Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Ferhat ABBAS – Sétif
UFAS - Algérie

THESE

présentée à la faculté de technologie

Département
Electronique

pour l'obtention du diplôme de
Doctorat Sciences

par
Mr. Abdelghani HARRAG

Thème

Extraction des données d'une base: Application à l'extraction des traits du locuteur

Soutenue publiquement le 26/06/2011 devant un jury composé de :

Mr.	Ameur Zegadi	Prof. à l'Université de Sétif	Président du jury
Mr.	Tayeb Mohamadi	Prof. à l'Université de Sétif	Rapporteur
Mr.	Moussa Benyoucef	Prof. à l'Université de Batna	Examinateur
Mr.	Rédha Benzid	M.C.A à L'Université de Batna	Examinateur
Mr.	Djamel Saigaa	M.C.A à L'Université de Msila	Examinateur
Mr.	Saad Bouguezel	M.C.A à l'Université de Sétif	Examinateur

A Lina ma fille,

A toute ma famille,

A tous ceux que j'aime.

Remerciements

J'ajoute quelques lignes à ce document pour remercier ceux qui ont participé de près où de loin à sa naissance.

Je pense tout particulièrement au Professeur Ameur Zegadi pour avoir accepté de présider mon jury de thèse. Au Professeur Moussa Benyoucef et aux Docteurs Rédha Benzid, Saad Bouguezel et Djamel Saigaa d'avoir bien voulu disséquer les quelques 150 pages de cette thèse sans m'en tenir trop rigueur.

Merci également à Mahmoud Drif, Ahmed Bouchlaghem, Ahmed Zouaoui, Madame Hikmat et mon frère Fouzi pour leur aide.

Je suis reconnaissant envers le Professeur T. Mohamadi pour m'avoir guidé durant ces années. Je le remercie d'avoir su rester présent malgré ses nombreuses obligations.

J'accorde une place à part dans ces remerciements au Dr. Jean-Francois SERIGNAT qui m'a accueillit à deux reprises au sein de son laboratoire et qui m'a fournit toute l'aide dont j'avais besoin que ça soit au niveau scientifique ou humain, de m'avoir donner la chance d'être accueilli au sein d'un laboratoire chaleureux et festif, je tiens à remercier tous ceux qui ont contribué à cette ambiance.

Et puis merci à ceux qui étaient là dès le commencement. Merci Maman et Papa pour votre soutien et votre confiance. Merci à ma famille de m'avoir accompagné tout au long de mon cursus.

Résumé

Motivé par l'amélioration de la précision de la reconnaissance par la fusion des différentes sources d'information, cette thèse se concentre sur l'exploitation de l'information de la source vocale spécifique au locuteur. Selon la théorie de la production de la parole, la parole est produite par la phonation des cordes vocales suivie par l'articulation du conduit vocal et du rayonnement aux lèvres. Les caractéristiques acoustiques représentant le conduit vocal ont été largement appliquées pour la reconnaissance du locuteur. Bien qu'il ait été révélé que la phonation glottique joue un rôle important dans la caractérisation du locuteur, et que la reconnaissance des personnes familières par les êtres humains en partie repose la-dessus, l'utilité des caractéristiques de la source vocale pour la reconnaissance automatique du locuteur, ainsi que sa technique efficace d'extraction des caractéristiques, n'a pas été pleinement exploitée.

Les paramètres de la source vocale sont généralement jugés moins discriminants et difficiles à extraire, ce qui explique la pré-pondération des paramètres issus du conduit vocal. Néanmoins, l'évolution de la technologie, des ressources de stockage et de calcul, des études dans le domaine de compréhension du phénomène de production et de perception de la parole, poussent les chercheurs à reconsidérer ces préjugés et essayer de tirer le maximum de ces informations complémentaires pour améliorer les performances du système de reconnaissance du locuteur.

L'objectif de ces travaux de recherche est d'améliorer les performances de la reconnaissance du locuteur en fusionnant des informations complémentaires de la source et du conduit. Pour atteindre cet objectif, nous essayons de répondre aux questions suivantes : Comment représenter efficacement les informations spécifiques au locuteur à partir du signal vocal ? Est-il vraiment utile de prendre en compte l'information de la source vocale pour la reconnaissance du locuteur ? Comment faire pour tirer pleinement parti de la fusion de l'information de la source vocale et de celle du conduit vocal pour la reconnaissance du locuteur ?

L'information spécifique au locuteur liée à la phonation glottique est analysée. L'utilité de l'information de la source vocale pour la reconnaissance du locuteur est discutée. En particulier, la complémentarité des informations locuteur, celle issue de la source vocale et celle issue du conduit vocal, est dressée.

Table des matières

Sommaire

Liste des Tableaux

Liste des Figures

Introduction

La reconnaissance du locuteur est une des branches de l'authentification biométrique, qui se réfère à la reconnaissance automatique de l'identité des personnes en utilisant certaines de leurs caractéristiques intrinsèques. Outre la voix, il y a beaucoup d'autres modèles physique et comportementaux pour l'authentification biométrique, par exemple : les yeux, le visage, les empreintes, la signature, …etc. Pratiquement, la sélection d'un modèle biométrique prometteur devrait prendre en compte au moins les considérations suivantes : la robustesse, la précision, l'accessibilité et l'acceptabilité. Par rapport à ces critères de sélection, parmi toutes les technologies d'authentification biométriques, la reconnaissance du locuteur est probablement la plus naturelle et économique pour les systèmes de communication homme-machine et cela en raison de 1) la collecte de données parole est beaucoup plus pratique que les autres motifs, et 2) plus important encore, la parole et le mode dominant d'èchange d'information pour les êtres humains et tend à être le mode dominant pour l'échange d'information pour les système de communication homme-machine.

Le développement de la technologie de traitement de la parole a boosté les nombreuses applications de reconnaissance du locuteur, plus particulièrement dans les domaines suivants:

- Contrôle d'accès aux installations physiques ou des réseaux de données ;
- Achats par téléphone ou d'autres transactions bancaires ;
- Recherche d'information, par exemple, renseignements sur les clients pour les centres d'appels et l'indexation audio ;
- La surveillance à distance ;
- Criminalistique.

La reconnaissance du locuteur peut être divisée en deux tâches: vérification et identification. La tâche de vérification est de décider si oui ou non une voix inconnue appartient au locuteur revendiqué. Il n'y a que deux décisions possibles : soit accepter la voix comme étant celle de la personne clamée ou bien la rejeter si la voix est considérée comme étant celle d'un imposteur. La tâche d'identification consiste à classer la voix inconnue comme appartenant à l'un des locuteurs de la base. Le nombre de décisions alternatives dans l'identification du locuteur est égal à la taille de la population de la base de locuteurs N, généralement la performance du système et inversement proportionnelle à N. Par conséquent, l'identification est généralement une

tâche plus difficile qu'une vérification. Cette identification est connue comme une identification en ensemble fermé (closed-set en anglais). Par contraste, l'identification en ensemble-ouvert (open-set en anglais) intègre la possibilité que la voix inconnue n'appartient à aucun locuteur de la base. Par conséquent, le nombre de décisions alternatives est de N+1, ce qui veut dire que la voix peut appartenir à un locuteur n'appartenant pas à la base de référence. L'identification en ensemble-ouvert est une combinaison entre l'identification et la vérification.

