs ont remarqué que l'utilisation de la dérivée des coefficients cepstraux améliorait les taux de reconnaissance. Ceci est à rapprocher de l'utilisation de modèles de Markov d'ordre 2 ou des modèles de trajectoire de Gong [1]. L'amélioration des performances s'explique par l'apport d'information sur l'évolution des paramètres. Dans un premier temps nous allons présenter l'influence de la dérivée des paramètres puis nous décrirons l'algorithme de quantification des formes de référence.

Influence de la dérivée des coefficients cepstraux

Comme l'ont remarqué certains auteurs [11], la prise en compte d'informations sur la cinétique des paramètres améliore souvent les taux de reconnaissance. Nous avons donc utilisé la dérivée des coefficients cepstraux. Une pondération $p_1 = 0,48$ a été calculée pour donner la même influence aux vecteurs cepstraux qu'à leur dérivée. La distance entre deux paramètres pour lesquels les vecteurs cepstraux sont respectivement $\overrightarrow{u}$ et $\overrightarrow{v}$, est $\sqrt{d^2(\overrightarrow{u}, \overrightarrow{v}) + p_1 d^2(\overrightarrow{u}', \overrightarrow{v}')}$. Cette distance est utilisée pour quantifier le nouvel espace et pour comparer les locutions aux références.

Le tableau 3.1 donne les taux d'erreur obtenus en fonction du nombre de paramètres prototypiques retenus. En outre, nous avons utilisé seulement onze coefficients cepstraux.

Nombre de classes	16	32	64
Taux d'erreur avec vecteurs cepstraux seuls	1.14	0.97	0.81
Taux d'erreur avec dérivée des vecteurs cepstraux	0.49	0.41	0.49

TAB. 3.1 – *Influence de la dérivée du vecteur cepstral*

Quantification des formes de référence

L'idée de cet algorithme est de choisir seulement certaines locutions dans l'espace d'apprentissage. Cela permet d'améliorer le temps de réponse du logiciel puisque le nombre de calculs de distance est réduit. Pour cela on cherche les points correspondant au diamètre maximum de l'espace des locutions, puis on choisit parmi les deux extrémités celle qui possède la plus proche locution voisine. On cherche un deuxième point de cette manière. Puis nous divisons l'espace en deux comme pour quantifier les vecteurs acoustiques. L'itération est similaire à celle de l'algorithme de quantification vectorielle.

Algorithme :

Soit E l'ensemble des locutions de l'espace d'apprentissage.

1. Chercher le premier prototype $\overrightarrow{p}_1 = \arg\min_{\overrightarrow{u} \in E} \max_{\overrightarrow{v} \in E} D(\overrightarrow{u}, \overrightarrow{v})$, $N = 1$, $E_N = E$.

2. Chercher le second prototype $\overrightarrow{p}_{N+1}$ de la même manière dans $E_N \setminus \{\overrightarrow{p}_N\}$, initialiser la distorsion $d = \infty$
3. Partitionner E en affectant à E_i les vecteurs $\overrightarrow{p}$ de E tels que $D(\overrightarrow{p}, \overrightarrow{p}_i) \leq D(\overrightarrow{p}, \overrightarrow{p}_j)$, $i, j \leq N$ et $j \neq i$
4. Calculer la distorsion du codage et réitérer les étapes 3 et 4 tant que celle-ci diminue.
5. Incrémenter N et retourner en 2 si le nombre de classes voulu n'est pas atteint.

Dans l'expérience réalisée nous utilisons 22 coefficients (vecteurs cepstraux et leurs dérivées), et 32 prototypes issus de la quantification vectorielle. Le tableau suivant synthétise les résultats obtenus en fonction du nombre de formes de référence retenues.

Nombre de formes	4	8	12	16	32
Taux d'erreur (%)	12.42	6.25	1.94	0.81	0.41
Temps CPU (ms)	394	496	605	709	1125

TAB. 3.2 – *Influence du nombre de références retenu*

Les tests ont été effectués sur un HP 9000/755. Nous constatons que le taux de reconnaissance croît avec le nombre de prototypes utilisés. Sachant que les locutions ont une durée de l'ordre de la seconde, nous pouvons effectuer la reconnaissance en temps réel avec 16 prototypes.

3.1.4 Conclusion

La méthode décrite ci-dessus a donné des résultats comparables à ceux obtenus en utilisant des modèles de Markov discrets. En outre, elle a l'avantage de ne pas nécessiter un corpus important pour l'apprentissage. Un second point fort de cette technique est la possibilité d'effectuer un apprentissage incrémental dans le sens où une locution mal reconnue peut facilement être insérée dans l'automate. Il serait ainsi possible d'envisager une adaptation au locuteur par exemple, mais pour cela il faudrait élaguer l'automate car la borne supérieure du nombre d'états est gigantesque.

Dans la partie suivante, nous allons décrire les modifications apportées à l'algorithme pour l'adapter à la reconnaissance de mots enchaînés.

3.2 Reconnaissance de mots enchaînés

La base TIDIGIT contient aussi des « phrases » qui sont des suites de chiffres[3]. Les mêmes locuteurs ont produit les mots isolés et les phrases.

Nous allons présenter la modification effectuée sur l'algorithme précédent pour reconnaître des phrases à partir des modèles créés avec des mots isolés. Cependant les résultats obtenus sont décevants du fait de l'absence de prise en compte des effets de coarticulation. En effet, ces derniers introduisent des modifications acoustiques importantes aux frontières des mots adjacents. Ensuite, nous montrerons comment nous avons segmenté les phrases de l'espace d'apprentissage pour obtenir des mots en contexte. En effet, ceux-ci permettent généralement d'obtenir de meilleurs scores de reconnaissance puisqu'ils permettent de prendre en compte la variabilité due à la coarticulation. Or la base utilisée n'a pas été segmentée préalablement. Enfin nous présenterons les contraintes supplémentaires imposées aux transitions entre deux modèles dans le but de mieux prendre en compte cette variabilité. Pour conclure, nous comparerons les avantages et les points faibles respectifs des modèles de Markov et de la méthode présentée dans ce chapitre.

3.2.1 Modification de l'algorithme de reconnaissance

Nous modifions l'algorithme présenté page 31 qui effectue la reconnaissance de mots en ajoutant un « niveau phrase ». En effet, reconnaître une phrase implique la comparaison d'une locution à l'ensemble des concaténations de mots possibles, pour cela il faut :

- déterminer le nombre de mots de la phrase,
- reconnaître chacun des mots.

Ces deux tâches peuvent être conduites simultanément comme l'ont fait remarquer Bridle et Nakagawa [15, 73]. Les modifications à apporter à l'algorithme sont :

- les modifications des contraintes de manière à ce que le chemin optimal puisse être composé d'une concaténation de mots,

3. Les « phrases » sont l'ensemble des concaténations de longueurs finies de chiffres (cf. page 36). Pour un lexique aussi petit, aucun modèle de langage n'est utilisé.

– la mémorisation à chaque instant du mot précédemment reconnu afin de retrouver la suite de mots prononcée.

Nouvel algorithme Le niveau phrase introduit des contraintes sur les transitions entre les différents modèles. Dans la base utilisée (choisie pour permettre une comparaison des résultats obtenus en reconnaissance de mots isolés avec d'autres travaux utilisant des modèles de Markov) toutes les transitions entre différents modèles sont possibles car il n'y a pas de contrainte syntaxique.

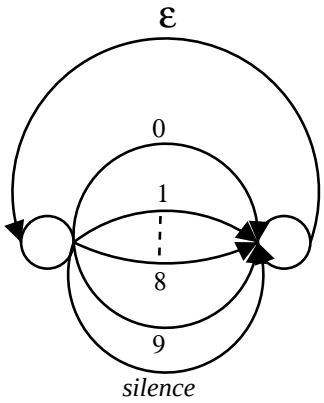


FIG. 3.7 – *Modèle des séquences de chiffres*

Le nouvel algorithme est semblable à celui qui permet de reconnaître les mots isolés. Seulement le taux de dissemblance à l'état 0 au temps t de chaque automate A n'est plus initialisé à l'infini, mais en respectant les contraintes du niveau phrase :

$$C_{i0}^{A} = \min_{A'} \min_{j'} C_{i-1,j'}^{A'} + P((i-1,j'),(i,j)) * d(c_i, p_j^{A'}) \tag{3.1}$$

où j' est un état final de A' et A' parcourt l'ensemble des automates. La figure 3.7 illustre le « langage » reconnu correspondant à cette nouvelle contrainte : toutes les séquences de chiffres sont possibles, le symbole ϵ représentant le mot vide. Le long des transitions étiquetées par les chiffres ou le silence, le calcul se conduit comme précédemment.

Par ailleurs, à chaque instant et pour tout les états finaux des automates, nous mémorisons le modèle précédemment reconnu, ce qui permet de retrouver la suite de mots prononcée. La figure 3.8 illustre le principe du décodage (*backtracking*) de la suite reconnue. Le dernier état final du chemin optimal (`Pred`) ainsi que l'instant où il a été

atteint (`Temps`) sont mémorisés à chaque étape afin qu'il soit possible de retrouver la suite de mots effectivement prononcée. Ainsi, à la fin de la deuxième locution (`3`), le meilleur état final du troisième modèle (A_3) référence l'état final du modèle A_1 au temps correspondant à la fin de la locution précédente (`1`).

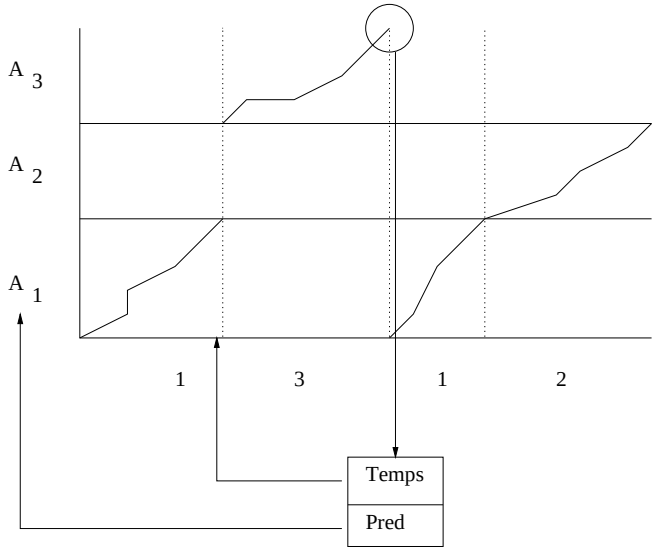


FIG. 3.8 – *Principe de la reconnaissance de mots enchaînés*

3.2.2 Nouvel apprentissage

Les locutions de TIDIGIT sont entourées de silence. Ceci n'est pas gênant pour la reconnaissance de mots isolés. En effet, dans ce cas le silence du modèle sera mis en correspondance avec celui de la locution à reconnaître. Cependant, les chiffres qui constituent les « phrases » ne sont pas tous séparés par du silence. Il faut donc le supprimer lors de la construction du modèle. Pour cela nous utilisons deux critères : le nombre de passages par zéro du signal ainsi que son énergie. Avec le silence ainsi extrait nous construisons un automate. Le vocabulaire est maintenant constitué des 11 chiffres et du silence.

3.2.3 Evaluation du taux de reconnaissance

Maintenant, l'évaluation du taux de reconnaissance doit tenir compte des substitutions, mais aussi des insertions et des omissions de mots. Toutefois le silence est

supprimé des phrases reconnues avant comparaison puisqu'il n'est pas transcrit dans le corpus. Ensuite, on définit deux taux de reconnaissance :

1. le taux de mots reconnus,
2. le taux de phrases reconnues, c'est-à-dire dans lesquelles il n'y a ni insertion, ni omission, ni substitution de chiffre.

Les tableaux 3.3, 3.4 et 3.5 illustrent la comparaison de deux chaînes. La première et la deuxième chaîne, respectivement en abscisse et en ordonnée, sont représentées en gras. Ces séquences de symboles sont comparées en utilisant la distance suivante : $d(s_1, s_2) = 1$ si $s_1 \neq s_2$, $d(s, s) = 0$. Les tableaux représentent le cumul des coûts le long des chemins optimaux (tableau 3.4), la longueur de ces chemins (tableau 3.3) et finalement le chemin optimal (tableau 3.5). En fixant le coût local d'une substitution (ou confusion) de symbole à 1, le nombre de substitutions entre les deux chaînes est égal à la somme des coûts le long du chemin optimal. Le nombre d'insertions de symboles dans la première chaîne par rapport à la deuxième correspond au nombre de transitions horizontales et le nombre d'omissions de symboles au nombre de transitions verticales dans ce chemin. Sur l'exemple, il y a donc une substitution (un **1** au lieu d'un **3**) et une insertion (un **3**). Le nombre de mots reconnus est égal à la longueur de la phrase du modèle moins le nombre de substitutions. Ici cela donne $6 - 1 = 5$ mots reconnus.

3	6	6	6	6	6	6	7
1	5	5	5	5	5	6	7
2	4	4	4	4	5	6	7
1	3	3	3	4	5	6	7
1	2	2	3	4	5	6	7
3	1	2	3	4	5	6	7
	1	**1**	**1**	**2**	**1**	**3**	**3**

TAB. 3.3 – *Longueurs*

3	3	3	3	3	2	1	1
1	2	2	2	2	1	2	3
2	2	2	2	1	2	3	4
1	1	1	1	2	2	3	4
1	1	1	1	2	2	3	4
3	1	2	3	4	5	5	5
	1	**1**	**1**	**2**	**1**	**3**	**3**

TAB. 3.4 – *Coûts*

3						x	x
1					x		
2				x			
1			x				
1		x					
3	S						
	1	**1**	**1**	**2**	**1**	**3**	**3**

TAB. 3.5 – *Chemin optimal*

Premiers résultats

Les premiers résultats ont donné de l'ordre de 60% de phrases reconnues. Ce faible score peut s'expliquer puisque les mots en contexte subissent de fortes modifications dues aux effets de coarticulation. Afin de prendre en compte cet effet, nous devons ajouter dans les automates des locutions extraites de phrases. Pour cela il nous faut segmenter celles-ci, c'est-à-dire déterminer le début et la fin de chacun des mots à l'intérieur d'une phrase. La partie suivante décrit la démarche.

3.2.4 Segmentation

Pour segmenter une phrase, on modifie légèrement l'algorithme de reconnaissance de manière à contraindre le système à reconnaître la suite de chiffres prononcée par le locuteur. Nous commençons par créer une instance de modèle pour chaque chiffre de la phrase. Entre chacun des chiffres, ainsi qu'en début et fin de phrase, il peut y avoir du silence. La figure 3.9 illustre les contraintes imposées à la reconnaissance

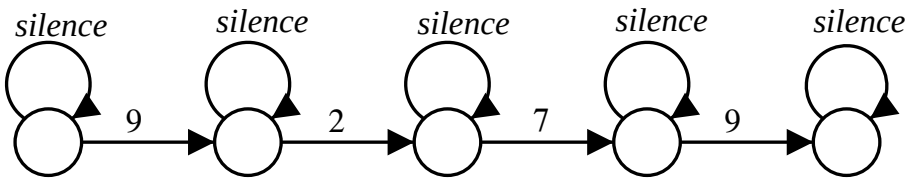


FIG. 3.9 – *Contraintes pour la segmentation*

pour segmenter le mot « 9279 ». La modification de l'algorithme se traduit donc par les mêmes contraintes que précédemment (cf. équation 3.1) où A' est maintenant une instance soit de silence, soit du modèle du mot précédent A et où j' parcourt ses états finaux. Le décodage permet alors de retrouver les début et fin des différents mots prononcés.

3.2.5 Problèmes rencontrés lors de la segmentation

L'algorithme de segmentation automatique ne nous a pas donné satisfaction puisqu'il a conduit à des mots de durée irréaliste : 4 échantillons soit 64ms pour certains. Nous avons donc modifié les contraintes à l'intérieur des modèles de mots afin de re-

médier à ce problème :

- suppression des boucles dans les automates,
- utilisation des contraintes de Sakoe et Shiba symétriques ; avec celles-ci, les dilatations temporelles ne peuvent varier que dans l'intervalle $[\frac{1}{2}, 2]$.

La segmentation a alors donné des résultats beaucoup plus réalistes. Cependant, il est à noter que l'évaluation de la qualité d'une segmentation est un problème difficile. On peut citer, par exemple, les travaux du CSLU [22] dans lesquels diverses segmentations de phrases en phonèmes faites par des experts de langues maternelles différentes sont comparées. Or il n'y avait pas d'accord parfait entre les experts, que ce soit à propos des étiquettes ou bien encore des frontières entre les phonèmes.

3.2.6 Résultats

Nous avons conservé les contraintes de Sakoe et Chiba symétriques lors de la reconnaissance. L'amélioration est significative : désormais le taux de reconnaissance s'élève à 93,3% pour les phrases entièrement correctes tandis que le taux de mots reconnu est de 97,7%. Le tableau 3.6 présente les substitutions, le tableau 3.7 les insertions ; tandis que les omissions sont consignées dans le tableau 3.8. Cependant, ils ne sont pas encore à la hauteur de ceux obtenus par Bocchieri et Wilpon [11]. En utilisant des modèles de Markov à 10 états comprenant 64 mixtures de vecteurs cepstraux, ils obtiennent en effet sur TIDIGIT des taux de reconnaissance supérieurs à 98% selon les caractéristiques de la paramétrisation cepstrale et atteignant même 98.6%.