La reconnaissance du locuteur peut aussi être divisée en reconnaissance dépendante du texte et indépendante du texte. En reconnaissance dépendante du texte, le système connait exactement le texte parlé qui peut être soit fixe ou prompté. En reconnaissance indépendante du texte, le système n'a aucune connaissance de l'énoncé, qui pourrait être des mots clés sélectionnés par le locuteur ou une conversation. Dans le cas de la reconnaissance dépendante du texte, le système peut exploiter des caractéristiques du locuteur associées à certains phonèmes ou syllabes. Ainsi, un système de reconnaissance dépendant du texte est plus performant qu'un système de reconnaissance indépendant du texte. Cependant, il requiert une coopération élevée et ne peut être utilisé que pour des applications avec un fort contrôle sur les entrées de l'utilisateur. Le système de reconnaissance indépendante du texte est plus convivial et plus facile à appliquer, mais sans la connaissance de l'énoncé parlé, il est également plus difficile d'atteindre des performances élevées. Dans les applications indépendantes du texte, un système de reconnaissance de la parole qui fournit les informations correctes sur le texte parlé, peut améliorer la précision du système de reconnaissance du locuteur. Bien que la tâche de la reconnaissance indépendante du texte a été acceptée comme une bonne plateforme pour l'évaluation des technologies de la reconnaissance du locuteur (par exemple, les évaluations annuelles NIST), beaucoup d'applications commerciales et industrielles se concentrent d'avantage sur les systèmes dépendants du texte, ou les systèmes de reconnaissance de locuteur texte-contraint.

Réalisations historiques dans la technologie de reconnaissance du locuteur

Les recherches sur la reconnaissance du locuteur ont été entreprises depuis plus de 40 ans, et continues d'être un domaine actif de traitement de la communication parlée. Le développement de la technologie de la reconnaissance du locuteur est étroitement concomitant avec l'avancement dans la connaissance de la parole, le traitement du signal et la technologie des ordinateurs.

La reconnaissance du locuteur par les humains a été largement étudiée dans les années 1960. La motivation de ces études a été d'apprendre comment l'homme reconnait les locuteurs et la fiabilité d'un humain à reconnaitre un locuteur. Le travail le plus important qui a stimulé la recherche sur la reconnaissance du locuteur par la machine a été réalisé par Kersta qui a introduit le spectrogramme (où il l'a noté comme empreinte vocale) en tant que moyen d'identification personnelle.

Dans les années 1970, l'attention a été tournée vers la reconnaissance du locuteur par ordinateur et devient la reconnaissance automatique du locuteur. A cette époque, les systèmes de reconnaissance du locuteur, en général, ne portait que sur une petite population (moins de 20 locuteurs). La transformée de Fourier, les techniques de prédiction linéaire et d'analyse cepstrale ont été appliquées pour générer des paramètres du locuteur. Les moyennes long-terme de ces paramètres ont été utilisées comme références des locuteurs.

Dans les années 1980, des méthodes statistiques de reconnaissance des formes plus compliquées ont été investiguées, par exemple, l'alignement temporel dynamique (DTW) et la quantification vectorielle (VQ), pour des systèmes de reconnaissance du locuteur à grande échelle (>100 locuteurs). La contribution des caractéristiques statiques et dynamiques pour la reconnaissance du locuteur a également été étudiée.

Depuis les années 1990, la mise à disposition de bases de données parole plus importantes (par exemple, corpus YOHO) a boosté les études sur des modèles plus compliqués pour la représentation des locuteurs. Ces modèles comprennent les modèles stochastiques (par exemple, les modèles de Markov cachés (HMM)), le modèle de mélange de Gaussiennes (GMM), les réseaux de neurones (par exemple, Perceptron Multicouches MLP), fonctions à base radiale (RBF)) et les machines à vecteurs de support (SVM)), …etc. Parmi ces techniques de modélisation, la GMM a été reconnue comme la plus efficace à caractériser la distribution de la densité des données de la parole et a été considérée comme la technique de modélisation dominante pour les systèmes de reconnaissance du locuteur. En ce qui concerne l'extraction des caractéristiques, les coefficients cepstraux incorporant le modèle auditif, connus sous le nom de Coefficients Cepstraux à Fréquence Mel (MFCC) et leurs coefficients dynamiques ont été les caractéristiques ou paramètres dominants. En outre, diverses techniques de normalisation des scores ont également été étudiées pour la reconnaissance du locuteur robuste. Un système avec les paramètres MFCC, une modélisation GMM et le modèle de base universel (UBM) pour la normalisation des scores, a été considéré comme celui qui obtient les meilleurs résultats et est accepté comme un système de base pour comparer les nouvelles technologies.

Challenge de l'état de l'art de la reconnaissance du locuteur

L'état de l'art des systèmes de reconnaissance du locuteur fonctionnent très bien dans des expérimentations de laboratoire ou sous certaines applications spécifiques avec des conditions d'apprentissage et de fonctionnement sophistiquées. Les résultats expérimentaux montrent que la reconnaissance automatique du locuteur dans un environnement idéal atteint un niveau de performance aussi bon, voire meilleur, que celui de la reconnaissance par les êtres humains. Toutefois, comme une technique orientée vers les applications, les performances des systèmes actuels de reconnaissance du locuteur dans les applications du monde réel sont loin d'être robustes et fiables en comparaison avec les systèmes de reconnaissance par l'homme. Le principal défi à la technologie de reconnaissance du locuteur a donc été d'améliorer la robustesse des systèmes dans des conditions incompatibles. Pour la reconnaissance du locuteur, les

disparités sont principalement causées par 1) la variation intra-locuteur du style de parler, et 2) les variations de l'environnement acoustique.

Notre système vocal fournit principalement les indices acoustiques pour la classification des phonèmes, et aussi la personnalité individuelle pour caractériser le locuteur. La variation interlocuteur pourrait être importante, même pour le même contenu parole. Un système de reconnaissance du locuteur essaye de comprendre ces variations interlocuteur sur lesquelles est basée la discrimination d'un locuteur par rapport aux autres. Dans le même temps, le système vocal produit un certain degré de variation intra-locuteur pour le même contenu parole réitéré à différents moments. La plupart des erreurs de reconnaissance sont causées par ces types de variations intra-locuteur.

D'autre part, la variation de l'environnement acoustique est causée par les diverses distorsions imprévisibles lors de la collecte des données et la transmission. Par exemple, dans les applications de reconnaissance du locuteur par téléphonie (par exemple, des transactions bancaires par téléphone), les données vocales pourraient être recueillies dans des environnements avec un bruit de fond différent, avec différents téléphones, et via différents canaux. Le bruit de fond et la combinaison combiné/distorsion du canal change la structure spectrale des données paroles et les paramètres acoustiques dérivés (par exemple, les paramètres MFCC) ne peuvent pas représenter les informations du locuteur correctement.

Motivation et objectif de cette thèse

Motivé par l'amélioration de la précision de la reconnaissance par la fusion des différentes sources d'information, cette thèse se concentre sur l'exploitation de l'information de la source vocale spécifique au locuteur pour la reconnaissance du locuteur. Selon la théorie de la production de la parole, la parole est produite par la phonation des cordes vocales suivie par l'articulation du conduit vocal et du rayonnement aux lèvres. Les caractéristiques acoustiques représentant le conduit vocal, (par exemple, MFCC ou LPCC) ont été largement appliquées pour la reconnaissance du locuteur. Bien qu'il ait été révélé que la phonation glottique joue un rôle important dans la caractérisation du locuteur, et que la reconnaissance des personnes familières par les êtres humains en partie repose la-dessus, l'utilité des caractéristiques de la source vocale pour la reconnaissance automatique du locuteur, ainsi que sa technique efficace d'extraction des caractéristiques, n'a pas été pleinement exploitée.

Beaucoup d'efforts ont été consacrés à l'étude des paramètres du conduit vocal ou ceux de la source vocale avec une pré-pondération pour les premiers. Les paramètres de la source vocale sont généralement jugés moins discriminants et difficiles à extraire, ce qui explique la pré-pondération des paramètres issus du conduit vocal. Néanmoins, l'évolution de la technologie, des ressources de stockage et de calcul, des études dans le domaine de compréhension du phénomène de production et de perception de la parole, poussent les chercheurs à reconsidérer ces préjugés et essayer de tirer le maximum de ces informations complémentaires pour améliorer les performances du système de reconnaissance du locuteur.

L'objectif de ces travaux de recherche est d'améliorer les performances de la reconnaissance du locuteur en fusionnant des informations complémentaires de la source et du conduit. Pour atteindre cet objectif, nous essayons de répondre aux questions suivantes :

- Comment représenter efficacement les informations spécifiques au locuteur à partir du signal vocal ?
- Est-il vraiment utile de prendre en compte l'information de la source vocale pour la reconnaissance du locuteur ?
- Comment faire pour tirer pleinement parti de la fusion de l'information de la source vocale et de celle du conduit vocal pour la reconnaissance du locuteur ?

comment le faire, en trouvant des réponses à ces questions, telle est la question ?

Cette thèse présente tout d'abord le processus et une introduction au domaine de la biométrie suivie par une bonne entrée en matière de la reconnaissance automatique du locuteur. La théorie acoustique de la production de la parole ainsi que son analyse numérique, les notions fondamentales, les spécificités du signal parole, les modèles associés …etc y sont introduits. L'information spécifique au locuteur liée à la phonation glottique est analysée. L'utilité de l'information de la source vocale pour la reconnaissance du locuteur est discutée. En particulier, la complémentarité des informations locuteur, celle issue de la source vocale et celle issue du conduit vocal, est dressée.

Structure du document

Dans l'introduction, la problématique de cette thèse, le cadre applicatif, le cadre théorique et les objectifs sont expliqués et fixés d'emblé. On y trouve aussi les différentes contributions de ces travaux de thèse. Une structure du document et une brève description pour chaque partie sert à faciliter le suivi du fil conducteur entre les deux parties de la thèse.

Le reste du document est déployé en 2 grandes parties en étudiant successivement:

Partie 1 : Dans cette partie, le cadre applicatif, le cadre théorique de la biométrie vocale (ou la Reconnaissance Automatique du Locuteur RAL) ainsi que le cadre théorique pour la production et l'analyse du signal parole sont introduits (Chapitres 1, 2 et 3).

- Le chapitre 1 est une bonne entrée en matière dans le domaine de la biométrie sous toutes ses formes tout en donnant les critères de choix d'une technologie biométrique parmi d'autres, telle que: l'accessibilité, l'acceptabilité, le coût, …etc. Ces critères poussent à considérer la biométrie vocale comme l'une des technologies biométriques la plus naturelle: elle n'est pas intrusive, n'exige aucun contact physique avec le système, disponible en plus d'être le moyen de communication le plus utilisé.

- Le chapitre 2 détaille la biométrie vocale vs la Reconnaissance Automatique du Locuteur (RAL). Il est consacré aux principes de la reconnaissance automatique du locuteur. Il souligne également les difficultés majeures associées à la RAL et expose les différentes tâches liées à la RAL. Nous exposons quelques approches utilisées pour les systèmes de RAL et présentons, enfin, les méthodes d'évaluation des performances des systèmes de RAL.
- Le chapitre 3 présente tout d'abord les mécanismes de production de la parole. Nous exposons ensuite les traitements numériques appliqués au signal audio dans un système de reconnaissance du locuteur. Différentes représentations sont données pour étoffer cette partie qui sera abordée plus en détail dans le chapitre suivant.

Partie 2 *:* Considérée comme le cœur de ces travaux de thèse, elle traite l'état de l'art de la caractérisation du locuteur et des méthodes de sélection, d'extraction ou de fusion des paramètres mises en place et testées dans la partie expérimentale (Chapitres 4, 5, 6 et 7).

- Le chapitre 4 donne un état de l'art exhaustif de la caractérisation du locuteur en soulignant les bases, l'évolution de cette caractérisation et les techniques de représentation utilisées au fil des décennies tout en mettant l'accent sur les développements récents et sur les tendances actuelles. Large éventail de travaux de recherche dans ce domaine sont cités qui couvrent l'ensemble de la littérature, ou presque, et donnent une bonne idée des différents niveaux d'information contenu dans le signal parole et la difficulté à les extraire et surtout la complexité à les utiliser dans le domaine de la RAL.
- Le chapitre 5 donne un aperçu sur les méthodes de sélection ou d'extraction des paramètres. Les méthodes de fusion de données et de réduction de la dimensionnalité, telle que l'Analyse Discriminante Linéaire (ADL) ou l'Analyse en Composantes Principales (ACP), sont introduites avec plus de détail pour la méthode ADL qui a été utilisée avec succès dans ces travaux de thèse. Cette fusion au niveau paramètres et celle au niveau classificateur sont actuellement le sujet de beaucoup de travaux de recherche qui essayent de combiner des informations complémentaires pour améliorer la performance des systèmes RAL.
- Les premiers résultats sur la base BDSONS y sont reportés dans le chapitre 6. Ces résultats inclus l'étude des paramètres prosodiques et les paramètres acoustiques ainsi que de leur fusion utilisant la méthode ACP ou la méthode ADL. Ces résultats ont fait l'objet de deux communications internationales (Annexe A). Le chapitre souligne aussi les différents problèmes rencontrés dans l'utilisation de la base BDSONS et qui sont dus principalement à sa qualité générique (non adaptée pour des recherches sur la reconnaissance du locuteur), à sa pauvreté au niveau des enregistrements et surtout au niveau des répétitions (pour des travaux comparatifs ou des travaux sur la durée), et par son nombre limité de locuteurs.
- Le chapitre 7 essaye de répondre aux problèmes de base de données cités dans le chapitre précédent. Il est consacré à la collecte et la mise en place de la nouvelle base de données nommée QSDAS (QSDAS pour Quranic Speech Database for

Arabic Speaker recognition). Le reste de ce chapitre résume les différents tests effectués sur cette nouvelle base et qui englobent les tests sur la meilleure représentation acoustique parmi plusieurs représentations utilisées, à savoir : LPC, LPCC, PARCOR et MFCC.