3.2.7 Prise en compte du contexte droit

Comme l'effet de coarticulation dépend évidemment du couple de mots adjacents, nous avons introduit des contraintes de manière à n'autoriser les transitions entre un état final e_f d'un automate et l'état initial de l'automate A seulement lorsque la locution dont est issue la branche se terminant en e_f était suivie d'une prononciation dont le modèle est A.

Pour cela il faut mémoriser le contexte droit des prononciations qui forment l'automate. La modification apportée à l'algorithme devient :

	o	1	2	3	4	5	6	7	8	9	zero
o	0	1	1	0	7	7	0	2	0	2	0
1	0	0	0	0	4	0	0	0	0	1	0
2	3	0	0	0	0	0	0	0	1	0	0
3	0	0	10	0	0	0	0	0	0	0	0
4	5	1	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0	0	0	0
6	0	0	13	0	0	0	0	11	12	0	0
7	5	1	0	0	0	0	0	0	0	2	0
8	1	2	3	6	0	1	0	0	0	2	0
9	0	0	0	0	0	17	0	0	0	0	0
zero	0	0	2	0	0	0	0	1	0	0	0

TAB. 3.6 – *Tableau de confusion*

o	1	2	3	4	5	6	7	8	9	zero
97	1	9	1	0	6	0	0	29	0	1

TAB. 3.7 – *Insertions*

o	1	2	3	4	5	6	7	8	9	zero
25	7	4	2	0	0	1	0	20	1	0

TAB. 3.8 – *Omissions*

$$C_{i0}^{A} = \min_{A'} \min_{j'} C_{i-1,j'}^{A'} + P((i-1,j'),(i,j)) * d(c_i, p_j^{A'})$$

où j' est un état final de A' ayant A pour contexte droit.

Résultats

Nous n'avons pas noté d'amélioration des résultats obtenus. Deux raisons peuvent expliquer cet échec. D'une part la quantification vectorielle ne prend pas en compte le contexte : la seule information temporelle utilisée est la dérivée des vecteurs et les transitions ne sont pas bien modélisées par la quantification vectorielle qui recherche au contraire des prototypes dans les zones denses. En second lieu, le nombre de locutions utilisées pour construire un modèle avec prise en compte du contexte est 12 fois plus grand que précédemment. Il est impossible pour des raisons d'espace mémoire de stocker autant de références, or un nombre insuffisant de locutions de référence conduit à une chute du taux de reconnaissance (cf. section 3.1.3.0).

3.2.8 Conclusion

Dans cette partie, nous avons présenté une méthode de reconnaissance de mots qui s'inspire des premiers travaux en RAP (Reconnaissance Automatique de Parole) mais qui a donné des résultats comparables à ceux obtenus en utilisant des modèles de Markov lors de la résolution du problème de la reconnaissance de mots isolés. Elle conduit cependant à des taux de reconnaissance plus faibles lorsqu'il s'agit de la reconnaissance de phrases. Il est maintenant intéressant de comparer les deux méthodes, pour d'une part apprécier leurs avantages respectifs et d'autre part mieux cerner les points faibles communs.

Les modèles de Markov utilisent une topologie fixée a priori, tandis que leurs probabilités de transition entre états ainsi que les densités d'observation à l'intérieur des états sont sujettes à un ajustement statistique. Au contraire, dans la méthode décrite ci-dessus, les observations sont quantifiées au départ tandis que l'apprentissage se réduit à ajouter des locutions codées dans le graphe. En utilisant des HMM, on admet l'hypothèse de l'indépendance Markovienne : « *la probabilité d'observer o à l'état q au*

temps t dépend de l'état où l'on se trouvait au temps t-1 ». Or cette hypothèse est vraiment mauvaise dans le cas de la parole réelle. Le nombre d'états est généralement petit et de ce fait, il suffit de relativement peu de prononciations pour entraîner correctement un modèle simple. On dispose de plusieurs choix pour effectuer l'estimation initiale des paramètres. L'estimation au maximum de vraisemblance conduit à l'amélioration du modèle pour chaque mot. L'inconvénient de cette démarche est de ne pas prendre en compte l'ensemble du lexique à reconnaître. Par conséquent, pour avoir des modèles plus discriminants, on estime les paramètres avec l'algorithme de MMIE ou *Corrective Training* (cf. page 21). La réestimation des paramètres peut se faire de manière bayésienne.

La démarche décrite dans cette partie ne requiert pas l'hypothèse markovienne. Cependant, le nombre d'états est grand, c'est pourquoi il a été utile de quantifier les locutions. En outre, la quantification vectorielle projette les locutions dans le domaine quéfrentiel et l'on perd ainsi de l'information temporelle. Ceci peut partiellement expliquer les résultats relativement mauvais obtenus en reconnaissance de mots enchaînés et l'échec de la prise en compte du contexte. En effet, les zones les plus affectées par la coarticulation sont les extrémités des mots. Les vecteurs acoustiques des extrémités ne sont pas bien représentés par les classes issues de la quantification vectorielle.

Les difficultés rencontrées lors de la reconnaissance de mots enchaînés sont la détermination du nombre de mots et la reconnaissance correcte malgré la coarticulation. Le premier problème est élégamment résolu avec l'algorithme de Viterbi pour les HMM, et par la programmation dynamique à deux niveaux dans la méthode décrite dans cette partie. Cependant la modélisation de la durée est un problème dans les HMM. En effet, sans modèle explicite de la longueur des observations, la probabilité de rester dans un état décroît exponentiellement—à cause de la probabilité de transition inférieure à 1—avec le temps ce qui n'est pas réaliste. La méthode décrite ci-dessus règle ce problème grâce à l'alignement temporel des séquences. Pour pallier l'effet de coarticulation, on peut utiliser plusieurs modèles par entrée lexicale ou utiliser des mixtures de gaussiennes dans le cas de modèles semi-continus. Ceci conduit à une augmentation du nombre de paramètres et nécessite donc un grand corpus d'apprentissage pour que l'estimation soit bonne. Avec les automates, la prise en compte du contexte échoue dans la prise en compte de la coarticulation.

Une autre difficulté est due au changement d'environnement. En effet, les taux de reconnaissance chutent lorsque le locuteur est mal représenté dans l'espace d'apprentissage ou que le bruit ambiant, le microphone... est différent lors de la reconnaissance. Le fait de représenter assez explicitement chaque locution permet aux automates de s'adapter rapidement comme l'a montré l'expérience réalisée en mots isolés. La faiblesse des HMM, qui pallient le manque de représentativité de l'espace d'apprentissage en réestimant de manière bayésienne les paramètres, est le « mélange » à l'intérieur d'une même observation du bruit inhérent à la mesure du signal et des informations concernant le locuteur, l'environnement... lorsqu'on ne dispose pas de modèles paramétriques pour cela.

Le chapitre suivant présente une démarche qui s'est répandue ces dernières années et qui a pour ambition de modéliser la production de la parole, afin de pouvoir extraire du signal acoustique les gestes des différents articulateurs.

Chapitre 4

Utilisation de modèles de production

4.1 Introduction

Les méthodes vues précédemment utilisent en général des techniques de paramétrisation du signal acoustique (analyse LPC, analyse homomorphique...) basées sur un modèle rudimentaire de la production de la parole : le modèle source/conduit. Cependant elles n'imposent pas de contraintes liées au fait que la parole est produite par un conduit vocal. Or le mouvement des articulateurs est soumis à des contraintes physiques dues à l'anatomie, d'où l'idée d'utiliser des modèles physiques de production de parole qui appréhendent correctement le conduit vocal. En outre, si psychologiquement la parole apparaît comme une succession d'unités discrètes, elle est produite par un flux d'air continu expulsé par les conduits nasal et vocal. La recherche d'invariants acoustiques permettant de classifier les unités discrètes laisse place à une recherche des causes de la variabilité lorsqu'on essaie d'intégrer les aspects de production et de perception de la parole [55]. Ainsi la phonologie articulatoire [61] fournit un cadre qui unifie les unités linguistiques et les unités de production par les phonèmes. La variabilité acoustique due aux phénomènes de coarticulation s'explique par des chevauchements de gestes articulatoires.

Rose, Schroeter et Sondhi décrivent aussi le rôle potentiel des modèles de production dans la reconnaissance automatique de parole [86]. Premièrement, les systèmes de RAP la représentent souvent comme une séquence de segments ne se recouvrant pas (systèmes classifiant des trames de parole issues d'une segmentation préalable du signal

acoustique). Ceci ne rend pas compte des phénomènes d'anticipation (par exemple lors de la transition /iy/ avec anticipation de la protrusion des lèvres). Il semble donc plus judicieux de représenter la parole par un flot continu de mouvements articulatoires. D'autre part, la connaissance a priori de l'organisation temporelle des gestes articulatoires permet d'exploiter les propriétés d'invariance liées à la notion d'articulateurs critiques [25]. Ces articulateurs sont ceux pour lesquels un faible déplacement entraîne une grande modification du signal acoustique. Enfin, dans les deux expériences citées par Rose, l'incorporation de données articulatoires a fortement amélioré les taux de reconnaissance (30% dans chaque cas). Dans la première expérience, les images des lèvres et de la langue ont été utilisées, dans la seconde, ce sont des trajectoires articulatoires mesurées grâce à l'EMA (*ElectroMagnetic Articulograph*). Cette dernière technique consiste à placer des bobines sur les articulateurs. Le locuteur doit prononcer les phrases dans un champ magnétique. Le courant induit dans les bobines permet de connaître les trajectoires des différents articulateurs.

4.2 La production de la parole

Dans « Speaking: from intention to articulation » Levelt [53] explique comment l'homme exprime ses intentions au moyen de la parole. La première étape constitue ce que l'auteur appelle la conceptualisation : le locuteur utilise les connaissances qu'il a de lui et du monde (dans sa mémoire à long terme) ainsi que des connaissances sur le discours qu'il conduit (dans sa mémoire à court terme) pour produire un message préverbal. Celui-ci est ensuite encodé à l'aide de la grammaire du langage du locuteur en accédant au lexique pour produire la structure de surface. Puis un plan phonétique ou plan articulatoire est construit grâce à l'encodage phonologique. Vient alors l'articulation, qui consiste à exécuter le plan articulatoire c'est-à-dire à déplacer les muscles qui contrôlent le flux d'air à l'intérieur du conduit vocal, ce qui s'accompagne de l'émission du signal sonore.

Cependant, dans cette partie nous nous intéresserons uniquement aux aspects articulatoires et acoustiques de la production. N'oublions pas en effet que notre but est de reconnaître la parole, en tenant compte des contraintes articulatoires inhérentes à l'anatomie du conduit vocal et ainsi de tenter de modéliser dans une certaine mesure la

variabilité du signal acoustique due à la coarticulation. Pour remonter aux intentions du locuteur, il faut comprendre les phrases prononcées ce qui nécessite de résoudre des problèmes liés à la référence, l'accès au lexique... qui sont d'autres vastes domaines de recherche.

4.2.1 Le conduit vocal

Une coupe sagittale de conduit vocal est représentée sur la figure 4.1. Bien que

FIG. 4.1 – *Coupe sagittale du conduit vocal (voyelle /i/)*

le conduit vocal soit tridimensionnel, la coupe sagittale donne une bonne description de la plupart des phonèmes. En outre l'acquisition de données anatomiques non invasive n'est pas encore chose aisée. Et les données les plus nombreuses sont encore les radiographies de conduits vocaux et par conséquent des vues bidimensionnelles [12]. Cependant on peut fonder de grands espoirs sur l'imagerie par résonance magnétique (IRM) [103]. En effet, si cette technique—qui permet de découper le conduit vocal selon n'importe quel plan—nécessite encore une longue exposition pour collecter une image, ces durées diminuent. Par ailleurs, la qualité de l'image ainsi obtenue est incomparable à la médiocrité des clichés radiographiques.

La parole est produite par l'air qui, expulsé des poumons via la trachée, traverse la cavité pharyngale puis les cavités orales et/ou nasales. Pour les sons voisés, la glotte vibre. Les sons fricatifs sont dûs à un rétrécissement du conduit qui génère des bruits de friction. Les occlusives sont caractérisées par une fermeture totale du conduit durant un bref instant. Les nasales sont produites par l'abaissement du vélum.

Par ailleurs, il semble que les configurations articulatoires dans le plan sagittal soient assez invariantes pour produire certains sons. Si les voyelles standard peuvent être décrites dans l'espace acoustique F1-F2, elles peuvent l'être aussi en fonction du degré d'ouverture (bien qu'il ne soit pas aisé à définir [87]) et de la position de la principale constriction du conduit. En outre, les occlusives aussi sont bien caractérisées par la position de sa fermeture.

Nous présenterons seulement le modèle du conduit vocal. Un modèle de production complet doit aussi intégrer les sources. Il en va ainsi dans l'optique de faire de la synthèse ou du codage articulatoire [97]. Or dans le cas de la reconnaissance de parole (au sens *speech to text*) nous pensons que l'encodage du texte est principalement contrôlé par la forme du conduit vocal, tandis que les sources (le contrôle de la glotte plus précisément) encodent plutôt des informations relatives au locuteur comme son identité et son état émotionnel par exemple.

4.2.2 Les équations acoustiques

Dans cette partie les quantités suivantes : distribution de vitesse volumique $\mathbf{v}$, pression p, densité ρ dépendent des 3 variables spatiales x, y, z et du temps t sauf indication contraire.

Équations de Navier-Stockes et d'Euler

L'écoulement de l'air dans le conduit vocal est régit par les équations de Navier-Stockes [44]

$$\rho[\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\nabla)\mathbf{v}] = -\mathbf{grad}p + \eta\triangle\mathbf{v} + (\zeta + \frac{\eta}{3})\mathbf{grad}\mathrm{div}\mathbf{v} \tag{4.1}$$

où ρ est la densité du fluide, ζ et η leurs coefficients de viscosité (qui sont supposés ne pas dépendre de la température ni de la pression du fluide), $\mathbf{v}$ est sa vitesse volumique

et p sa pression.

Or résoudre cette équation n'est pas une mince affaire, c'est pourquoi des approximations sont faites. Dans un premier temps on peut considérer que l'on a affaire à un fluide parfait ce qui annule ζ et η. Les équations obtenues :

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\nabla)\mathbf{v} = -\frac{\mathbf{grad}p}{\rho} \tag{4.2}$$

sont encore appelées équations d'Euler, ou équations de conservation de mouvement.

Ensuite, la vitesse volumique est faible et on peut négliger le terme $(\mathbf{v}\nabla)\mathbf{v}$ dans ces équations. D'autre part, si la vitesse volumique est faible, alors les variations de pression et de densité sont petites elles aussi. Soit p' la variation de la pression p autour de la pression de référence p_0 et ρ', ρ_0 les grandeurs équivalentes pour la densité, les équations d'Euler se réduisent alors à :

$$\frac{\partial \mathbf{v}}{\partial t} + \frac{\mathbf{grad}p'}{\rho_0} = 0 \tag{4.3}$$

où $p' = p - p_o$ et $\rho_0 = \rho - \rho'$.

Équation de continuité

Cette équation traduit le fait que la matière se conserve. Soit un volume V_o d'air de masse $\int_{V_0} \rho dV$, la masse d'air qui sort de ce volume par unité de temps vaut $\oint_{V_0} \rho \mathbf{v}\mathbf{df}$ où $\mathbf{df}$ est un vecteur normal à la surface du volume V_o orienté vers l'extérieur de V_o dont la norme égale l'aire de l'élément de surface. Or la masse d'air sortant du volume vaut aussi $-\frac{\partial}{\partial t}\int_{V_0} \rho dV$ par unité de temps. On a donc $\frac{\partial}{\partial t}\int_{V_0} \rho dV = -\oint_{V_0} \rho \mathbf{v}\mathbf{df}$. En transformant l'intégrale de surface en intégrale de volume on obtient $\frac{\partial}{\partial t}\int_{V_0} \rho dV = -\int_{V_0} \mathrm{div}\rho\mathbf{v}\mathbf{df}$. Or cette équation doit être vérifiée pour tout volume V, il faut donc que les quantités sous le signe d'intégration soient nulles. Ceci conduit à l'équation de continuité, aussi appelée équation de conservation de masse :

$$\frac{\partial \rho}{\partial t} + \mathrm{div}\rho\mathbf{v} = 0 \tag{4.4}$$

En utilisant les approximations faites sur v, ρ et p nous obtenons l'équation de continuité :

$$\frac{\partial \rho'}{\partial t} + \rho_0 \mathrm{div}\mathbf{v} = \mathbf{0} \tag{4.5}$$

Équation d'état du fluide parfait

Arrivés à ce stade, nous avons 3 équations d'Euler et une équation de continuité pour 5 inconnues: les 3 composantes du vecteur vitesse, la variation de pression et la variation de densité. Il nous faut donc une dernière équation pour résoudre le problème. Dans un fluide parfait le mouvement est un processus adiabatique, ce qui permet de lier ρ' et p' :

$$p' = \left(\frac{\partial p}{\partial \rho}\right)_s \rho' \tag{4.6}$$

où s est l'entropie du fluide.