Des tests complémentaires sur les paramètres prosodiques (Pitch, Durée et énergie) ont été effectués. Une étude comparative de trois méthodes de sélection des paramètres (ACP, SFS et ADL) a été réalisée et donne la méthode ADL comme une méthode viable pour la fusion des données, l'extraction de l'information pertinente et la réduction de la dimension, réduction qui s'avère cruciale dans les applications à faibles ressources. Ces travaux ont fait l'objet de deux publications internationales (Annexe A).

Nous concluons cette thèse par une récapitulation des principales conclusions et discussions de ces travaux de recherches tout en donnant les perspectives des futurs travaux qui auront pour but d'utiliser d'autres méthodes de sélection telle que les algorithmes évolutionnaires, d'autres types de paramètres tels que les paramètres haut niveau ainsi que des travaux sur la fusion des scores de différents classificateurs.

Chapitre

1 Introduction à la biométrie

Sommaire

Résumé

Ce chapitre propose une introduction à la biométrie. Il introduit la notion d'identité et les questions inhérentes à la reconnaissance d'un individu. Il présente ensuite les différentes technologies biométriques, leurs applications et les tendances actuelles du marché mondial. Les critères de choix d'une biométrie sont donnés.

1.1 Le contexte

La croissance internationale des communications, tant en volume qu'en diversité (déplacement physique, transaction financière, accès aux services...), implique le besoin de s'assurer de l'identité des individus. L'importance des enjeux, motive les fraudeurs à mettre en échec les systèmes de sécurité existants. En 2004, cette fraude a couté à un pays comme la France, où le commerce électronique est loin d'être à la mode, 241,6 millions d'euros avec 210.000 identités usurpées. Les taux de fraude sur les transactions internationales sont beaucoup plus élevés que celui de la France.

RSA Security a publié en 2005 une enquête sur les problèmes rencontrés par l'employé dans la gestion de ses mots de passe ainsi que des risques potentiels pour la sécurité de l'entreprise. L'étude réalisée aux Etats-Unis dans 1700 entreprises technologiques montre que plus du quart des personnes interrogées doivent gérer plus de 13 mots de passe. Neuf personnes sur dix s'estiment agacées par la gestion de cette quantité de mots de passe. Et c'est cette frustration qui peut donner naissance à des comportements dangereux pour la sécurité. 60 % des employés interrogés gèrent plus de 6 mots de passe différents. La plupart d'entre eux (88 %) parle de frustration liée à la quantité de codes à retenir.

Les entreprises, pour respecter les normes de sécurité, ont été obligées de renforcer l'usage des mots de passe, devenant bientôt un fardeau pour l'utilisateur final. Le mot de passe doit être modifié de plus en plus fréquemment et les caractéristiques auxquelles il doit répondre sont de plus en plus complexes - alliant majuscule, minuscule, chiffre, ponctuation...

Une autre étude réalisée par une université anglaise, montre que 91% des mots de passe utilisés par des internautes sont connus c'est-à-dire issus de l'environnement familier de la personne et jugés non viables par des spécialistes du cryptage.

- 21 % utilisent leur prénom ou celui d'un membre de la famille
- 15 % leur date de naissance ou d'anniversaire
- 15 % les noms de leurs animaux
- 14 % le prénom d'un membre de leur famille
- 7 % ont un lien avec une date clé
- 2 % utilisent "password"
- 30 % des personnes partagent leur mot de passe avec leur partenaire
- 50 % seulement affirment être les seuls à connaître leur mot de passe

Il y a donc un intérêt grandissant pour les systèmes d'identification et d'authentification. Leur dénominateur commun, est le besoin d'un moyen simple, pratique, fiable, pour vérifier l'identité d'une personne, sans l'assistance d'une autre personne.

1.2 Pourquoi la biométrie ?

Le niveau de sécurité d'un système est toujours celui du maillon le plus faible. Ce maillon faible, c'est bien souvent l'être humain : mot de passe aisément déchiffrable ou noté à coté de l'ordinateur. Dans la plupart des entreprises, on exige que les mots de passe soient modifiés régulièrement et comportent au moins 8 caractères, mélangeant lettres majuscules, minuscules et chiffres. L'objectif est d'échapper aux logiciels de décodage qui peuvent en peu de temps, balayer tous les mots du dictionnaire. Une protection qui peut s'avérer insuffisante pour l'accès à des applications sensibles.

Le défaut commun à tous les systèmes d'authentification est que l'on identifie un objet (ordinateur, carte, code...) et non la personne elle-même. Il est pourtant plus acceptable d'authentifier une personne, plutôt qu'une machine.

D'une manière générale, les technologies de l'identification humaine automatique sont essentiellement caractérisées (Liu et Silverman, 2001) par l'élément qui permet l'identification. Celui-ci peut être :

- Quelque chose de détenu par l'utilisateur (clef, carte, etc.),
- Quelque chose de connu par l'utilisateur (mot de passe, code, etc.),
- Quelque chose de spécifique à l'utilisateur (ADN, empreinte, voix, etc.). Il s'agit de la biométrie.
- Quelque chose spécifique au comportement de l'utilisateur

Les 2 premiers moyens d'identification peuvent être utilisés pour usurper l'identité d'un tiers. La relation entre le niveau de sécurité de tout système et ses trois principes est illustrée par la figure 1.1.

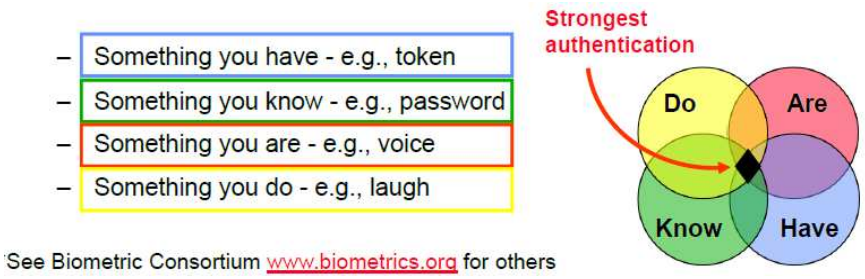


Figure 1.1 Relation entre le niveau de sécurité et le moyen utilisé.

1.3 Qu'est ce que la Biométrie ?

La biométrie est la science issue de la biologie et des mathématiques qui implique une analyse de caractéristiques physiologiques des êtres vivants (dimensions, croissance, …). C'est l'étude de caractéristiques humaines pour la reconnaissance ou la vérification de l'identité d'un individu. De la même manière, on peut définir un système d'authentification biométrique comme étant un système automatique de vérification d'identité basé sur une caractéristique physique propre à l'individu ou une caractéristique décrivant son comportement.

1.3.1 L'identité d'un individu

L'identité est une notion complexe, difficile à définir. Cette première section propose une définition des enjeux et les limitations relatifs au concept d'identité, dans le cadre des applications biométriques.

L'identité renvoie à ce qu'un sujet a d'unique. D'un point de vue personnel, la caractérisation de l'identité prend en compte tout ce que l'individu considère comme faisant partie intégrante de lui et qui ne peut lui être enlevé. Cette définition inclut un certain nombre de facteurs qui peuvent évoluer dans le temps comme, d'ailleurs, la conscience de soi.