Équation d'onde

Maintenant posons

$$\mathbf{v} = \mathbf{grad}\varphi \tag{4.7}$$

φ étant le potentiel des vitesses. On déduit alors des équations d'Euler

$$-\frac{p'}{\rho_o} = \frac{\partial \varphi}{\partial t} \tag{4.8}$$

puis l'équation de continuité conduit à l'équation d'onde

$$\frac{\partial^2 \varphi}{\partial t^2} - c^2 \triangle \varphi = 0 \tag{4.9}$$

en posant $c = \sqrt{\left(\frac{\partial p}{\partial \rho}\right)_s}$ (célérité du son).

L'équation 4.9, vérifiée par le potentiel des vitesses, l'est aussi pour chaque composante de la distribution de vitesse (appliquer à l'équation 4.7 l'opérateur **grad**) ainsi que pour p' (appliquer à l'équation 4.8 l'opérateur $\frac{\partial}{\partial t}$) et par conséquent pour ρ' (d'après l'équation 4.6).

Équation de Webster

Jusqu'à présent, nous avons fait des hypothèses sur le flux d'air. Il faut maintenant préciser la géométrie du conduit vocal. Il est souvent modélisé par un tube droit de section circulaire $A(x,t)$. Souvent on considère le tube statique à parois rigides: A ne dépend que de x qui varie de la glotte aux lèvres. La vitesse à l'intérieur du tube vaut

alors $v = \frac{u}{A}$ où u est le débit dans le conduit. Dans ce cas, les équations d'Euler et de continuité mènent à :

$$\frac{\rho_0}{A}\frac{\partial u}{\partial t} + \frac{\partial p'}{\partial x} = 0$$

$$\frac{\partial p'}{\partial t} + \rho_0 c^2 \frac{\partial}{\partial x}\frac{u}{A} = 0$$

En dérivant la seconde équation par rapport au temps nous dérivons la célèbre équation de Webster :

$$\frac{1}{A}\frac{\partial}{\partial x}\left[A\frac{\partial p'}{\partial x}\right] + \frac{1}{c^2}\frac{\partial^2 p'}{\partial t^2} = 0 \tag{4.10}$$

Conduit nasal et cavité sous-glottique

Pour modéliser le conduit nasal, il suffit de considérer 3 tubes. Le premier correspond au pharynx, les deux autres respectivement à la bouche et au conduit nasal. Les conditions à la jonction du pharynx et des deux autres tubes permettent de résoudre l'équation 4.10.

On peut de la même manière prendre en compte la cavité sous-glottique qui introduit des zéros dans le spectre de parole. Cependant, lorsque la glotte est fermée, le couplage entre les cavités sous et supra-glottique est faible et il est possible de négliger la partie inférieure du conduit.

Les sources

Lorsque l'on veut simuler la production de la parole, il faut aussi prendre en compte les sources. Celles-ci sont :

1. la vibration des cordes vocales. Pour les sons voisés, la source quasi-péridodique se situe au niveau de la glotte. Celle-ci peut être simulée grâce au modèle à deux masses vibrantes d'Ishizaka [40] par exemple.
2. le bruit de friction lorsque la section du tube est très faible. Dans ce cas il faut introduire une source au niveau de la constriction la plus étroite dans le conduit. Les zéros dûs à la cavité postérieure à la constriction compensent les pôles pour les fréquences inférieures à 5kHz environ.

3. le bruit des occlusives. Il est dû au relâchement soudain de la pression occasionnée par la fermeture du conduit. L'occlusion complète du conduit vocal s'accompagne d'un bref silence. Puis le conduit vocal s'ouvre et l'air sous pression derrière la constriction s'échappe en créant un bruit turbulent.

4.2.3 Méthodes de résolution

Modèles du conduit vocal

Différents choix peuvent être faits pour la fonction d'aire A. Si l'on considère le tube uniforme (ou une concaténation de tubes uniformes), l'analogie électrique permet d'effectuer les simulations. Celles-ci peuvent s'effectuer dans le domaine temporel, dans le domaine fréquentiel ou encore dans les deux à la fois (en modélisant les sources dans le domaine temporel et le conduit dans le domaine fréquentiel [97]).

Toutefois, quand on s'intéresse plus particulièrement aux modes de résonance du conduit, il n'est pas toujours nécessaire de résoudre l'équation de Webster en calculant le spectre de transfert du conduit vocal.

Si l'on développe en série de Fourier la fonction d'aire [108] ou son logarithme [107], la théorie des perturbations offre un cadre efficace pour calculer les fréquences propres de l'équation de Webster.

Plus généralement, le calcul variationnel [41] n'impose pas de contraintes à la fonction d'aire et fournit une manière efficace pour calculer les fréquences des formants.

Nous allons présenter la simulation par analogie électrique dans le domaine fréquentiel car les formants (fréquences, largeurs de bande et amplitudes) peuvent être calculés efficacement à partir du spectre produit.

Analogie électrique

Représentons le conduit vocal par une concaténation de tubes uniformes. En cherchant les solutions de l'équation d'onde sous la forme d'ondes monochromatiques $\phi(x,t) = \psi(x)e^{j\omega t}$, nous obtenons pour chaque tube :

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\omega^2}{c^2}\psi = 0$$

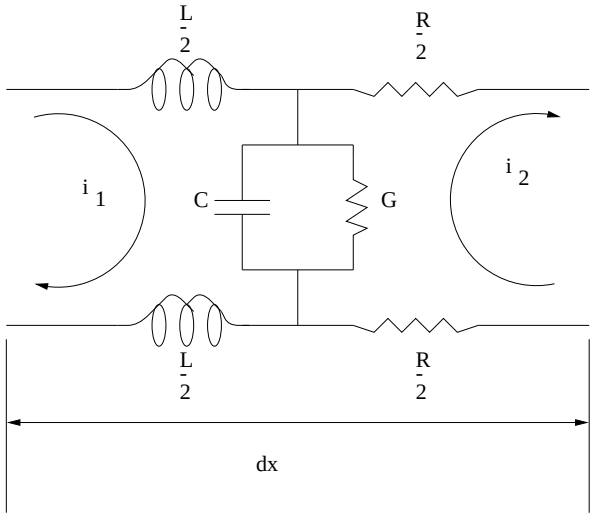


FIG. 4.2 – *Elément de circuit électrique*

Maintenant, en posant $y = G + j\omega C$ et $z = R + j\omega L$ et en considérant les 2 branches dans lesquelles circulent respectivement les courant i_1 et i_2, les équations suivantes sont vérifiées pour la ligne de transmission représentée sur la figure 4.2 :

$$\begin{cases} \mathrm{d}I = -Ey\,\mathrm{d}x \\ \mathrm{d}E = -Iz\,\mathrm{d}x \end{cases}$$

d'où le système :

$$\begin{cases} \frac{\mathrm{d}^2E}{\mathrm{d}x^2} - yzE = 0 \\ \frac{\mathrm{d}^2I}{\mathrm{d}x^2} - yzI = 0 \end{cases}$$

en posant $\gamma^2 = -\frac{\omega^2}{c^2} = yz$ (γ est appelé constante de propagation) nous constatons que ces équations sont analogues à l'équation d'onde, la pression correspondant à la tension E et le débit au courant I. Dans le cas où il n'y a pas de pertes, $\gamma = \pm j\frac{\omega}{c} = \pm j\omega\sqrt{LC}$ est imaginaire pur. Les composants résistifs R et G permettent de tenir compte des pertes thermiques et de celles dues à la viscosité de l'air. Les solutions sont :

$$E = A_1 e^{\gamma x} + B_1 e^{-\gamma x}$$

$$I = A_2 e^{\gamma x} + B_2 e^{-\gamma x}$$

où A_1, A_2, B_1 et B_2 sont des constantes déterminées par les conditions initiales. En considérant une section de longueur l de la ligne de transmission, dont les différences

de potentiel sont E_1 en entrée et E_2 à la sortie, et parcourue par les courants I_1 et I_2 nous obtenons :

$$\begin{cases} E_1 = \sqrt{\frac{z}{y}} I_1 \coth \gamma l + \sqrt{\frac{z}{y}} I_2 \csc \gamma l \\ E_2 = \sqrt{\frac{z}{y}} I_1 \csc \gamma l + \sqrt{\frac{z}{y}} I_2 \coth \gamma l \end{cases}$$

Ce système relie les tensions (équivalent électrique de la pression acoustique) aux courants (équivalent de la vitesse volumique) observés aux bornes du quadripôle (équivalent d'un élément de conduit vocal) sous forme simple. En concaténant de simples tubes, on obtient alors un modèle du conduit vocal. Pour les voyelles, les pressions et les vitesses peuvent ainsi être calculées de proche en proche le long du conduit lorsque l'on dispose d'un modèle de la glotte.

4.2.4 Conclusion

À partir des équations d'Euler (équations 4.2), de continuité (équation 4.5) et d'état (équation 4.6), différentes dérivations ajoutées aux hypothèses sur la fonction d'aire permettent de calculer les fréquences des modes de résonance du conduit vocal sans forcément passer par le calcul de la fonction de transfert. Les approximations faites lors de la dérivation des équations n'influent pas beaucoup sur les fréquences des formants. Comme l'inversion acoustique-articulatoire est souvent basée sur les formants extraits du signal acoustique, toutes ces méthodes peuvent être utilisées en première approximation pour résoudre ce problème. Cependant pour faire de la synthèse de parole il faut prendre en compte les sources ainsi que les pertes qui sont primordiales dans l'évaluation des largeurs de bandes associées aux formants afin d'obtenir un timbre naturel. Ceci peut être accompli en faisant d'autres approximations (voir par exemple [96, 71]) ou en introduisant des impédances pour chacune des pertes dans l'analogie électrique [29]. La section suivante présente différents modèles ayant pour but de contrôler l'aire du conduit avec peu de paramètres.

4.3 Différents modèles articulatoires

Boë, Maeda et Perrier [9] ont présenté une histoire des modèles articulatoires aux XXèmes Journées d'Etude sur la Parole. Ils distinguent dans cet article trois types de

modélisations : géométriques, physiques et articulatoires. Dans cette section, nous allons commencer par expliquer pourquoi il est préférable d'utiliser un modèle articulatoire pour générer la fonction d'aire qui intervient dans les équations acoustiques. Puis nous présenterons brièvement quelques modèles. Pour conclure, nous donnerons les raisons qui nous ont conduit à choisir le modèle de Maeda.

4.3.1 Introduction : Insuffisance des modèles de la fonction d'aire pour la reconnaissance

Comme nous l'avons vu dans la section précédente, il est possible de modéliser le conduit vocal par un ensemble de tubes acoustiques. Or toutes les fonctions d'aire produites par ce type de modèle ne peuvent certainement pas l'être par un humain. Par conséquent, l'espace acoustique résultant de la simulation peut recouvrir l'espace acoustique d'un locuteur quelconque. L'avantage en codage de parole est évident : à chaque son produit par le locuteur humain, il correspond au moins une fonction d'aire et en conséquence, tous les sons de la parole peuvent ainsi être codés. Cependant, l'utilisation de modèles articulatoires en reconnaissance a pour but de retrouver les trajectoires des articulateurs du locuteur. Comme Atal l'a remarqué [3], il existe une infinité de fonctions d'aire qui produisent le même ensemble de formants. Le problème est de choisir parmi elles celle que le locuteur a utilisée pour parler. Un moyen d'éliminer les fonctions d'aire les plus irréalistes est par exemple de fixer la longueur du tube représentant le conduit. Mais dans ce cas, comment rendre compte de la protrusion de lèvres ou de l'abaissement du larynx? En revanche, l'utilisation d'un modèle articulatoire construit à partir d'images de conduits vocaux permet de fixer des contraintes beaucoup plus réalistes. Par exemple le fait que la lèvre inférieure est attachée à la mâchoire... Néanmoins ces modèles doivent être adaptés au locuteur précis qu'ils modélisent afin de pouvoir faire coïncider les espaces vocaliques du locuteur et du modèle. Nous avons ainsi été conduits à adapter le modèle de Maeda pour notre locuteur, à l'aide d'images IRM comme cela est décrit dans ce qui suit. Toutefois, les caractéristiques du conduit du locuteur peuvent aussi être calculées en fonction du signal acoustique notamment grâce à l'étude de l'affiliation des formants aux cavités buccale et pharyngale du conduit vocal [28]. Candille [16] a ainsi adapté le modèle DRM (*Distinctive Region*

Model) et celui de Maeda en calculant un facteur d'échelle optimisé selon le critère de la meilleure adéquation entre les formants du modèle et du locuteur pour la voyelle /y/. Payan et Perrier [79] utilisent les voyelles /i/, /u/, /y/ et /a/ pour normaliser le modèle de Maeda.

4.3.2 Modèles physiques

Certains auteurs représentent la forme des articulateurs par une équation mathématique. Par exemple, Liljencrants [54] représente la langue par les premiers coefficients de la décomposition en série de Fourier de la forme de la langue dans un repère semi-polaire. Ces modèles mettent en évidence le fait que la géométrie de la langue dans le plan sagittal peut être contrôlée avec peu de paramètres. Ainsi les deux premiers coefficients de la série de Fourier suffisent à Liljencrants pour définir le contour du corps de la langue. En utilisant aussi la seconde harmonique, la pointe de la langue peut être modélisée.

Bien que ces modèles donnent une bonne représentation de la forme de la langue, ils ont toutefois quelques difficultés à rendre compte de certains aspects anatomiques du conduit, comme par exemple le fait que la langue soit fixée à la mâchoire.

4.3.3 Modèles géométriques

Ces modèles sont construits à partir de coupes sagittales de conduits vocaux. Les différents articulateurs sont représentés par des formes géométriques (par exemple un cercle pour la langue...). Parmi ces modèles ont peut citer ceux de Cocker [21, 20] et de Mermelstein [67]. L'inconvénient majeur de cette représentation géométrique est qu'elle ne prend pas correctement en compte la souplesse de la langue, principal articulateur.

Pour pallier cette lacune, des modèles statistiques ont vu le jour, comme par exemple celui de Maeda [57, 58]. Ceux-ci sont aussi bâtis à l'aide de radiographies de conduits vocaux. Cependant les paramètres du modèle sont issus d'une analyse statistique des clichés. Ainsi les déformations des articulateurs sont bien représentées dans le plan sagittal. En outre, l'analyse des images a permis de construire un modèle contrôlé à l'aide d'un nombre restreint de paramètres. Dans le modèle de Maeda seulement sept variables suffisent à obtenir une bonne adéquation entre les conduits humains

et synthétiques. Une difficulté rencontrée lors de la construction de tels modèles est bien sûr la lecture précise sur les radiographies des contours servant à construire le modèle. Néanmoins, ils représentent bien la géométrie effective du conduit vocal. Il est toutefois nécessaire de calculer l'aire du conduit à partir du diamètre mesuré dans le plan sagittal afin de résoudre les équations acoustiques. Or la forme du conduit ne varie pas seulement dans le plan sagittal, et d'autre part elle dépend certainement du locuteur. Il est par conséquent impossible de faire exactement coïncider les espaces vocaliques du locuteur et du modèle.

4.3.4 Modèles biomécaniques

Plus récemment, des modèles tridimensionnels ont vu le jour. Les données utilisées proviennent en général de l'imagerie ultra-sonique ou par résonance magnétique. La méthode des éléments finis permet de contrôler les différents muscles. Ces modèles permettent de prendre en compte des contraintes telles que la conservation de la masse de la langue et par conséquent de son volume (en effet on peut considérer qu'elle est incompressible étant donnée l'abondance d'eau qu'elle contient). Or les modèles bidimensionnels sont incapables d'intégrer ces contraintes puisque la langue ne se déforme pas uniquement dans le plan sagittal. En particulier, la profondeur du sillon varie fortement, comme on peut l'observer sur des images IRM mais difficilement sur des radiographies. Par conséquent, la reconstruction de l'aire—qui entre en jeu dans les équations de l'acoustique—à partir des coupes sagittales reste assez approximative malgré les diverses formules proposées [49]. Les modèles biomécaniques n'ont pas cet inconvénient. En revanche ils nécessitent plus de paramètres pour contrôler finement la forme du conduit.

4.3.5 Conclusion

Parmi les nombreux modèles existants, nous avons choisi celui de S. Maeda principalement pour ses qualités anthropomorphiques. En outre il a l'avantage—dans une optique d'inversion acoustique-articulatoire—d'être contrôlé par peu de paramètres. Par ailleurs, l'analyse statistique dont il est issu permet de représenter correctement les liens entre différents articulateurs. Néanmoins, il demeure un modèle de coupes sa-

gittales ce qui rend impossible une adéquation parfaite entre le modèle et le locuteur.

Les modèles tridimensionnels ont comme point fort de fournir directement la fonction d'aire le long du conduit, contrairement aux modèles de coupes qui s'appuient sur une application pour calculer l'aire à partir du diamètre sagittal. Néanmoins, les techniques d'acquisition ne permettent pas encore d'étudier la dynamique du conduit dans les trois dimensions. Or celle-ci est primordiale pour résoudre le problème de l'inversion de manière réaliste.

4.4 Contrôle des paramètres

La parole n'est pas une suite de phonème adjacents. C'est pourquoi il n'est pas suffisant d'avoir un modèle qui puisse produire tous les sons stationnaires de la parole. Il doit aussi être capable de modéliser les phénomènes de coarticulation. C'est le but de la représentation du conduit par une fonction analytique contrôlée par un petit nombre de variables dont les trajectoires sont recherchées pour la reconnaissance. Les paramètres ne varient pas indépendamment les uns des autres, c'est pourquoi leur dynamique est modélisée.