D'un point de vue externe à l'individu, son identité est la façon dont il est perçu par le monde qui l'entoure. Cette identité, en tant qu'entité, est associée à une appellation. L'individu se nomme « moi », son environnement lui associe un nom. La reconnaissance d'un individu se heurte à la caractérisation de son unicité. Il n'est pas imaginable d'obtenir une description exhaustive d'un individu qui engloberait sa description physiologique complète ainsi que la description de ses connaissances, de ses possessions, de son vécu et de son expérience. Il faut alors, pour obtenir une description unique d'un individu, la restreindre aux informations nécessaires et suffisantes à sa reconnaissance au sein d'un groupe. Dans le cadre de la reconnaissance d'une identité par un système automatique, cette description ne doit intégrer que des informations susceptibles d'être vérifiées dans le contexte applicatif choisi.

Les informations décrivant un individu peuvent être de nature variable. Il est commun de décrire une personne par ses caractéristiques physiques, comme la couleur de ses cheveux, de ses yeux ou d'autres détails de son anatomie. Ce type de description nécessite cependant d'avoir déjà vu cet individu. Lors d'une conversation téléphonique, il est naturel de reconnaître son interlocuteur à sa voix ou la façon dont il s'exprime.

Deux principaux types d'information peuvent être utiles pour décrire une personne: les informations liées à ses possessions et celles qui décrivent sa nature même. Pour les premières, il peut s'agir d'une possession matérielle comme une clef ou un passeport, mais également d'une possession intellectuelle, comme un code, un mot de passe ou, plus généralement, un souvenir. Ces informations présentent l'intérêt d'être facilement vérifiables, mais peuvent être perdues, oubliées ou usurpées.

Les informations obtenues par la mesure des caractéristiques d'une personne, ou données biométriques, font référence aux caractéristiques intrinsèques de l'individu. Leur utilisation nécessite la prise en compte de la nature changeante de l'être humain. Ces changements peuvent être dus au vieillissement, à la maladie ou à un état émotionnel différent et doivent être pris en compte dans la description biométrique d'un individu. Pour la reconnaissance des individus, les systèmes biométriques permettent d'atteindre des niveaux de performance qui sont inaccessibles aux êtres humains.

1.4 Histoire de la biométrie

"Perhaps the most beautiful and characteristics of all superficial marks are the small furrows with the intervening ridges and their pores that are disposed in a singularly complex yet even order on the under surfaces of the hand and the feet."[1].

Cette affirmation concernant les empreintes digitales, plutôt banale à notre époque, marquait le premier pas vers l'élaboration d'un système universel d'identification des criminels au service des policiers du monde entier. Sir Francis Galton (1822-1911) n'était pas le premier à remarquer les sillons et les creux existant à l'intérieur de nos mains et sous nos pieds, ni même à leur trouver d'utiles applications. Certains auteurs mentionnent que le procédé de reconnaissance anthropométrique, bien que les empreintes de mains laissées par nos ancêtres sur les parois des cavernes, ne nous aient pas livré tous leurs secrets, l'empreinte du pouce servait déjà de signature lors d'échanges commerciaux à Babylone (-3000 av. JC) et dans la Chine antique (7ème siècle) pour la signature de documents.

Les caractéristiques de ces empreintes attirèrent aussi l'attention de l'anatomiste Marcello Malpighi (1628-1694) qui les étudia alors avec un nouvel instrument nommé microscope. Puis le Physiologiste tchèque Jan Evangelista Purkinge (1787-1869) s'affaira à catégoriser les empreintes selon certaines caractéristiques. Une application pratique de prise d'empreinte fut réalisée par Sir William Herschel (1738-1822), fonctionnaire britannique au Bengale, excédé par le peu d'empressement des marchands locaux à respecter les contrats qu'il concluait avec eux. Il exigea alors de ceux-ci l'apposition de leurs empreintes digitales sur les documents contractuels. Puis le Dr Henry Faulds (1843-1930), chirurgien à Tokyo, donna une sérieuse impulsion au développement d'un système de classification par la prise d'empreintes. En octobre 1880, il écrivit dans la revue Nature : "*When bloody finger-marks or impression on clay, glass, etc., exist, they may lead to the scientific identification of criminals*"[2]. A cette époque, le Dr Faulds écrivit au naturaliste Charles Darwin (1809-1882) pour l'informer de ses découvertes sur les empreintes digitales. Darwin, vieux et malade, déclina l'offre et transmit son courrier à son cousin Sir Francis Galton. Galton était à la fois physiologiste, anthropologue et psychologue. Il s'affaira notamment à appliquer la méthode statistique à l'étude de l'hérédité et des différences individuelles[3]. Sa contribution fut de démontrer que les empreintes digitales ne changent pas de façon notoire avec le vieillissement des personnes[4].

Au moment où Galton travaillait sur les empreintes, un de ses contemporains, le Français Alphonse Bertillon (1853-1914), testait à la préfecture de police de Paris une méthode d'identification des prisonniers nommé anthropométrie judiciaire ou

---

[1] Personal identification and description, Nature, Sir Francis Galton, 28 juin 1888.

[2] BBC-History-Science and discovery-By people- Henry Faulds, British Broadcasting Corporation.

[3] Petit Robert 2, dictionnaire universel des noms propres, Les dictionnaires Robert – Canada SCC, Montréal, Canada, 1990.

[4] "Galton's formulation gives the probability that particular fingerprint configuration in an average size fingerprint (containing 24 regions as defined by Galton) will be observed in nature." Cette probabilité est de $(1/16 \times 1/256 \times (1/2)^R$ ou 1.45×10^{-11}, On the individuality of fingerprints, Sarah Pankanti, IBM T.J.

bertillonnage[5]. Bertillon procédait à la prise de la photographie de sujets humains, mesurait certaines parties de leur corps (tête, membres, etc.) et en notait les dimensions sur les photos et sur des fiches à des fins d'identification ultérieure.

La dactyloscopie (procédé d'identification par les empreintes digitales) et le bertillonnage furent des techniques rapidement adoptées par les corps de polices du monde entier. Un policier argentin fut le premier à identifier un criminel par ses empreintes en 1892. Par la suite, la dactyloscopie s'imposa comme technique anthropométrique et le bertillonnage s'efface graduellement.

"IAFIS became operational on July 28, 1999, and provides the FBI with a totally electronic environment in which to process fingerprint submissions 24/7/365. Today over 42.8 million digitized criminal fingerprint records reside in the IAFIS database, which is far the world's largest biometric repository of any kinds. It is at least four times larger than all of fingerprint repositories in Europe Combined"[6].

L'utilisation de l'empreinte digitale à des fins d'identification des criminels par les policiers est prédominante, mais parallèlement elle est de plus en plus utilisée par des entreprises et des gouvernements à des fins de sécurité, d'identification et d'authentification. Cette technique est aussi en compétition avec plusieurs autres qui s'imposent de plus en plus sur le marché.

1.5 Marché de la biométrie

La biométrie connaît un engouement sans précédent. La croissance mondiale de la biométrie depuis quelques années est incontestable, tant le nombre d'intervenants est grand, même s'il existe peu d'informations publiques concernant ce marché. On peut toutefois considérer certaines données et certains chiffres sur son évolution au fil des années, tant à l'échelle mondiale, qu'américaine, européenne ou française.