Le modèle DRM

Les commandes transversales permettent de contrôler le déplacement d'une constriction dans le conduit en faisant varier conjointement les sections de certains tubes. La commande longitudinale fait varier également les longueurs des tubes [19]. George, Jospa et Soquet utilisent un réseau de neurones entraîné à l'aide de données acoustiques et articulatoires [33].

Le modèle de Maeda

Bouabana et Maeda ont modélisé la dynamique de la langue dans [14, 13] en utilisant la LPC multi-impulsionnelle, dans le cadre d'une discussion sur l'unité phonétique. Candille [16] a utilisé une sigmoïde par articulateur afin de rendre compte de leur dynamique. Cette fonction a été choisie après l'étude de mesures des mouvements des articulateurs effectuées au moyen d'un articulatograhe électro-magnétique (EMA).

4.5 Inversion acoustique-articulatoire

L'inversion acoustique-articulatoire a pour but le calcul des paramètres du modèle articulatoire à partir du signal acoustique. Ce travail se heurte à plusieurs difficultés. Outre les problèmes liés à l'extraction des indices pertinents dans la parole (en l'occurrence les fréquences de résonance du conduit vocal), plusieurs configurations des articulateurs peuvent produire un signal acoustique équivalent.

L'inversion a été étudiée depuis plusieurs décennies [32], page 46. Outre l'intérêt en reconnaissance, l'extraction de paramètres articulatoires qui évoluent lentement dans le temps permet de coder très efficacement la parole [97]. Différentes méthodes sont utilisées pour conduire l'inversion acoustique-articulatoire.

4.5.1 Inversion par tabulation

Cette technique consiste à créer une table reliant les paramètres acoustiques (formants) aux formes du conduit vocal (paramètres articulatoires).

Par exemple, Larar a utilisé le modèle de Mermelstein pour construire cette table [48], Schroeter, Meyer et Parthasarathy ont utilisé ceux de Mermelstein et de Cocker dans [95]. Yu [109] construit un dictionnaire en utilisant son modèle de fonction d'aire. Candille et Méloni en construisent un à partir du modèle DRM [17].

Ensuite, pour chaque forme acoustique, on recherche la configuration correspondante des articulateurs. Or comme l'explique Atal [3], la correspondance n'est pas biunivoque. Il faut donc imposer des contraintes sur la cinématique et/ou la dynamique des articulateurs afin de choisir les paramètres articulatoires qui donnent la meilleure trajectoire.

4.5.2 Inversion par optimisation

Ces méthodes itératives optimisent des paramètres d'un modèle articulatoire jusqu'à ce que les caractéristiques acoustiques de la parole prononcée et de celle produite par le modèle soient proches. Les paramètres initiaux peuvent être choisis au hasard, mais il est aussi possible de partir d'une solution approchée obtenue par exemple par inversion par tabulation.

Schoentgen et Ciocea [93, 94] utilisent un modèle à tubes simplifié qui leur permet de relier directement les dérivées des trajectoires des longueurs et sections aux dérivées des trajectoires des formants. Afin d'obtenir une solution unique, ils ajoutent au système d'équations qu'ils obtiennent des contraintes.

Dans [33], George, Jospa et Soquet ont appris à un réseau neuronal la relation acoustique-articulatoire. À partir d'une trajectoire acoustique, une trajectoire articulatoire est calculée et optimisée de manière à avoir une bonne correspondance entre les formants calculés et la parole effectivement prononcée.

4.6 Inversion acoustique-articulatoire dans le modèle de Maeda

Cette section décrit les expériences que nous avons menées afin d'obtenir les trajectoires des paramètres du modèle de Maeda à partir des formants extraits de la parole d'un locuteur. Elle présentera successivement une description plus détaillée du modèle, puis la normalisation effectuée pour adapter ce modèle à notre locuteur. Enfin elle décrira la méthode d'inversion et les résultats obtenus.

4.6.1 Le modèle de Maeda

Comme nous utilisons ce modèle dans la suite, nous allons le présenter brièvement ici. Maeda a noté dans [57] que les deux premières composantes principales d'une analyse de tracés de contours de la langue extraits de clichés radiographiques et projetés sur une grille semi-polaire (cf. figure 4.3) permettaient d'expliquer 90% de la variance des données originales. En utilisant les 3 premières composantes principales, il a remarqué que ce taux s'élevait à 98%. Mieux, il a donné une interprétation des 3 composantes principales calculées : l'une d'elles correspond à un déplacement avant/montant et arrière/descendant de la langue, la deuxième à une contraction dans une direction arrière/montante. La troisième correspond à une compression/dilatation de la langue dans le plan sagittal plus difficilement interprétable. C'est pourquoi dans [58] il a utilisé une connaissance a priori du système articulatoire : la langue comporte 3 articulateurs (dorsal, apical et corps de la langue). Elle est fixée à la mâchoire inférieure qui est

contrôlée par un articulateur. Pour utiliser cette connaissance il a utilisé le modèle général de composantes linéaires décrite dans [78]. Stone et al. [99] utilisent aussi une analyse en composantes principales pour les sections transversales de la langue. Ceci leur permet de caractériser certains traits articulatoires des voyelles. Le modèle que nous utilisons comprend en outre deux paramètres pour déterminer l'ouverture et la protrusion des lèvres, et un paramètre pour fixer la hauteur du larynx. Cependant le modèle a été construit à partir des clichés d'une locutrice précise, or les tailles du pharynx et de la bouche différent entre les hommes, les femmes et les enfants [28, 34]. Le modèle est donc muni de deux facteurs d'échelle pour contrôler leur longueur. À partir d'images de notre locuteur, nous avons déterminé ces deux coefficients.

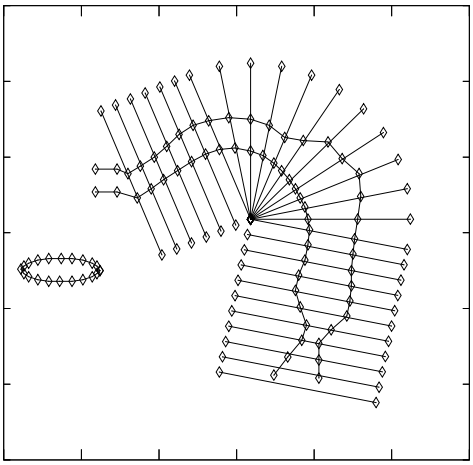


FIG. 4.3 – *Modèle articulatoire original sur sa grille semi-polaire*

4.6.2 Acquisition des images et de la parole

La première étape de notre expérience consiste à adapter le modèle à notre locuteur. Cette étape a pour but de permettre au modèle de recouvrir l'espace acoustique du locuteur [63, 64]. Pour cela, nous avons utilisé des images IRM de son conduit vocal et enregistré la parole correspondante.

Acquisition des images

Nous avons utilisé une machine SIGNA de General Electic pour l'acquisition des images et retenu les mêmes paramètres que Yang [106]. Nous avons réalisé trois séries

d'images pour les voyelles orales françaises /i, y, e, ɛ, œ, ø, a, ɑ, ɔ, o, u/ : une série de coupes sagittales passant par le sillon de la langue et deux dans des plans parallèles à 5mm de distance. En effet, les clichés ayant servi à construire le modèle étant des images aux rayons X, elles ne laissent pas apparaître ce sillon. La série d'images en meilleure adéquation avec le modèle était la série du plan de gauche et nous l'avons retenue pour la suite.

À partir de ces images, nous avons développé un logiciel pour extraire manuellement les contours et les projeter sur la grille semi-polaire permettant de repérer les contours dans le modèle de Maeda. Nous avons ainsi extrait les vecteurs correspondant au contour de la langue, à la paroi dure, la protrusion de lèvres, leur aperture et la hauteur de larynx pour chacune des voyelles.

Enregistrement de la parole

L'inconvénient majeur de l'imagerie IRM est l'intensité du bruit émis par la machine qui interdit l'analyse de la parole produite lors de l'acquisition des images. Nous avons donc enregistré le locuteur dans le bruit, puis nous lui avons demandé de reproduire chacune des voyelles dans le silence tout en écoutant à l'aide d'un casque l'enregistrement du bruit. C'est à partir de cette parole que nous avons extrait les caractéristiques acoustiques utilisées pour l'adaptation. Celles-ci sont les valeurs des fréquences des trois premiers formants, car elles suffisent à différencier les voyelles du français (niveau de la perception) et sont affiliées à des cavités du conduit vocal [60] (niveau de la production). Ces valeurs sont extraites du signal de parole grâce à l'analyse par prédiction linéaire (cf. section 2.2.2). Pour comparer ces valeurs à celles obtenues par simulation, nous avons utilisé la même procédure pour extraire les fréquences des formants produits par le modèle bien que d'autres techniques puissent être utilisées : dérivée seconde du spectre de transfert (cf. [102]), modèle acoustique de la parole (cf. [98], chapitre 4) par exemple.

4.6.3 Adaptation des échelles et estimation du contour extérieur

Le modèle de Maeda a été construit d'après des images d'une locutrice adulte. Or chez les hommes adultes le conduit vocal est plus long que chez les femmes, et par ailleurs le rapport des tailles du pharynx et de la cavité antérieure est plus grand [34]. Notre locuteur étant un homme la première adaptation a porté sur les facteurs d'échelle pour le pharynx et pour la bouche du modèle. Pour cela, nous avons superposé le modèle et les images et modifié ces deux paramètres en se basant sur les contours extérieurs du conduit vocal (la paroi dure). Nous avons trouvé une taille du pharynx de 18% supérieure à celui du modèle, et une bouche 8% plus grande. Ces valeurs ne sont pas en désaccord avec la littérature. Ainsi Fant [28] donne un rapport typique de 11% pour la cavité buccale et 18% pour la cavité pharyngale. Ensuite le contour extérieur du modèle a été recalculé en se basant sur le contour extérieur moyen des images.

L'estimation de la qualité de l'adaptation s'effectue selon deux critères. Le premier concerne la fidélité de la modélisation : la parole synthétisée à partir des paramètres articulatoires extraits des images IRM doit correspondre à celle du locuteur. Toutefois, une erreur est tolérable puisque la parole utilisée n'est pas celle qui a effectivement été prononcée lors de l'acquisition. Le second critère a trait à la couverture de l'espace vocalique du locuteur par le modèle. Ce dernier doit être en mesure d'atteindre les voyelles prononcées par le locuteur. Les deux paragraphes suivants décrivent les résultats obtenus.

4.6.4 Fidélité du modèle

Nous avons extraits les contours des images, puis nous avons estimé les paramètres articulatoires correspondants en cherchant les paramètres qui minimisent l'erreur $|\sum_i (c_i^{extrait} - c_i(\overrightarrow{p}))^2|$ où c_i est la $i^{ème}$ coordonnée du contour intérieur du conduit vocal sur la grille semi-polaire, et $\overrightarrow{p}$ est le vecteur des sept paramètres articulatoires. Comme nous avons plus d'équations que d'inconnues, nous utilisons une décomposition en valeurs singulières afin de résoudre le système [81]. À partir de ces paramètres nous avons effectué une synthèse. Les formants obtenus sont proches de ceux produits par le locuteur.

Les différences peuvent s'expliquer en partie par le fait que le locuteur n'a pas été capable, lors de l'enregistrement de sa parole, de reproduire fidèlement celle qu'il avait produite lors de l'acquisition des images. La table 4.1 illustre les résultats obtenus. Les

	Fs1	Fs2	Fs3		Fp1	Fp2	Fp3		ΔF1	ΔF2	ΔF3
/i/	241	2298	3082		200	2291	2759		41	-7	323
/y/	238	1909	2325		335	1642	2332		97	-267	7
/e/	406	1964	2788		445	2042	2759		39	78	-29
/ɛ/	507	1902	2803		494	1927	2623		-13	25	-180
/œ/	411	1581	2476		469	1547	2360		58	-34	-116
/ø/	500	1517	2599		511	1739	2464		11	222	-135
/a/	594	1157	2659		644	1361	2452		50	204	-207
/ɑ/	652	1178	2757		545	1498	2267		-107	320	-490
/ɔ/	493	1034	2651		536	1202	2324		43	-102	-327
/o/	418	991	2823		326	765	2456		-92	-226	-367
/u/	246	751	2458		197	775	2441		-49	24	-17

TAB. 4.1 – *Comparaison des formants synthétisés et prononcés*

erreurs moyennes respectivement pour F1 F2 et F3 sont de 54Hz, 136Hz et 199Hz. Cependant on remarque que l'erreur maximale est obtenue pour chacun des formants pour une même voyelle. En fait, le locuteur ne l'a évidemment pas correctement reproduite en dépit des précautions que nous avions prises. En retirant cette voyelle, nous obtenons les erreurs moyennes suivantes : 46Hz pour F1, 111Hz pour F2, et 155Hz pour F3. On remarque que pour les voyelles d'arrière le F3 synthétique est systématiquement plus élevé que le F3 naturel, ce que notent aussi Candille et Story [17, 100].

4.6.5 Couverture de l'espace vocalique du locuteur

Les figures 4.4 et 4.5 représentent les projections de l'espace vocalique du locuteur respectivement dans les plans F1-F2 et F2-F3 après le calcul des deux paramètres qui contrôlent les tailles respectives du pharynx et de la bouche. Ils ont été générés en faisant varier les paramètres articulatoires autour de leur position moyenne de manière aléatoire, uniformément de -3 à $+3$ écarts type. Cet espace a ensuite été projeté sur

les plans F1-F2 et F2-F3.

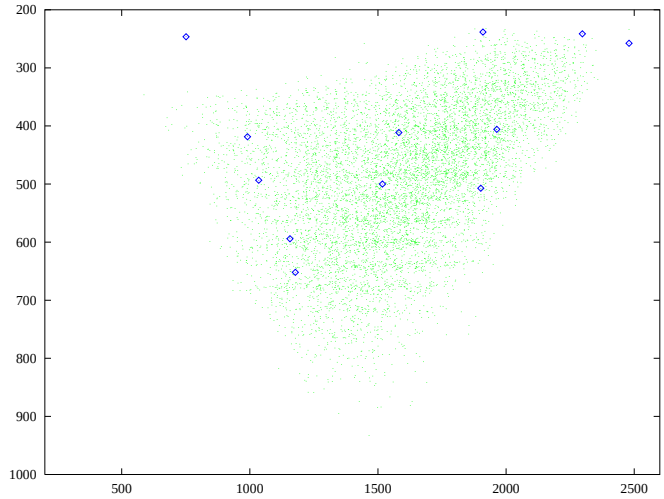


FIG. 4.4 – *Espace acoustique F1-F2 après modification des facteurs d'échelle. Les losanges représentent les voyelles prononcées par notre sujet ; les points représentent l'échantillonnage aléatoire de l'espace articulatoire.*

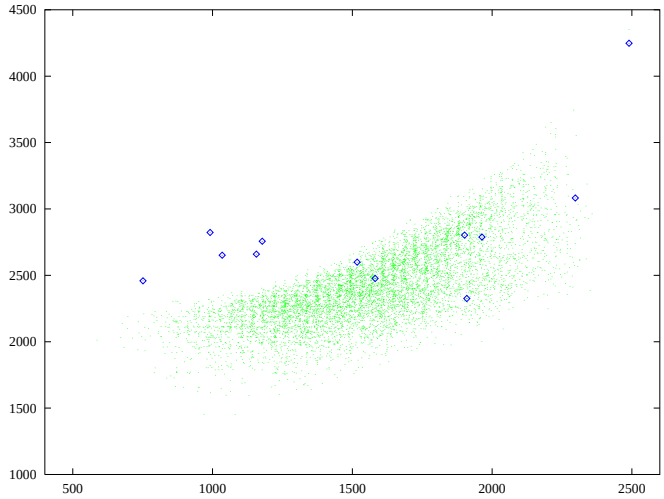


FIG. 4.5 – *Espace acoustique F2-F3 après modification des facteurs d'échelle*

Nous avons recommencé la génération après avoir ré-estimé le contour de la paroi dure. Les figures 4.6 et 4.7 montrent les nouveaux espaces vocaliques. Une amélioration sensible de la couverture est constatée, mais certaines voyelles ne sont toujours pas atteintes.