Le marché de la sécurité informatique est encore atomisé, peu de fournisseurs peuvent prétendre offrir une gamme complète de produits. Les spécialistes estiment que ce marché est en pleine croissance et qu'il va également se concentrer. Internet et le commerce électronique sont des marchés porteurs pour la sécurité, mais ils ne sont pas les seuls. Le télétravail, la mise à dispositions d'informations aux clients et sous traitants sont également des facteurs de risque pour les entreprises qui ouvrent leur système d'informations.

Le besoin grandissant de sécurité sur les terminaux mobiles a été mis en exergue par une enquête récente, publiée par Toshiba. Celle-ci soutient que 90% des cadres dirigeants et chefs d'entreprise européens stockent des données sensibles, voire confidentielles sur leur outil de communication et parmi eux, 22% admettent avoir pourtant déjà perdu cet outil.

---

[5] La criminologie et la criminalistique, VIIe colloque de l'Association internationale des criminologues de langue française, 15 mai 2001, SUN Fu Ph.D., Alphonse Bertillon, "Le père de l'anthropométrie".

[6] Hearing How New Technologies (Biometrics) Can be Used To Prevent Terrorism, Michael D. Kirkpatrick, FBI, devant le United States Senate Committee on The Judiciary Subcommittee on Technology, Terrorism, and Government Information, Washington, 14 novembre 2001.

Le groupe IBG (International Biometric Group[7]) édite régulièrement un rapport sur le marché de la biométrie. Cette étude est une analyse complète des chiffres d'affaires, des tendances de croissance, et des développements industriels pour le marché de la biométrie actuel et futur. La lecture de ce rapport est essentielle pour des établissements déployant la technologie biométrique, les investisseurs dans les entreprises biométriques, ou les développeurs de solutions biométriques.

On s'attend à ce que le chiffre d'affaires de l'industrie biométrique incluant les applications judiciaires et celles du secteur public, se développe rapidement. Une grande partie de la croissance sera attribuable au contrôle d'accès aux systèmes d'information (ordinateur/réseau) et au commerce électronique, bien que les applications du secteur public continuent à être une partie essentielle de l'industrie.

On prévoit que le chiffre d'affaires des marchés émergents (accès aux systèmes d'information, commerce électronique et téléphonie, accès physique, et surveillance) dépasse le chiffre d'affaires des secteurs plus matures (identification criminelle et identification des citoyens). On s'attend à ce que l'Asie et l'Amérique du Nord soient les plus grands marchés globaux pour les produits biométriques et les services. La figure 1.2 donne les prévisions d'évolution du marché mondial de la biométrie sur la période 2007-2015.

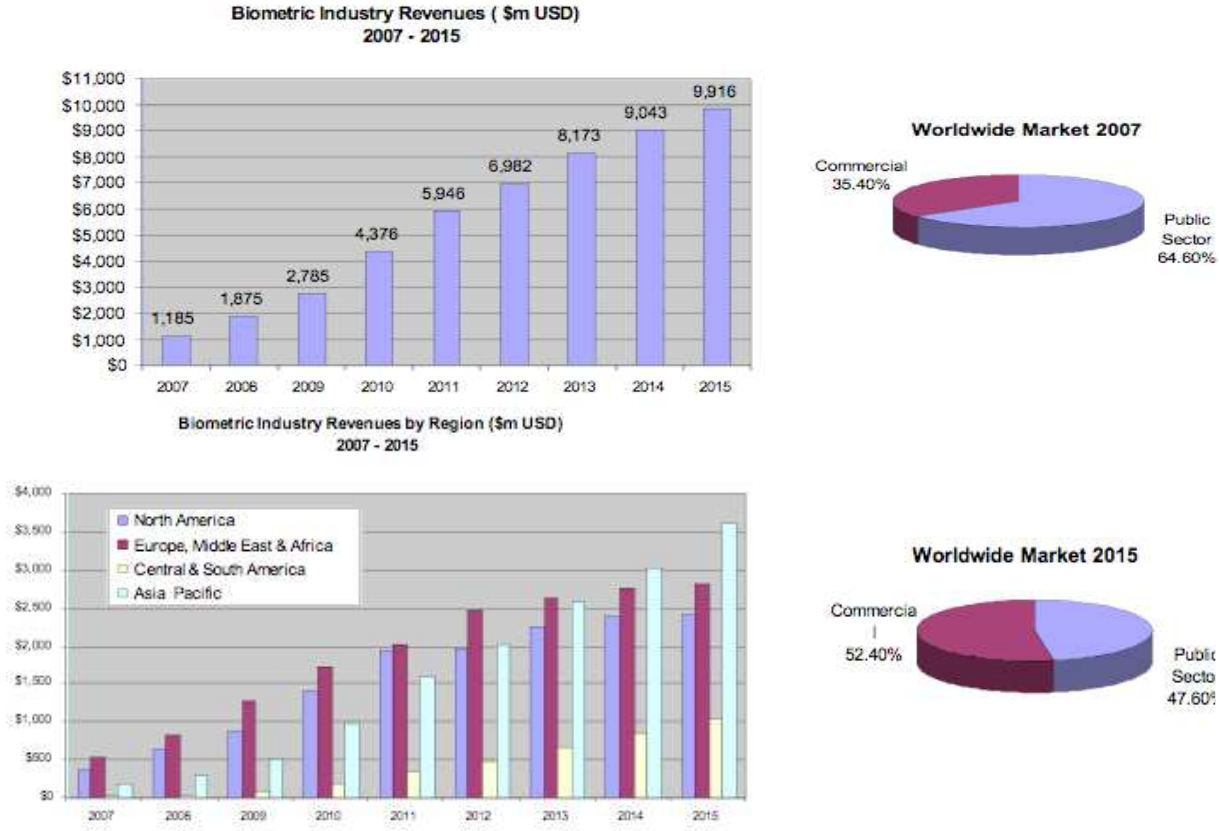


Figure 1.2 Prévisions d'évolution du marché mondial de la biométrie entre 2007-2015.

[7] http://www.biometricgroup.com

Selon le cabinet IBG, les revenus annuels de la technologie de l'empreinte digitale représenteraient 467 millions de dollars en 2002, soit la plus grande part de marché parmi toutes les technologies. Cette croissance est attribuable au grand choix d'applications fonctionnant avec des solutions basées sur l'empreinte digitale. La figure 1.3 donne les parts de marché par technologie biométrique.

Les empreintes digitales continuent à être la principale technologie biométrique en terme de part de marché, près de 50% du chiffre d'affaires total (hors applications judiciaires). La reconnaissance du visage, avec 12% du marché (hors applications judiciaires), dépasse la reconnaissance de la main, qui avait avant la deuxième place en terme de source de revenus après les empreintes digitales.

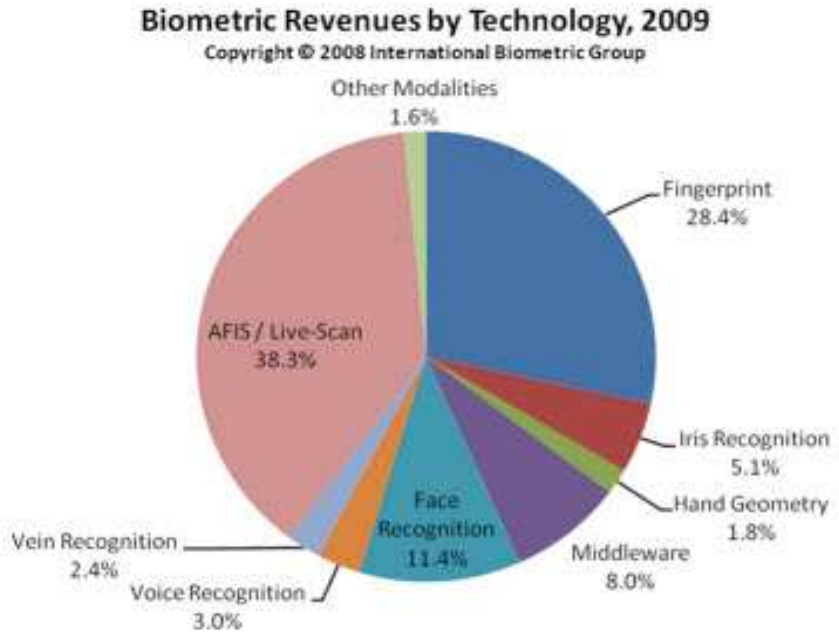


Figure 1.3 Parts du marché par technologie biométrique.

1.6 Les applications de la biométrie

Le champ d'application de la biométrie couvre potentiellement tous les domaines de la sécurité où il est nécessaire de connaître l'identité des personnes. Aujourd'hui, les principales applications sont la production de titres d'identité, le contrôle d'accès à des sites sensibles, le contrôle des frontières, l'accès aux réseaux, systèmes d'information, stations de travail et PC, le paiement électronique, la signature électronique et même le chiffrement de données. Cette liste n'est pas exhaustive, et de nouvelles applications vont très certainement voir rapidement le jour.

La liste des applications pouvant utiliser la biométrie pour contrôler un accès (physique ou logique), peut être très longue. La taille de cette liste n'est limitée que par l'imagination de chacun :

1. Contrôle d'accès physiques aux locaux
   - Salle informatique
   - Site sensible (service de recherche, site nucléaire)
2. Contrôle d'accès logiques aux systèmes d'informations
   - Lancement du système d'exploitation
   - Accès au réseau informatique
   - Commerce électronique, paiement en ligne
   - Transaction (financière pour les banques, données entre entreprises)
   - Signature de document (lot de fabrication de médicaments)
   - Tous les logiciels utilisant un mot de passe
3. Equipements de communication
   - Terminaux d'accès à Internet
   - Téléphones portables
4. Machines & Equipements divers
   - Coffre fort avec serrure électronique
   - Distributeur automatique de billets
   - Casier sensible (club de tir, police)
   - Cantine d'entreprise, cantine scolaire (pour éviter l'utilisation d'un badge par une personne extérieure et améliorer la gestion)
   - Casier de piscine (plus d'objet à porter sur soi)
   - Contrôle des adhérents dans un club, carte de fidélité
   - Contrôle des temps de présence
   - Voiture (anti-démarrage)
5. Etat / Administration
   - Fichier judiciaire
   - Le Fichier national automatisé des empreintes digitales
   - Le fichier national automatisé des empreintes génétiques
   - Titres d'identité (carte d'identité, passeport, permis, titre de séjour)
   - Services sociaux (sécurisation des règlements)
   - Services municipaux
   - Système de vote électronique

1.7 Caractéristiques de la biométrie

La biométrie ou mesure (metron) du vivant (bios) est, d'après l'encyclopédie Larousse[8], « l'étude statistique des dimensions et de la croissance des êtres vivants ». L'extension de la biométrie au domaine de la reconnaissance des personnes consiste à déterminer l'identité d'un individu grâce à des mesures quantitatives. Ces mesures peuvent avoir pour objet les caractéristiques morphologiques ou les caractéristiques comportementales de cette personne.

[8] http://www.larousse.fr/encyclopedie/

3 catégories de technologies biométriques :

- ***Analyses biologiques:*** Odeur, sang, salive, urine, ADN…
- ***Analyses morphologiques:*** empreintes digitales, forme de la main, traits du visage, dessin du réseau veineux, oeil (l'iris ou la rétine), géométrie de l'oreille, dessin de lèvres, forme de pores de la peau….
- ***Analyses comportementales:*** La dynamique de la signature (la vitesse de déplacement du stylo, les accélérations, la pression exercée, l'inclinaison), la façon d'utiliser un clavier d'ordinateur (la pression exercée, la vitesse de frappe), la voix, la manière de marcher (démarche)...

1.7.1 La biométrie morphologique

La biométrie morphologique décrit les individus par des mesures de leurs caractéristiques biologiques ou physiologiques. Ces mesures sont moins sujettes à l'influence du stress que la biométrie comportementale. Elles sont également plus difficiles à falsifier.

Les caractéristiques mesurables qui permettent de décrire un individu sont nombreuses (Jain et al., 1999). Chaque modalité présente des avantages et inconvénients qu'il faut considérer en parallèle de ses performances et, donc, du degré de sécurité qu'elle propose. Les biométries morphologiques les plus courantes mesurent les empreintes digitales, le réseau veineux de la rétine, l'iris, l'empreinte de la main ou certaines caractéristiques du visage. La biologie permet, quant à elle, de caractériser un individu par son ADN à travers une analyse de sa salive, de son sang ou de tout échantillon corporel.

La biométrie morphologique est, à l'heure actuelle, un des moyens les plus fiables pour reconnaître un individu, car elle mesure des caractéristiques qui sont indissociables de cet individu. Elle présente néanmoins certains inconvénients. Elle doit, par exemple, pour être utilisable, intégrer les changements temporels intrinsèques de l'individu. L'acquisition de certaines données biométriques peut également être compliquée par des difficultés physiques ou sociétales.

1.7.2 La biométrie comportementale

La biométrie comportementale mesure et caractérise des éléments qui sont propres aux comportements d'un individu. De nombreux comportements peuvent être observés et analysés afin de caractériser une personne.

La signature dynamique constitue un exemple de biométrie comportementale. Elle consiste à mesurer certaines variables qui interviennent lorsqu'un individu signe un document. Les systèmes de biométrie utilisant cette méthode enregistrent la vitesse et les accélérations du stylo ou la pression exercée. Ils permettent aussi d'analyser, de façon plus naturelle, la forme de la signature. Il est alors possible de différencier les parties qui sont identiques à chaque réalisation de la signature de celles qui varient.

Cette biométrie présente l'avantage d'être historiquement une méthode d'identification très utilisée et adaptée à l'authentification de documents manuscrits.

L'utilisation de matériels informatiques a également suscité un intérêt pour la biométrie. Par exemple, les travaux de Monrose et Rubin (Monrose et Rubin, 2000) ont montré qu'il est possible de reconnaître une personne au rythme de sa frappe sur un clavier. Cette méthode présente l'avantage de permettre une identification continue de l'utilisateur et de détecter un changement d'utilisateur en temps réel et de façon transparente.