Plusieurs causes peuvent expliquer cette difficulté qu'a le modèle à atteindre cha-

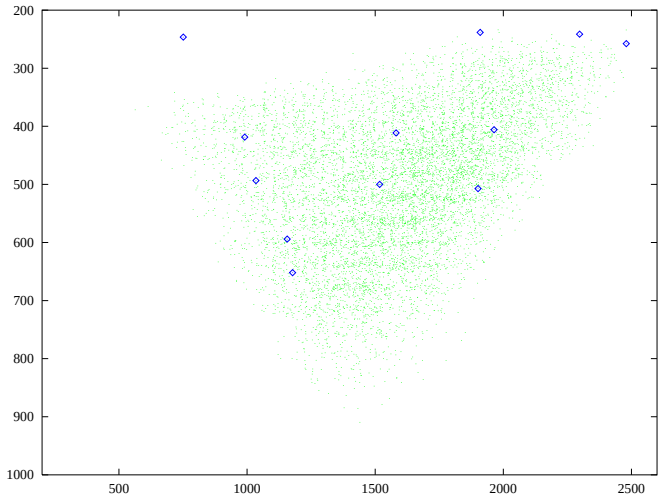


FIG. 4.6 – *Espace acoustique F1-F2 après ré-estimation de la paroi dure*

FIG. 4.7 – *Espace acoustique F2-F3 après ré-estimation de la paroi dure*

cune des voyelles. Une première cause de ces différences est certainement la formule de Heinz et Stevens [36] utilisée pour passer du diamètre sagittal à la fonction d'aire : $A_i = \alpha_i.d_i^{\beta_i}$. En effet, cette formule a été calculée sous l'hypothèse d'une langue plate à partir de cinéradiographies. Il n'y a cependant pas de consensus sur le choix d'une telle fonction [49]. Par exemple, Beautemps et al. [6] ont utilisé une formule différente lors de l'adaptation de leur modèle de coupes sagittales :$A_i = \alpha_{1,i}.d+\alpha_{2,i}.d^{1,5}+\alpha_{3,i}.d^2+\alpha_{4,i}.d^{2.5}$. Une autre raison peut être l'extraction du contour extérieur du conduit vocal. Les images utilisées n'étant pas de même nature que celles ayant servi à construire le modèle, nous avons obtenu un contour extérieur assez différent. Par ailleurs, ne disposant que de onze images, l'estimation du contour moyen doit aussi être entachée d'erreur. C'est pourquoi, au lieu de modifier les coefficients de la formule de Heinz, nous avons opté pour une perturbation de la forme de la paroi dure du conduit.

4.6.6 Adaptation de la paroi dure

Nous avons donc décidé d'optimiser itérativement le contour extérieur dans le but d'obtenir une meilleure adéquation entre la parole naturelle et celle du modèle. Pour cela, nous modifions le contour extérieur à chaque pas de l'itération d'après la fonction de sensibilité donnée par la matrice jacobienne $\frac{\partial F_i}{\partial c_j}$ où c_j est la $j^{ème}$ coordonnée sur la grille semi-polaire. Nous obtenons respectivement pour chacun des formants les erreurs résiduelles suivantes : 49Hz, 125Hz, 170Hz. Les modifications apportées à la paroi dure, à partir de la glotte et en direction des lèvres, sont résumées dans le tableau 4.2. On constate que les diminutions affectent principalement la région centrale du conduit dans la région de la luette (coefficients 10 à 13), ainsi qu'un petit peu le niveau de la glotte (premier coefficient). En revanche, la région pharyngale a dû être « élargie ». Quant à la région buccale, elle n'a pas beaucoup été affectée par cette adaptation.

N° section	1	2	3	4	5	6	7	8	9	10	11	12
Δd (%)	-10,8	-6,4	3,6	8,4	10,5	9,7	5,6	1,2	-5,2	-15,4	-24,9	-14,0
13	14	15	16	17	18	19	20	21	22	23	24	25
-15,8	3,5	6,8	2,0	0,1	0,2	-3,2	-4,2	-1,3	2,4	1,8	1,0	-3,0

TAB. 4.2 – *Nouvelle paroi dure*

Bien que des écarts résiduels subsistent entre les formants du locuteur et ceux de synthèse qui correspondent aux paramètres articulatoires extraits des images IRM, l'espace acoustique du locuteur est maintenant entièrement couvert par le modèle comme le montrent les figures 4.8 et 4.9.

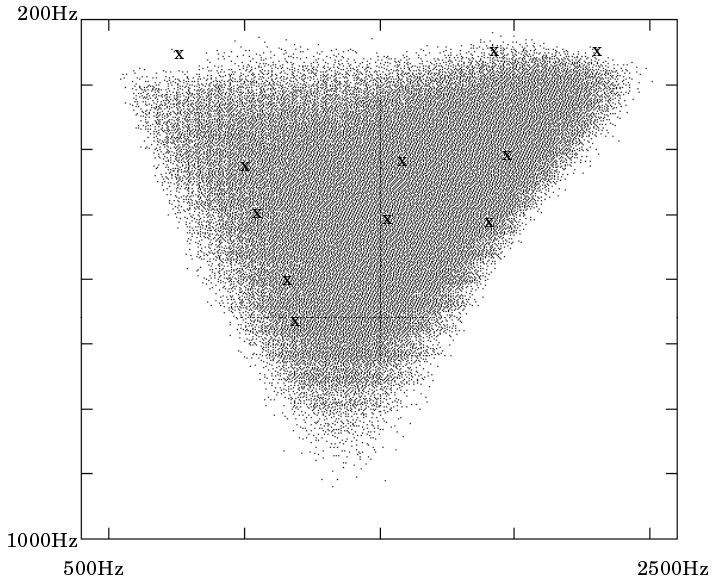


FIG. 4.8 – *Espace acoustique F1-F2 finalement obtenu*

À partir du moment où le modèle parvient à reproduire la parole du locuteur, nous pouvons entreprendre l'inversion proprement dite.

4.6.7 Inversion

Les données utilisées sont des transitions entre voyelles dont les formants ont été extraits par l'algorithme de Laprie [45]. Nous n'avons pas voulu imposer une trajectoire paramétrique, ni des configurations articulatoires initiales et finales pour chaque paramètre contrairement à Candille [16] ou George [32]. Au lieu de cela un dictionnaire nous a fourni des trajectoires initiales que nous avons optimisées à l'aide d'une approche variationnelle.

Nous allons commencer par décrire la construction du dictionnaire, puis nous décrirons l'obtention des trajectoires initiales. Enfin nous présenterons la méthode variationnelle que nous utilisons pour optimiser les trajectoires ainsi obtenues.

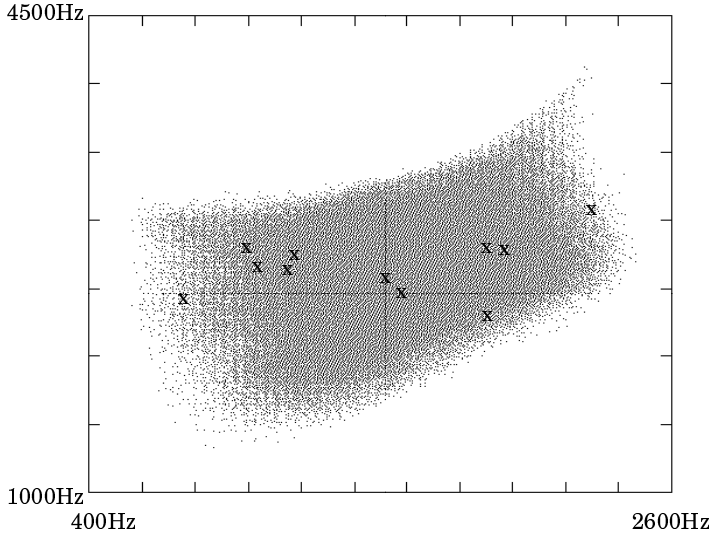


FIG. 4.9 – *Espace acoustique F2-F3 finalement obtenu*

Construction du dictionnaire

Deux méthodes ont été testées. La première consiste à faire varier les sept paramètres aléatoirement entre -3 et $+3$ écarts type. Ne retenant que les entrées pour lesquelles la constriction la plus faible est d'au moins $0,2cm^2$ comme Boë et al.[10], nous en conservons 300000. Nous avons fait varier les paramètres au voisinage de trajectoires rectilignes entre chaque couple de configurations articulatoires obtenues (cf. page 65) pour construire le deuxième dictionnaire. De cette façon nous espérons éliminer des configurations qui ne seraient pas réalistes. Candille [17] a construit un dictionnaire semblable en choisissant des paramètres articulatoires autour de configurations articulatoires de référence associées à chaque voyelle, puis en ajoutant des points situés le long de lignes droites entre couples de configurations articulatoires de référence.

Obtention des trajectoires initiales

À partir de la parole dont nous avons extrait les formants, nous choisissons à chaque instant $t \in [1..T]$ les N meilleures entrées du dictionnaire au sens où elles produisent des formants calculés F_i proches de ceux extraits de la parole f_i. Pour cela nous utilisons une distance euclidienne entre les triplets de formants. Beautemps et Gabioud [7] utilisent une erreur basée sur les fréquences des cinq premiers formants f_i pondérée par une estimation de leurs largeurs de bande S_i mesurées : $E = \sum_{i=1}^{5} \left(\frac{f_i - F_i}{S_i}\right)^2$. Or d'une part il

est difficile d'extraire les cinq premiers formants et d'autre part les trois premiers sont suffisants pour distinguer les voyelles du français.

À partir de ces N^Ttrajectoires potentielles (cf. figures 4.10 et 4.11), nous utilisons l'algorithme de lissage de Ney [74] pour trouver la meilleure trajectoire, c'est-à-dire la plus lisse [47, 46]. En effet, un lissage linéaire ne préserve pas les discontinuités et est sensible au bruit. Dans notre cas il a pour but d'éliminer des points acoustiquement proches de la parole du locuteur, mais pour lesquels il n'y a pas de configuration articulatoire dans le dictionnaire qui permette d'obtenir une trajectoire lisse. Soit $S = (s(0), ..., S(i), ..., S(T))$ l'évolution des paramètres articulatoires en fonction du temps i. Les deux objectifs de l'algorithme sont :

- d'éliminer les instants i pour lesquels il n'existe pas de $S(i)$ permettant d'avoir une trajectoire articulatoire assez lisse,
- de choisir les $K + 1$ points $S(j_k)$ pour chaque instant $j_k \neq i$ de manière à ce que l'évolution des paramètres articulatoires en fonction du temps $\overline{S} = (S(j_0), ..., S(j_K))$ soit lisse.

Bien sûr, une contrainte de monotonie est imposée pour refléter l'écoulement du temps : $j_k < j_{k+1}$. En outre, les conditions aux limites sont choisies telles que $j_0 \geq 0$ et $j_K \leq N$ puisque les points extrêmes peuvent ne pas être conservés. Afin d'obtenir une trajectoire lisse, on pourrait choisir comme critère à minimiser :

$$D = \sum_{k=1}^{K} d(\alpha(j_{k-1}), \alpha(j_k))$$

où d est la distance entre les paramètres articulatoires α aux temps j_k et j_{k-1}. Or tous les termes de la somme sont positifs, donc ce critère ne permettrait que d'avoir une trajectoire plate voire vide ! Pour pallier cet inconvénient, un terme de bonus B est retranché au coût ce qui conduit à minimiser $\sum_{k=1}^{K} d(\alpha(j_{k-1}), \alpha(j_k)) - B$. Plus B est grand, moins la courbe est lissée et plus elle contient de points. Pour l'étude de transitions entre deux voyelles, on choisit B de manière à éliminer les évolutions des paramètres articulatoires non monotones. Le problème de minimisation peut être résolu efficacement en utilisant le principe de la programmation dynamique qui consiste à introduire un coût partiel $D(i) = min_j \left\{ \sum_{k=1}^{K} d(\alpha(j_{k-1}), \alpha(j_k)) - B \, tel \, que \, j_K = i \right\}$.

Cet algorithme a l'avantage de supprimer les points les plus éloignés ce qui permet de remédier dans une certaine mesure aux mauvaises entrées du dictionnaire pour lesquelles un triplet de formants proches de ceux du locuteur existe mais sans paramètre articulatoire correspondant dans le voisinage de ceux trouvés aux instants voisins. Ces trous sont ensuite comblés par interpolation linéaire, ce qui se justifie puisque les trajectoires articulatoires sont très régulières.

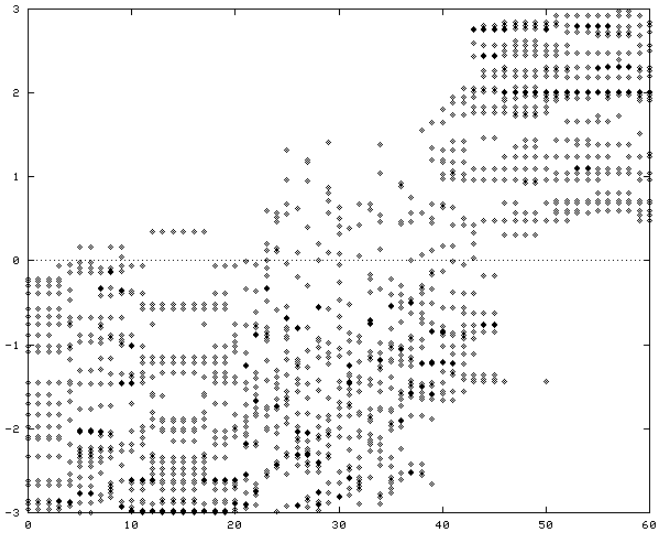


FIG. 4.10 – *Ensemble de trajectoires possibles pour la position de la langue lors d'une transition /iu/ avec le premier dictionnaire*

Les trajectoires possibles en fonction du dictionnaire utilisé sont représentées par les figures 4.10 et 4.11. La figure 4.10 a été obtenue en utilisant le premier dictionnaire pour retrouver les trajectoires possibles du paramètre « position de la langue » lors d'une transition /iu/ en conservant les $N = 25$ meilleures configurations articulatoires. Nous remarquons que les phénomènes de compensation sont bien conservés (la dispersion des positions de la langue s'explique par la compensation due à l'ouverture de la mâchoire [59]) tandis que les trajectoires possibles sont cohérentes : /i/ est bien une voyelle avant comme l'indique la valeur négative du paramètre, tandis que /u/ est une voyelle arrière. La figure 4.11 illustre les trajectoires potentielles en utilisant le deuxième dictionnaire. Comme prévu, les configurations articulatoires ne sont pas aussi dispersées que dans le cas de l'utilisation du premier dictionnaire.

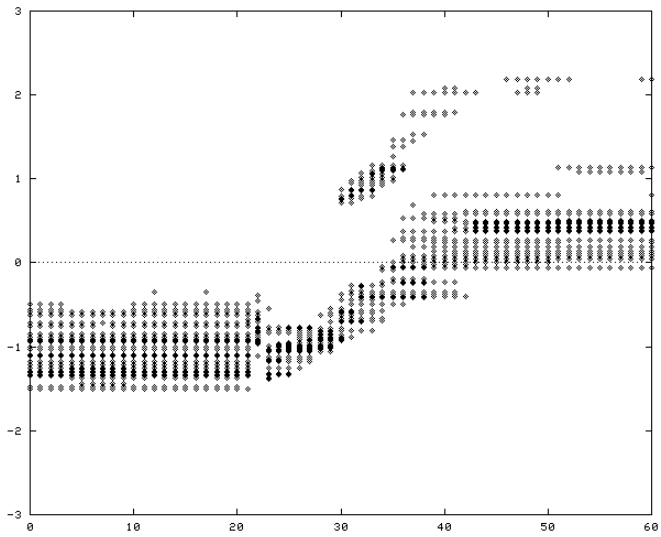


FIG. 4.11 – *Ensemble de trajectoires possibles pour la position de la langue lors d'une transition /iu/ avec le deuxième dictionnaire*

Optimisation des trajectoires initiales

À partir des trajectoires obtenues grâce à l'inversion par tabulation et l'algorithme de Ney, nous allons itérativement améliorer la qualité de l'inversion en utilisant une approche variationnelle. En effet, les trajectoires résultant de l'inversion par tabulation et lissage ne prennent pas en compte simultanément des critères acoustiques et articulatoire. La méthode variationnelle permet de combiner de manière élégante les deux critères [47]. Le problème à résoudre est le suivant :

- obtenir une bonne qualité acoustique, c'est-à-dire minimiser $\sum_{j=1}^{3}(f_j(t) - F_j(\alpha(t))^2$,
- obtenir une trajectoire lisse, c'est-à-dire minimiser $\sum_{i=1}^{7} m_i \alpha_i'^2(t)$; où m_i est la pseudo-masse pour l'articulateur i,
- obtenir une trajectoire articulatoire réaliste, en minimisant $\sum_{i=1}^{7} k_i \alpha_i^2(t)$; où k_i est la pseudo-constante de raideur pour l'articulateur i.

La fonction de coût à minimiser se présente donc sous la forme

$$I = \int_{t_i}^{t_f} \sum_{j=1}^{3}(f_j(t) - F_j(\alpha(t))^2 + \lambda \sum_{i=1}^{7} m_i \alpha_i'^2(t) + \beta \sum_{i=1}^{7} k_i \alpha_i^2(t)\ \mathrm{d}t \tag{4.11}$$

où λ et β sont des constantes qui contrôlent le compromis entre la qualité acoustique et la qualité articulatoire, t_i et t_f sont les bornes de l'intervalle de temps sur lequel on conduit l'optimisation.

L'équation 4.11 peut s'écrire sous la forme $I = \int_{t_i}^{t_f} \Phi(\alpha(t), \alpha'(t), t)\, \mathrm{d}t$ et le calcul variationnel [92] peut être utilisé pour minimiser I. Le minimum de I est atteint lorsque les dérivées relatives aux α_i est nulle. Les équations d'Euler-Lagrange conduisent au système :

$$\begin{cases} \frac{\partial \Phi}{\partial \alpha_1} - \frac{\partial}{\partial t}\frac{\partial \Phi}{\partial \alpha_1'} & = & 0 \\ \ldots & & \\ \frac{\partial \Phi}{\partial \alpha_7} - \frac{\partial}{\partial t}\frac{\partial \Phi}{\partial \alpha_7'} & = & 0 \end{cases} \tag{4.12}$$

Considérant la définition de Φ, chacune des équations de ce système s'écrit :

$$\sum_{j=1}^{3}(f_j(t) - F_j(\alpha(t))\frac{\partial F_j}{\partial \alpha_i}(t) - \lambda m_i \alpha_i''(t) + \beta k_i \alpha_i(t) = 0 \tag{4.13}$$

où $\alpha_i''(t)$ est la dérivée seconde du $i^{ème}$ paramètre articulatoire par rapport au temps.

Considérons maintenant une unique équation de ce système par soucis de clarté ; et supposons que nous disposons d'une solution initiale grossière $\alpha_i^{\tau=0}(t)$ pour la trajectoire de l'articulateur i entre les temps t_i et t_f. Nous définissons un processus itératif tel que $\lim_{\tau\to\infty} \alpha_i^\tau(t) = \alpha_i(t)$ en utilisant l'équation d'évolution :

$$-\sum_{j=1}^{3}(f_j(t) - F_j(\alpha^\tau(t))\frac{\partial F_j}{\partial \alpha_i^\tau}(t) = -\lambda m_i \alpha_i^{\tau\prime\prime}(t) + \beta k_i \alpha_i^\tau(t) + \gamma \frac{\partial \alpha_i^\tau}{\partial \tau}(t) \tag{4.14}$$

où $\frac{\partial \alpha_i^\tau}{\partial \tau}$ représente l'évolution du paramètre α_i^τ durant le processus itératif, dont la vitesse de convergence est contrôlée par γ. Une solution de l'équation 4.13 est obtenue quand le terme $\gamma \frac{\partial \alpha_i^\tau}{\partial \tau}$ est nul. On utilise maintenant une approximation aux différences finies pour calculer les dérivées premières et secondes des paramètres articulatoires. Soit $\alpha_i^\tau = (\alpha_{i,0}^\tau, \ldots, \alpha_{i,k}^\tau, \ldots, \alpha_{i,N}^\tau)$ la représentation discrète de la trajectoire articulatoire du paramètre i au temps $t = t_i + k\frac{t_f - t_i}{N}$, $(f_0, \ldots, f_k, \ldots, f_N)$ la trajectoire acoustique observée et $(F_0, \ldots, F_k, \ldots, F_N)$ la trajectoire acoustique produite par la simulation. L'équation 4.14 ainsi discrétisée conduit à :

$$-\sum_{j=1}^{3}(f_{j,k}-F_{j,k})\left.\frac{\partial F_j}{\partial \alpha_i^\tau}\right|_{\alpha_{1,k},\ldots,\alpha_{7,k}^\tau} = -\lambda m_i(\alpha_{i,k-1}^\tau - 2\alpha_{i,k}^\tau + \alpha_{i,k+1}^\tau) + \beta k_i \alpha_i^\tau(t) + \gamma(\alpha_{i,k}^\tau - \alpha_{i,k}^{\tau-1}) \tag{4.15}$$

Le premier membre de l'équation 4.15 contient l'information acoustique tandis que le second membre représente le comportement articulatoire. Pour obtenir une unique solution au système d'équations 4.12 il nous faut des conditions aux limites. Celles-ci sont choisies de manière à ne pas fixer la position des articulateurs aux extrémités de la fenêtre temporelle, mais plutôt la régularité de la trajectoire articulatoire. On prend donc $\alpha'(0) = \alpha'(N) = 0$. L'équation 4.15 peut donc s'écrire avec ces conditions aux limites sous la forme matricielle :

$$B\alpha^\tau = c^\tau \tag{4.16}$$

où :

$$\left\{\begin{array}{l} B = \begin{bmatrix} \gamma + k_i\beta + m_i\lambda & -\lambda & 0 & \cdots & 0 \\ -\lambda & \gamma + k_i\beta + 2m_i\lambda & -\lambda & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\lambda & \gamma + k_i\beta + 2m_i\lambda & -\lambda \\ 0 & \cdots & 0 & -\lambda & \gamma + k_i\beta + m_i\lambda \end{bmatrix} \\ C = \begin{bmatrix} \gamma\alpha_{i,0}^\tau - \sum_{j=1}^3 (f_{j,0} - F_{j,0})\frac{\partial F_j}{\partial \alpha} \\ \gamma\alpha_{i,1}^\tau - \sum_{j=1}^3 (f_{j,1} - F_{j,1})\frac{\partial F_j}{\partial \alpha} \\ \vdots \\ \gamma\alpha_{i,N}^\tau - \sum_{j=1}^3 (f_{j,N} - F_{j,N})\frac{\partial F_j}{\partial \alpha} \end{bmatrix} \end{array}\right.$$

L'équation 4.16 conduit à un calcul itératif dans lequel on calcule α^τ en fonction de $\alpha^{\tau-1}$. La solution initiale ainsi proposée à l'algorithme tend à s'améliorer dans le sens où d'une part les trajectoires articulatoires sont de plus en plus régulières et l'adéquation entre la parole prononcée et la parole de synthèse augmente. Néanmoins il n'est nullement garanti que le minimum atteint soit celui recherché, c'est pourquoi une solution initiale pour laquelle l'erreur acoustique est déjà faible tout en étant réaliste sur le plan articulatoire est intéressante. Cela justifie que nous ayons utilisé l'algorithme de Ney précédemment décrit. D'autre part, lors des expériences, nous avons fixé les pseudo-masses et pseudo-constantes de raideur à 1. Or ce point est améliorable puisque

la mâchoire a une masse plus importante que les lèvres par exemple. Cet articulateur doit donc se mouvoir plus lentement. Par conséquent la pseudo-masse m_1 correspondant à la mâchoire devrait être supérieure à m_5 qui correspond aux lèvres. Toutefois, pour ajuster ces coefficients, il aurait été intéressant d'obtenir des films articulatoires ce qui était impossible du fait des temps d'acquisition nécessaires à l'obtention d'une image IRM.

Les trajectoires articulatoires obtenues pour les paramètres « position de la mâchoire », « position du dos de la langue » et « ouverture des lèvres » lors des expériences sur la transition /iai/ sont représentées sur la figure 4.12 tandis que la figure 4.13 illustre la précision acoustique obtenue pour cette transition. On peut noter une exagération de l'effet compensatoire entre la mâchoire et les lèvres : ceci peut s'expliquer par le fait que pour passer d'une voyelle d'arrière (/a/) à une voyelle d'avant (/i/) sans tenir compte des inerties respectives de chaque articulateur, le mouvement de la mâchoire est favorisé par rapport à celui des lèvres.

La figure 4.14 présente des trajectoires articulatoires obtenues pour les paramètres « position de la mâchoire », « position du dos de la langue » et « forme de la langue » dans le cas d'une transition /iui/. Les trajectoires des trois premiers formants correspondants sont superposées au spectrogramme de la parole prononcée par le locuteur sur la figure 4.15.

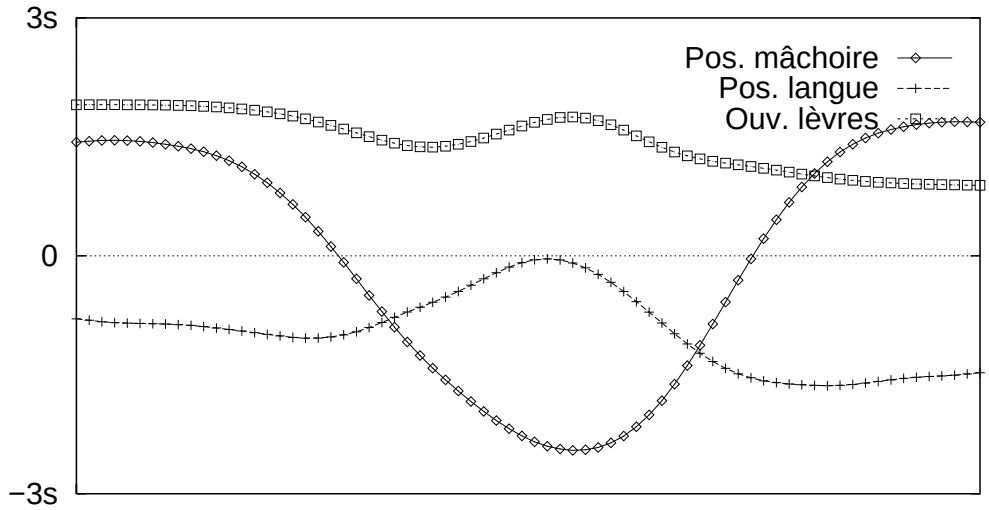


FIG. 4.12 – *Trajectoires articulatoires pour /iai/*

FIG. 4.13 – *Trajectoires formantiques pour /iai/*

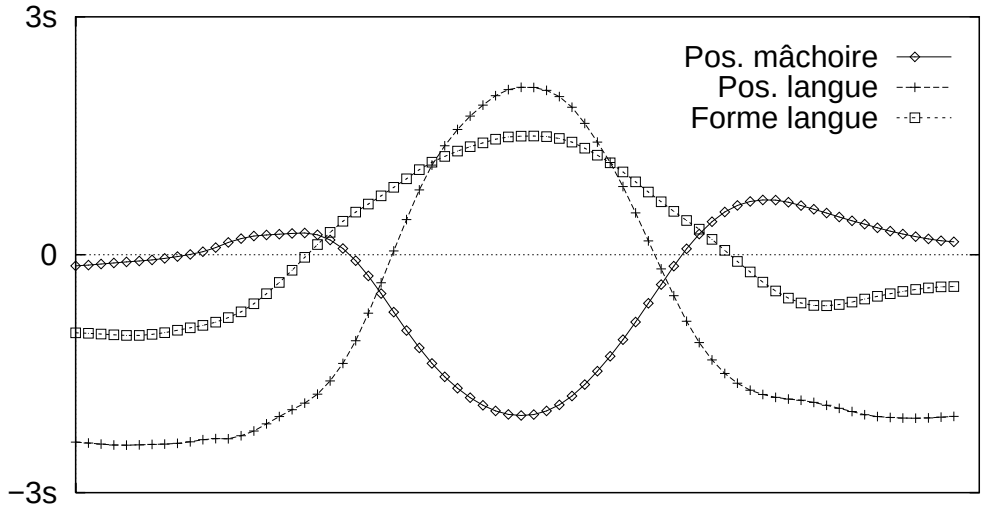


FIG. 4.14 – *Trajectoires articulatoires pour /iui/*

FIG. 4.15 – *Trajectoires formantiques pour /iui/*

4.6.8 Conclusion

Notre méthode d'inversion possède l'avantage de travailler simultanément dans les domaines acoustique et articulatoire. Sans imposer de contraintes paramétriques comme, par exemple, des morceaux de trajectoires sigmoïdales, ni de position initiale et finale des articulateurs, elle conduit à des trajectoires articulatoires lisses correspondant assez fidèlement à la parole du locuteur.

Toutefois, pour pouvoir vérifier le représentativité des résultats obtenus, il nous faudrait disposer de films ce qui n'est pas encore possible avec la technologie utilisée pour l'acquisition d'images. Ceci permettrait par ailleurs de fixer les pseudo-masses m_i et pseudo-constantes de raideur k_i de manière plus réaliste. Un autre point à développer est la recherche de valeurs optimales pour λ et β qui contrôlent le compromis entre la qualité acoustique et la qualité articulatoire dans l'équation 4.11 de manière analogue à ce qu'ont réalisé Richards et al. [85].

Cependant, la méthode décrite fait preuve d'une grande flexibilité ce qui la rend adéquate pour l'étude des effets de compensation. La section suivante décrit brièvement les logiciels utilisés et ceux réalisés pour mener à bien ces expériences d'inversion

acoustique-articulatoire.

4.6.9 Réalisations logicielles

Nous disposions d'une librairie de routines écrites en langage C qui implantent le modèle de Maeda[4] : contrôle des sept paramètres et des deux facteurs d'échelle, calcul des diamètres le long du conduit vocal et passage à la fonction d'aire à l'aide des coefficients de Heinz et Stevens, simulation fréquentielle, simulation temporelle utilisant un modèle de la source glottale. Nous avions aussi à notre disposition le logiciel Snorri (cf. figure 4.16) qui est un outil d'analyse de la parole et permet en particulier d'extraire les formants en implantant différents algorithmes d'analyse de la parole.

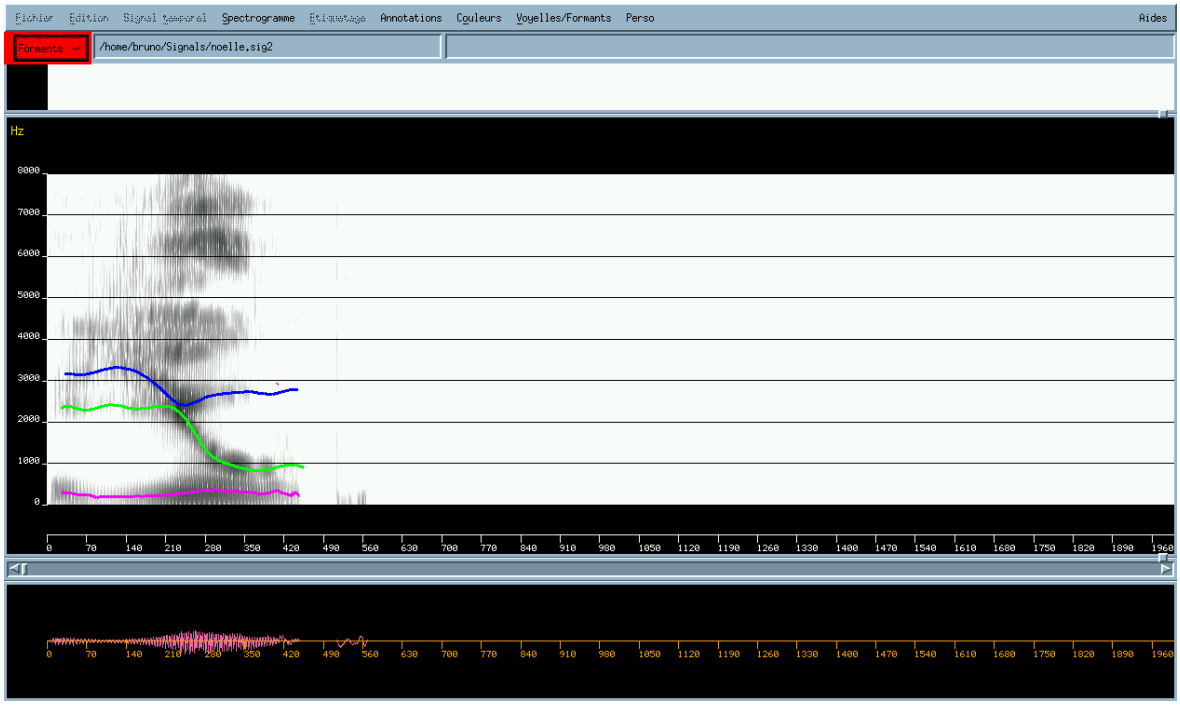


FIG. 4.16 – *Snorri*

Sur cette base, nous avons réalisé une classe en C++ qui encapsule les routines C. Une interface graphique a aussi été développée. L'utilisateur dispose alors d'une boite de dialogue pour changer les valeurs des constantes physiques utilisées lors de la simulation (densité de l'air, pressions...), une autre pour contrôler les paramètres articulatoires du modèle et modifier les facteurs d'échelle respectifs de la bouche et du

4. Nous tenons donc vivement à remercier Shinji Maeda qui nous a fourni cette librairie.

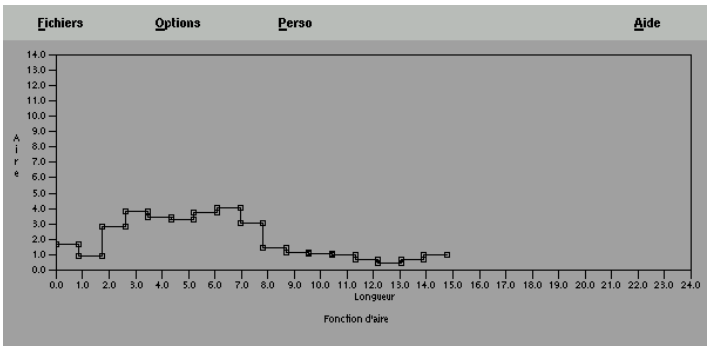


FIG. 4.17 – *Fenêtre principale*

pharynx. La fenêtre principale affiche la fonction d'aire (figure 4.17). Elle comporte une barre de menus à partir de laquelle il est possible d'accéder aux boites de dialogues et de synthétiser puis d'écouter la voyelle correspondant aux paramètres articulatoires. Depuis ces menus, il est encore possible de donner un triplet de formants à atteindre à partir de la configuration articulatoire courante, soit en « optimisant » la paroi dure, soit en recherchant des paramètres articulatoires proches de la position initiale. Le calcul des fonctions de sensibilité (cf. [8]) et leur affichage (ΔF_i en fonction de Δp_j) est aussi disponible. La fonction d'aire est synchronisée avec la coupe sagittale mais l'utilisateur peut aussi la modifier interactivement. Les valeurs des formants (qui sont synchronisées avec la fonction d'aire) sont affichées avec le spectre de transfert et la coupe sagittale du conduit dans une autre fenêtre (figure 4.18). Enfin, les résultats de l'inversion peuvent être récupérés pour animer la coupe sagittale.

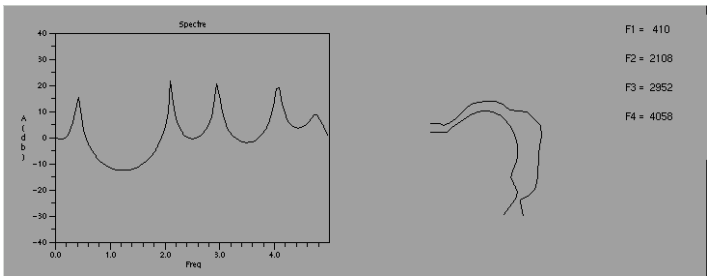


FIG. 4.18 – *Coupe sagittale, spectre et formants*

Un autre programme a été développé pour extraire manuellement les contours du conduit vocal à partir des images IRM du conduit vocal, projeter les contours ainsi obtenus sur la grille semi-polaire puis calculer les paramètres articulatoires du modèle

de Maeda qui produisent le contour le plus proche.

Le filtrage non-linéaire de Ney a été implanté par un programme qui prend en entrée un dictionnaire et une trajectoire acoustique. Il est paramétré par la valeur du bonus (cf. section 4.6.7). Il fournit le fichier initial utilisé par le programme de calcul variationnel.

L'optimisation a été implantée sous la forme d'un autre programme qui prend en entrée un fichier contenant la trajectoire initiale, c'est-à-dire pour chaque pas de temps, d'une part un ensemble de paramètres articulatoires initialement filtré par l'algorithme de Ney, et d'autre part les trois formants extraits de la parole du locuteur. Elle fournit un nouveau fichier à chaque pas d'itération, dans lequel les paramètres articulatoires ont évolué. Ce fichier peut être utilisé pour animer le modèle dans l'interface graphique.

Chapitre 5

Conclusion et perspectives

Ce document a tout d'abord présenté les traitements acoustiques les plus couramment utilisés en vue de reconnaître la parole ; basés sur la transformée de Fourier ou la prédiction linéaire, ils sont utilisés pour extraire des indices permettant de classifier les différents sons de parole ou comme un simple moyen de représenter de manière concise l'information pertinente pour la RAP. Cependant d'autres représentations existent. Heitz et Becker [37] utilisent ainsi une représentation temps-fréquence qui réduit un signal de parole à un ensemble de points interprétables aussi bien pour les sons stationnaires que les sons non stationnaires.

Nous avons ensuite décrit quelques architectures et les différents principes appliqués aux problèmes spécifiques de la reconnaissance de la parole. Ainsi, le principe de la programmation dynamique utilisé dans les premiers systèmes de reconnaissance permet de prendre en compte les différences de durée lors de diverses prononciations d'un mot. Il a été repris dans l'algorithme de Viterbi pour les modèles de Markov. Mais ces derniers ont l'avantage de disposer d'un algorithme d'apprentissage qui permet d'ajuster automatiquement les paramètres du modèle à partir d'exemples. Les réseaux neuromimétiques disposent eux aussi de tels algorithmes, qui conduisent à des modèles fortement discriminants. Mais il est plus difficile de modéliser les dilatations temporelles de la parole dans ce cadre. C'est pourquoi ils sont souvent couplés aux autres architectures. D'un autre côté, l'approche analytique consistant à rechercher des invariants dans le signal acoustique puis à modéliser l'expertise humaine a beaucoup contribué à une meilleure compréhension des problèmes de perception et compréhension de la pa-

role. Ces architectures pourront être utilisées avec des paramètres articulatoires issus d'une inversion au lieu d'utiliser directement le signal acoustique. Ainsi, Li Deng [25] utilise déjà un modèle de Markov dont les états renferment des informations articulatoires. Les réseaux neuromimétiques pourront servir à apprendre la relation inverse comme cela est fait par George, Jospa et Soquet [33] entre autres.

Le chapitre suivant a détaillé les premières expériences. Quoique basées sur des travaux anciens, elles permettent d'obtenir pour les mots isolés et en l'absence de bruit des taux de reconnaissance tout à fait honorables et comparables à ceux obtenus à l'aide de modèles de Markov. En revanche, la reconnaissance de mots enchaînés n'a pas abouti à des taux de reconnaissance équivalents. L'estimation statistique des paramètres acoustiques du modèle semble donc favorable à la prise en compte de certains effets de coarticulation. Malgré cela, lorsque le vocabulaire est de grande taille, les modèles de Markov utilisant des paramètres purement acoustiques ont des difficultés à résoudre le problème de la reconnaissance de phrases indépendamment du locuteur. Ainsi les premiers logiciels grand public se voulant par définition indépendants du locuteur ont laissé place à des systèmes « facilement entraînables par n'importe quel locuteur ». Pour cela l'utilisateur doit prononcer des phrases qui permettent au système de ré-estimer les paramètres du modèle.

La variabilité acoustique de la parole, et en particulier celle due aux effets de coarticulation, est mieux appréhendée par la modélisation de sa production. En effet, la parole n'est pas un signal acoustique quelconque et les contraintes anatomiques peuvent expliquer les effets de coarticulation, par exemple, dans le cadre de la phonologie articulatoire. Par ailleurs, des expériences confirment le fait que la prise en compte d'informations articulatoires (reconnaissance multimodale, par exemple en filmant le visage du locuteur) améliore sensiblement les taux de reconnaissance. C'est pourquoi la deuxième série d'expériences a porté sur l'utilisation d'un modèle articulatoire. Dans le cadre de la reconnaissance de la parole, il s'agit de retrouver les paramètres du modèle à partir du signal acoustique : c'est le problème de l'inversion acoustique-articulatoire. Avant de présenter quelques méthodes de résolution de ce problème, nous avons décrit les équations acoustiques régissant l'état de l'air dans le conduit vocal, puis nous avons présenté brièvement quelques modèles. Celui de Maeda a été retenu pour ses qualités anthropomorphiques et le nombre réduit de paramètres qui le contrôlent. Or il a été

construit à partir d'images aux rayons X d'une locutrice. Nous avons donc commencé par l'adapter à partir d'images IRM d'un locuteur. Les premières difficultés rencontrées tiennent à la nature différente des images utilisées et à l'impossibilité d'exploiter la parole produite durant l'acquisition des images. Cependant, Mohommad et al. [68] ont récemment proposé une technique pour filmer le conduit vocal en mouvement.

La mesure des tailles du pharynx et de la bouche sur les images a conduit à des valeurs conformes à la littérature. Mais cette étape n'a pas suffit à remplir les critères de qualité considérés qui sont, d'une part, une bonne couverture par le modèle de l'espace vocalique du locuteur et d'autre part, une bonne fidélité du modèle. Nous avons donc modifié le contour extérieur du conduit vocal à partir des images IRM ce qui a sensiblement amélioré la qualité. Enfin, nous avons abordé le problème du passage de la coupe sagittale à la fonction d'aire. Au lieu de modifier encore les coefficients de Heinz et Stevens, nous avons opté pour une optimisation du contour extérieur puisque le peu d'images en notre possession ne nous a pas permis de le considérer comme parfaitement digne de confiance. Le passage de la fonction d'aire à la coupe sagittale, étudié par Lecuit dans [50], est un épineux problème qui pourra se résoudre grâce à l'amélioration de la qualité des images. En effet, l'utilisation de profils sagittaux est dû en grande partie au fait que seules des images aux rayons X étaient disponibles. Stone [99] a classifié les voyelles en utilisant une coupe transversale dans la région du palais. Ainsi, les profils des coupes sagittales et transversales permettent de classifier les voyelles. Cependant Story et al. [100] ont mesuré le volume total du conduit vocal en utilisant l'imagerie IRM pour des sons stationnaires. Or ils mettent en évidence le besoin de modèles physiques plus précis. En effet, une concaténation de tubes ne suffit pas pour rendre compte de la forme du conduit vocal lors de la prononciation d'un /l/ par exemple.

Après toutes ces adaptations, le modèle est capable de produire les voyelles du locuteur, et la parole produite à partir des paramètres extraits des images est assez proche de celle du locuteur. Nous avons donc pu entreprendre l'inversion. La méthode proposée est basée sur le calcul variationnel et permet d'optimiser une solution initiale en tenant compte simultanément des contraintes acoustiques et articulatoires. La première étape de l'inversion a été la recherche d'une solution initiale. Pour cela nous avons construit des dictionnaires de formes articulatoires et sélectionné un ensemble

de trajectoires articulatoires potentielles qui correspondent à un signal proche de celui prononcé. La diversité des trajectoires possibles correspond à la capacité du modèle à prendre en compte les effets de compensation. Puis nous avons utilisé le lissage non-linéaire de Ney pour choisir parmi cet ensemble une trajectoire assez régulière. Nous avons ensuite décrit l'algorithme d'optimisation. Les résultats obtenus montrent qu'il est possible d'obtenir une trajectoire articulatoire lisse et dans le même temps une bonne correspondance entre les formants de synthèse et les formants extraits du signal, et ceci sans imposer de formule analytique ni de conditions portant sur les positions des articulateurs aux extrémités du segment temporel sur lequel s'effectue l'inversion. Yehia et Itakura [107] ont aussi utilisé une méthode combinant des contraintes acoustiques et articulatoires dans un même espace de paramètres. Ils soulignent le besoin d'inclure des contraintes dynamiques. Dans notre approche, ceci se traduira par une évaluation des pseudo-masses et pseudo-constantes de raideur.

Si l'on souhaite se placer dans le cadre de la dynamique des tâches [66, 65, 90], il faudra ensuite interpréter les combinaisons de paramètres articulatoires contrôlant le modèle en termes de *variables* articulatoires pertinentes au niveau psychoacoustique (endroits et degrés des points de constriction dans le conduit vocal en fonction de paramètres du modèle de Maeda). Il faudra ensuite déterminer les trajectoires de ces variables et leurs instants d'activation. Les différentes synchronisations possibles expliquant la variabilité de coarticulation.

On pourra aussi simplement utiliser directement les paramètres articulatoires pour effectuer une reconnaissance « multimodale » de la parole. Au lieu de filmer le conduit vocal en mouvement, on utiliserait les résultats de l'inversion...

Annexes

Images IRM du conduit vocal

Pour chaque voyelle, trois coupes ont été réalisées :

- une coupe mid-sagittale (au centre),
- une coupe à $5mm$ à gauche face au locuteur,
- une coupe à $5mm$ à droite (image retenue).

Voyelle /i/

Voyelle /y/

Voyelle /e/

Voyelle /ɛ/

Voyelle /œ/

Voyelle /ø/

Voyelle /a/

Voyelle /ɑ/

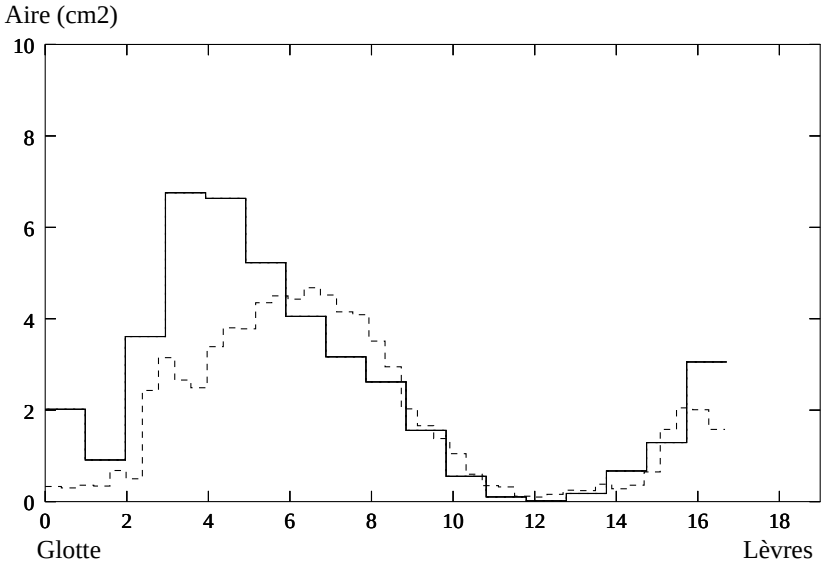
Voyelle /ɔ/

Voyelle /o/

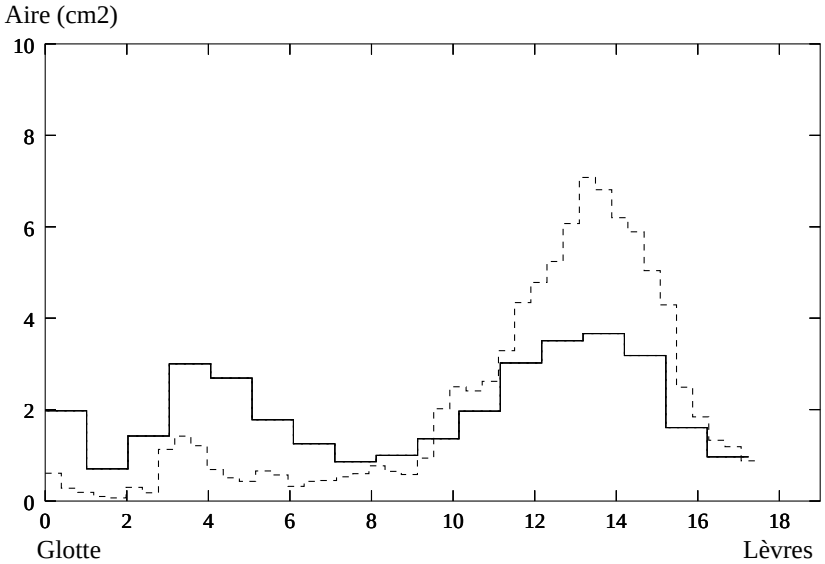
Voyelle /u/

Fonctions d'aire

Les figures suivantes donnent une comparaison des fonctions d'aire obtenues par reconstruction du volume du conduit (d'après Story et Titze [100]) avec celles obtenues grâce au modèle de Maeda normalisé, à partir des paramètres articulatoires extraits des images IRM.



Voyelle /i/



Voyelle /ɔ/

Bibliographie

[1] M. Afify, Y. Gong, J.-P. Haton. « Estimation of Mixtures of Stochastic Dynamic Trajectories : Application to Continuous Speech Recognition ». *Computer, Speech and Language*, 1995. to appear.

[2] Y. Anglade. « *Robustesse de la reconnaissance automatique de la parole : étude et application dans un système d'aide vocal pour une standardiste mal-voyante* ». PhD thesis, CRIN&INRIA, Nancy, France, 1994.

[3] B.S. Atal, J.J. Chang, M.V. Mathews, J.W. Tukey. « Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique ». *J. Acoust. Soc. Amer.*, 63, 1535–1555, 1978.

[4] L.R. Bahl, P.F. Brown, P.V. de Souza, R.L. Mercer. « A New Algorithm for the Estimation of Hidden Morkov Model Parameters ». *ICASSP-88*, 493–496, 1988.

[5] L.E. Baum, T. Petrie, G. Soules, N. Weiss. « A Maximization Technique occurring in the Statistical Analysis of Probabilistic Functions of Markov Models ». *The Annals of Mathematical Statistics*, 41, 164–171, 1970.

[6] D. Beautemps, P. Badin, G. Bailly, A. Galvan, R. Laboissière. « Evaluation of an Articulatory-Acoustic Model Based on a Reference Subject ». *1st ESCA Tutorial and Research Workshop on Speech Production Modeling.*, 45–48, Autrans, France, 1996.

[7] D. Beautemps B. Gabioud. « Adaptation d'un modèle articulatoire à un locuteur dans le but de contraindre l'inversion articulatori-acoustique ». *XXèmes Journées d'Etude sur la Parole*, 119–124, Trégastel, France, 1994.

[8] L.-J. Boë, P. Badin, P. Perrier. « Fron sensitivity functions..... to macro-variations ». *ICPhS*, 2, 234–237, Stockholm, 1995.

[9] L.-J. Boë, S. Maeda, P. Perrier. « La modélisation articulatoire : un demi siècle d'évolution entre fonctionnel, physique et biomécanique ». *XXèmes Journées d'Etudes sur la Parole*, 41–54, 1994.

[10] L.-J. Boë, P. Perrier, G. Bailly. « The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion ». *Journal of Phonetics*, 20:27–38, 1992.

[11] E.L. Bocchieri J.G. Wilpon. « Disciminative Feature Selection for Speech Recognition ». *Computer Speech and Language*, 7:229–246, 1993.

[12] A. Bothorel. *Cinéradiographie des voyelles et consonnes du français.* Publications de l'Institut de Phonétique de Strasbourg, 1986.

[13] S. Bouabana. « Modélisation des mouvements articulatoires de la langue par la méthode de la LPC Multi-impulsionnelle ». Master's thesis, ENST, 1995.

[14] S. Bouabana S. Maeda. « Modélisation des mouvements articulatoires par la méthode de la MLPC et ses effets sur les caractéristiques acoustiques ». *XXèmes Journées d'Etude sur la Parole*, 439–444, Trégastel, France, 1994.

[15] J. S. Bridle, MD. Brown, R. M. Chamberlain. « An algorithm for connected word recognition ». *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing 1982*, 899–902, Paris, France, 1982.

[16] L. Candille. « *Modèles de Production et Reconnaissance Automatique de la Parole* ». PhD thesis, Univ. d'Avignon et des Pays de Vaucluse, 1996.

[17] L. Candille H. Méloni. « Automatic speech recognition using procuction models ». *ICPhS 95*, 4, 256–259, 1995.

[18] R. Cardin, Y. Normandin, E. Millien. « Inter-Word Coarticulation Modeling and MMIE Training for Improved Connected Digit Recognition ». *ICASSP-83*, 243–246, 1993.

[19] S. Chennoukh R. Carré. « Vowel-Vowel production on a Distinctive Region Model. A new command strategy ». *ICPhS*, 1, 78–81, Stockholm, 1995.

[20] C. H. Coker. Synthesis by Rule from Articulatory Parameters. J. L. Flanagan L. R. Rabiner, , *Speech Synthesis*, 396–397. Dowden, Hutchinson & Ross, 1973.

[21] C. H. Coker. « A model of articulary dynamics and control ». *Proc. of the IEEE*, 64(4):452–460, 1976.

[22] R. Cole, B. T. Oshika, M. Noel, T. Lander, M.Fanty. « Labeler agreement in phonetic labeling of continuous speech ». *ICSLP*, 2131–2134, 1994.

[23] S. B. Davis P. Mermelstein. « Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences ». *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-28(4):357–366, 1980.

[24] J. de Veth, G. Gallopyn, H. Bourlard. « Limited Parameter Hidden Markov Models for connected Digit Speaker Verification over Telephone Channels ». *ICASSP-93*, 2, 247–250, Minneapolis, 1993.

[25] L. Deng D. X. Sun. « A statistical approach to automatic speech recognition using the atomic speech units contructed from overlapping articulatory features ». *J. Acoust. Soc. Am*, 95(5):2702–2719, 1994.

[26] J. di Martino, J.-F. Mari, B. Mathieu, K. Perot K., K. Smaïli. « Which Model for Future Speech Recognition Systems: Hidden Markov Models or Finite-State Automata ». *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1, 633–635, Adelaïde, Australia, 1994.

[27] J. L. Elman. « Finding structure in Time ». *Cognitive Science*, 179–211, 1990.

[28] G. Fant. « Vocal-Tract Area and Length Perturbations ». *STL-QPSR*, 1–14, 1975.

[29] J. L. Flanagan. *Speech Analysis Synthesis and Perception.* Springer-Verlag, New-York, 1972.

[30] D. Fohr. « APHODEX : Un système expert en décodage acoustico-phonétique de la parole continue ». *Thèse de Doct. Univ. de NANCY 1*, 1986.

[31] D. François. « *Détection et identification des occlusives et fricatives au sein du système indépendant du locuteur APHODEX.* ». PhD thesis, CRIN&INRIA, Nancy, France, 1995.

[32] M. George. « *Analyse du signal de parole par modélisation de la cinématique de la fonction d'aire du conduit vocal* ». PhD thesis, Univ. Libre de Bruxelles, 1997.

[33] M. George, P. Jospa, A. Soquet. « Estimation de trajectoires articulatoires à l'aide d'un réseau de neurones ». *XXèmes Journées d'Etude sur la Parole*, 427–432, Trégastel, France, 1994.

[34] U. G. GOLDSTEIN. « *An Articulatory Model for the Vocal Tracts of Growing Children* ». PhD thesis, Massachusetts Institute of Technology, 1980.

[35] J.-P. HATON. « Les modèles neuronaux et hybrides en reconnaissance automatique de la parole : état des recherches ». *Fondements et perspectives en reconnaissance automatique de la parole*, 139–153, 1995.

[36] J. M. HEINZ K. N. STEVENS. « On the relations between lateral cineradiographs, area functions and acoustic spectra of speech ». *5th International Congress on Acoustics*, A44, Liège, 1965.

[37] C. HEITZ J. D. BECKER. « An optimized time-frequency distribution for speech analysis ». *Speech Communication*, 1994.

[38] H. HERMANSKY, B.A. HANSON, H. WAKITA. « Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain ». *Speech communication*, 4, 181–187, 1985.

[39] H. HERMANSKY, N. MORGAN, A. BAYYA, P. KOHN. « Compensation effect of the communication channel in auditory-like analysis of speech ». *Eurospeech*, 1367–1370, 1991.

[40] K. ISHIZAKA J. L. FLANAGAN. « Acoustic properties of a two-mass model of the vocal cords ». *Bell Syst. Technol. J.*, 51:1233–1268, 1972.

[41] P. JOSPA. « Formulation variationnelle du lien acoustico-articulatoire ». *XXèmes Journées d'Etude sur la Parole*, 113–118, Trégastel, France, 1994.

[42] Biing-Huang JUANG Shigeru KATAGIRI. « Discriminative Learning for Minimum Error Classification ». *IEEE Trans. on Signal Proc.*, 40, 3043–3054, 1992.

[43] R. KNESER J. PETERS. « Semantic Clustering for Adaptive Language ModelingI ». *Proc. of ICASSP*, 779–782, 1997.

[44] L. LANDAU E. LIFCHITZ. « *Mécanique des fluides* », Chapitre 2, page 69. Librairie du globe, 1989.

[45] Y. LAPRIE M.-O. BERGER. « Cooperation of Regularization and Speech Heuristics to Control Automatic Formant Tracking ». *Speech Communication*, 19(4):255–270, 1996.

[46] Y. LAPRIE B. MATHIEU. « Inversion acoustique articulatoire par une méthode

variationnelle ». *Actes des 22èmes Journ/'ees d'Etude su la Parole*, 295–298, Martigny, Switzerland, 1998.

[47] Y. LAPRIE B. MATHIEU. « A variational approach for estimating vocal tract shapes from the speech signal ». *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2, 929–932, Seattle, USA, 1998.

[48] J.N. LARAR, J. SCHROETER, M.M. SONDHI. « Vector Quantization of the Articulatory Space ». *IEEE trans. on Acoustic, Speech and Signal Processing*, 36, 1812–1818, 1988.

[49] V. LECUIT. « *Sagittal cut to area function transformation: a comparative study* ». PhD thesis, Mémoire de licence en sciences physiques, Univ. Libre de Bruxelle, 1995.

[50] V. LECUIT. « Conséquences acoustiques du passage de la coupe sagittale à la fonction d'aire ». *XXIèmes JEP*, 75–78, Avignon, 1996.

[51] K.-F. LEE, H.-W. HON, M.-Y. HWANG, S. MAHAJAN, R. REDDY. « The SPHINX Speech Recognition System ». *ICASSP-89*, 445–448, 1989.

[52] Kai-Fu LEE Sanjoy MAHAJAN. « Corrective and reinforcement learning for speaker-independent continuous speech recognition ». *Computer Speech and Language*, 231–245, 1990.

[53] W.J.M. LEVELT. *Speaking: From intention to articulation.* MIT Press, 1989.

[54] J. LILJENCRANTS. « Fourier series description of the tongue profile ». *Speech Transmission Laboratory, QPSR*, (4):9–18, 1971.

[55] B. LINDBLÖM. « Approche intégrée de la production et de la perception ». *Fondements et perspectives en reconnaissance automatique de la parole*, 1–8, 1995.

[56] Y. LINDE, A. BUZO, R. M. GRAY. « An algorithm for the vector quantizer design ». *IEEE. Trans. on Communication*, com. 28(1), 1980.

[57] S. MAEDA. « Une analyse statistique sur les positions de la langue : Etude préliminaire sur les voyelles françaises ». *IXèmes Journées d'Etude sur la Parole*, 191–199, Lannion, France, 1978.

[58] S. MAEDA. « Un modèle de la langue avec des composantes linéaires ». *Xèmes Journées d'Etude sur la Parole*, 152–162, Lannion, France, 1979.

[59] S. MAEDA. « Articulatory co-ordination and its neurobiological aspects: an essay ». *ICPhS*, 2, 76–83, Stockholm, 1995.

[60] S. MAEDA R. CARRÉ. « Modèles de production ». *Fondements et perspectives en reconnaissance automatique de la parole*, 51–72, 1995.

[61] A. MARCHAL. « Eléments de phonologie articulatoire ». *Fondements et perspectives en reconnaissance automatique de la parole*, 33–39, 1995.

[62] J. Di MARTINO. « Contribution à la Reconnaissance Globale de la Parole : Mots Isolés et Mots Enchaînés ». *Thèse de 3ème cycle*, 1984.

[63] B. MATHIEU Y. LAPRIE. « Speaker Normalization of the Maeda's Model ». *Proceeding of International Workshop on Speech and Computer, SPECOM'96*, 167–170, St. Petersburg, Russia, 1996.

[64] B. MATHIEU Y. LAPRIE. « Adaptation of Maeda's model for acoustic to articulatory inversion ». *Proceedings of the 5th European Conference on Speech Communication and Technology*, 4, 2015–2018, Rhodes, Greece, September, 1997.

[65] R.S. MCGOWAN. « Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algoritm: Preliminary model tests ». *Speech Communication*, 14:19–48, 1994.

[66] R.S. MCGOWAN M. LEE. « Task dynamic and articulatory recovery of lip and velar approximations under model mismatch conditions ». *J. Acoust. Soc. Am.*, 99:595–608, 1996.

[67] P. MERMELSTEIN. « Articulatory Model for the Study of Speech Production ». *J. Acoust. Soc. Am.*, 53:1070–1082, 1973.

[68] M. MOHOMMAD, E. MOORE, J. N. CARTER, C. H. SHADLE, S. J. GUNN. « Using MRI to image the moving vocal tract during speech ». *Proc. of EuroSpeech*, 2027–2030, 1997.

[69] D. MONTANA L. DAVIS. « Training Feedforward Neural Networks Using Genetic Algorithms ». *Eleventh International Joint Conference on Artificial Intelligence*, 762–767, Palo Alto, USA, 1989.

[70] R. De MORI. « Modèles stochastiques de langage ». *Fondements et perspectives en reconnaissance automatique de la parole*, 109–118, 1995.

[71] M. Mrayati B. Guerin. « Étude des caractéristiques acoustiques des voyelles orales françaises par simulation du conduit vocal avec pertes ». *Revue d'Acoustique*, 36:18–32, 1976.

[72] Y. K. Muthusamy R. A. Cole. « A review of research in automatic language identification ». , Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, 1992.

[73] S. Nakagawa. « A Connected spoken word Recognition Method by $O(n)$ Dynamic Programming Pattern Matching Algorithm ». *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 296–299, Boston, 1983.

[74] H. Ney. « A Dynamic Programmation Algorithm For Nonlinear Smoothing ». *Signal Processing*, 5(2):163–173, 1983.

[75] T. R. Nielser P. C. Woodland. « Variable-length category-based n-grams for language modelling ». , Cambridge Univ. Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, 1995.

[76] Y. Normandin. « Optimal splitting of HMM gaussian mixture components with MMIE training ». *ICCASP-95*, 449–452, 1995.

[77] Y. Normandin, R. Lacouture, R. Cardin. « MMIE Training foe Large Vocabulary Continuous Speech Recognition ». *ICSLP*, 1367–1370, 1994.

[78] J.E. Overall. « Orthogonal Factors and Uncorrelated Factor Scores ». *Psychological Reports*, 10, 651–662, 1962.

[79] Y. Payan P. Perrier. « Vowel normalization by articulatory normalization: First attempts for vowel transitions ». *Proc. of the 3rd European conference on Speech communication and technology*, 1, 417–420, Berlin, 1993.

[80] F. Pellegrino R.-A. Obrecht. « Stratégies en identification automatique des langues. Vers une classification automatique des systèmes vocaliques ». *JEP-96*, 409–412, Avignon, 1996.

[81] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery. *Numerical Recipes in C, The Art of Scientific Computing.* Cambridge University Press, 1992.

[82] L. R. Rabiner S. E. Levinson. « Isolated and Connected Word Recognition

Theory and Selected Applications ». *IEEE Trans. on Communication*, 29(5):621–659, 1981.

[83] L. R. RABINER R. W. SCHAFER. *Digital Processing of Speech.* Prentice-Hall, Englewood Cliffs, N.J., 1978.

[84] L.R. RABINER. « Mathematical Foundations of Hidden Markov Models ». *NATO ASI Series*, F46:183–205, 1988.

[85] H. B. RICHARDS, J. S. BRIDLE, M. J. HUNT, J. S. MASON. « Dynamic constraint weighting in the context of articulatory parameter estimation ». *Proceedings of the 5th European Conference on Speech Communication and Technology*, 2535–2538, Rhodes, Greece, September, 1997.

[86] R.C. ROSE, J. SCHROETER, M. M. SONDHI. « An investigation of potential role of speech production models in automatic speech recognition ». *ICSLP*, 575–578, Yokohama, Japan, 1994.

[87] M. ROSSI. « Niveaux de l'analyse phonétique : nature et structuration des indices et des traits ». *Speech Communication*, 2:91–106, 1983.

[88] D. E. RUMELHART, G. E. HINTON, R. J. WILLIAMS. Learning Internal Representation by Error Propagation. D. E. RUMELHART J. L. MCCLELLAND, , *Parallel Distributed Processing. Exploration of the Microstructure of Cognition*, 1: Foundations, page . MIT Press, 1986.

[89] H. SAKOE S. CHIBA. « A Dynamic Programming Approach to Continuous Speech Recognition ». *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Budapest, Hungary, 1971.

[90] E.L. SALTZMAN K.G. MUNHALL. « A Dynamical Approach to Gestural Petterning in Speech Production ». *Ecological Psychology*, 14:333–382, 1989.

[91] George SAON Abdel BELAÏD. High Performance Unconstrained Word Recognition System Combining HMMs and Markov Random Field. S IMPEDEVO, P. S. WANG, H BUNKE, , *Automatic Bankcheck Processing*, 309–325. World Scientific, 1997.

[92] R.S. SCHECHTER. *The variational Method in Engineering.* McGraw-Hill Book Comp., New York, 1967.

[93] J. SCHOENTGEN S. CIOCEA. « Direct calculation of the vocal tract area function from measured formant frequencies ». *EUROSPEECH*, Madrid, 1995.

[94] J. SCHOENTGEN S. CIOCEA. « Kinematic formant-to-area mapping ». *Speech Communication*, 21:227–244, 1997.

[95] J. SCHROETER, P. MEYER, S. PARTHASARATHY. « Evaluation of improved articulatory codebooks and codebook access distance measure ». *ICASSP*, 1 *1*, 393–396, New Mexico, 1990.

[96] M. M. SONDHI. « Model for wave propagation in a lossy vocal tract ». *J. Acoust. Soc. Amer.*, 55(5):1070–1075, 1974.

[97] M.M. SONDHI J. SCHROETER. « A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer ». *IEEE Trans. on Acoust., Speech and Signal Processing*, 35(7):955–967, 1987.

[98] A. SOQUET. « *Etude comparée de représentations acoustiques et articulatoires du signal de parole pour le décodage acoustico-phonétique* ». PhD thesis, Univ. Libre de Bruxelles, 1995.

[99] M. STONE, M.H. Goldstein JR., Y. ZHANG. « Principal Component analysis of cross sections of tongue shapes in vowel production ». 1st *ESCA Tutorial and Research Workshop on Speech Production Modeling: From Control Strategies to Acoustics & 4th Speech Production Seminar : Models and Data*, 37–40, Autrans, France, 1996.

[100] B. H. STORY, I. R. TIZE, E. A. HOFFMAN. « Vocal tract area functions from magnetic resonance imaging ». *Journal of Acoustical Society of America*, 100(1):537–553, 1996.

[101] H. STRIK. « Using articulatory knowledge in automatic speech recognition ». *Newsletter of the Center for Language Studies (CLS)* [25], 8–13.

[102] W. J. STRONG E. P. PALMER. « Formant Estimation from Linear Prediction Spectra and Their Second Derivatives ». *J. Acoust. Soc. Am.*, 55, page 396, 1974.

[103] F.-R. VERDUN A.G. MARSHALL. *Transformée de Fourier Applications en RMN et IRM.* Masson, 1995.

[104] J.-F. WANG, C.-H. WU, C.-C. HAUNG, J.-Y. LEE. « Integrating Neural Nets and

One-Stage Dynamic Programming for Speaker Independent Continuous Mandarin Digit Recognition ». *ICASSP*, 69–72, 1991.

[105] R.L. WATROUS L. SHASTRI. « Learning Phonetic Features using Connectionnist Networks ». *Proc. of the 10th International Joint Conference on Artificial Intelligence*, 851–854, 1987.

[106] C.-S. YANG H. KASUYA. « Accurate measurement of vocal tract shapes from magnetic resonance images of child, female and male subjects ». *ICSLP*, 623–626, Yokohama, Japan, 1994.

[107] H. YEHIA F. ITAKURA. « A method to combine acoustic and morphological constrints in the production inverse problem ». *Speech Communication*, 18:151–174, 1996.

[108] Z. YU. « A method to determine the area function of speech based on perturbation therory ». *STL-QPSR*, 4:77–94, 1993.

[109] Z.L. YU P.C. CHING. « Geometrycally and Acoustically Optimized Codebook for Unique Mapping from Formants to Vocal-Tract Shape ». *Proc. of EuroSpeech*, 2551–2554, Rhodos, Greece, 1997.