L'analyse de la démarche (Cunado et al., 1997) ou celle du contact du pied sur le sol (Orr et Abowd, 2000 ; Rodriguez et al., 2007 ; Rodríguez et al., 2008) permettent également d'authentifier un individu.

Les principaux inconvénients des biométries comportementales sont liés à la grande variabilité que peuvent générer des changements émotionnels ou environnementaux chez l'utilisateur. Le stress ou un environnement perturbé peuvent, par exemple, affecter les comportements et ainsi perturber le résultat du test de reconnaissance.

1.7.3 Les biométries mixtes

Certaines modalités se situent à la croisée des biométries morphologiques et comportementales. La voix, qui est utilisée de façon naturelle par les êtres humains pour reconnaître un individu, est une modalité comportementale qui peut subir les influences d'une pathologie, du stress ou même d'un changement émotionnel. Elle peut également être modifiée selon la volonté du locuteur. Elle garde cependant des caractéristiques constantes qui peuvent permettent d'identifier le locuteur dans le cas où il contrefait sa voix. En effet, le phénomène complexe de la production vocale fait intervenir un grand nombre de caractéristiques intrinsèques au locuteur, qui seront abordées plus précisément dans la partie 3.1. La morphologie de l'appareil respiratoire du locuteur influe, par exemple, sur les caractéristiques de sa voix. Or, cette morphologie ne peut être modifiée de façon volontaire par l'individu.

L'analyse des battements du coeur par l'intermédiaire des signaux d'un électrocardiogramme (Israel et al., 2005), ou l'analyse de l'activité électrique du cerveau mesurée par Électro-encéphalographie (Marcel et del R. Millan., 2007) sont d'autres modalités biométriques mixtes. Les signaux acquis dans ces deux modalités sont sujets aux changements émotionnels, physiologiques et environnementaux, mais contiennent cependant des informations caractérisant respectivement le muscle cardiaque et le cerveau qui sont propres à l'individu considéré.

La biométrie vidéo exploite également les informations morphologiques et comportementales des individus. Elle permet de décrire les traits du visage de ses sujets d'étude, leur apparence physique, aussi bien que leurs mouvements.

1.8 Comment choisir un moyen biométrique ?

Les principales contraintes liées à la biométrie sont dues à l'ergonomie et à l'acceptabilité de certaines modalités. Mais si la reconnaissance d'iris ou d'empreintes digitales sont généralement mal perçues par le public, il existe d'autres modalités, moins intrusives, comme la Reconnaissance Automatique du Locuteur (RAL) et les biométries du visage. Ces modalités présentent l'avantage d'être naturelles aux êtres humains, tout en apportant un niveau de sécurité suffisant pour un grand nombre d'applications. De plus, le matériel nécessaire - microphone et caméra - est actuellement intégré à la plupart des systèmes embarqués. D'une manière générale, les propriétés souhaitées pour tout système biométrique sont:

- ***La robustesse*** : la caractéristique biométrique doit être la plus stable possible au cours du temps et la plus difficilement altérable par le contexte d'utilisation,
- ***La distinctibilité*** : la caractéristique biométrique doit être la plus fortement dépendante de l'utilisateur,
- ***L'accessibilité*** : elle doit être facilement et efficacement mesurable par un capteur,
- ***L'acceptabilité*** : elle ne doit pas être perçue comme intrusive par l'utilisateur. Cette propriété relativement subjective dépend du contexte culturel voir politique dans lequel le système d'authentification biométrique est mis en œuvre,
- ***La disponibilité*** : pour chaque utilisateur, une quantité suffisante de mesure de la caractéristique biométrique doit être simplement disponible.

1.8.1 Comparaison des différentes technologies biométriques

Plutôt que de comparer les performances des diverses technologies (empreintes, visage, main…), il faut surtout tenir compte de l'environnement de leur usage, (facilité de : saisie, d'analyse, de stockage, de vérification). Chaque technologie possédant des avantages et des inconvénients, acceptables ou inacceptables suivant les applications. Ces solutions ne sont pas concurrentes, elles n'offrent ni les mêmes niveaux de sécurité ni les mêmes facilités d'emploi. La figure 1.4 appelée analyse de Zephyr présente une évaluation des avantages et inconvénients applicatifs de différents systèmes d'authentification biométrique basés sur différents attributs biométriques courants. Sur cette figure, les quatre critères d'évaluation des systèmes sont :

- ***La précision (accuracy)*** : mesure objective des performances obtenues par le système,
- ***Le coût (cost)*** : mesure objective du coût de mise en œuvre et de fonctionnement du système,
- ***L'ergonomie (effort)*** : mesure subjective du nombre et de la difficulté des démarches à suivre par l'utilisateur lors de son utilisation,
- ***Le caractère intrusif (intrusiveness)*** : mesure subjective, dépendant fortement de la culture de l'utilisateur et permettant d'évaluer sa perception du système.

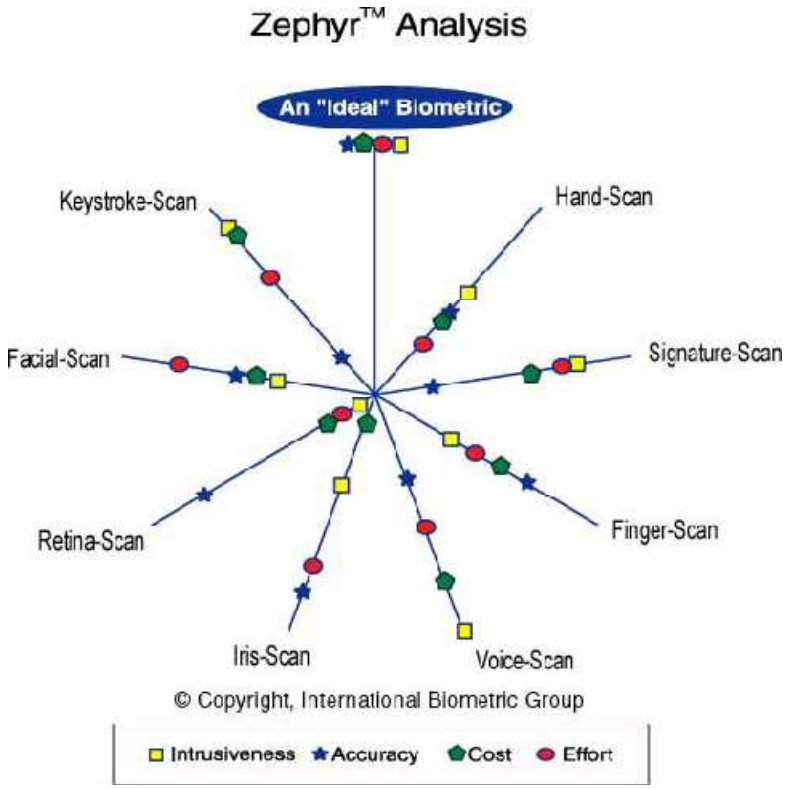


Figure 1.4 Comparaison des avantages et inconvénients applicatifs de différentes technologies biométriques d'après www.biometricgroup.com.

Selon les critères d'évaluation choisis, aucune caractéristique biométrique n'est globalement supérieure à une autre.

1.9 Qu'en est-il de la biométrie vocale ?

Une étude approfondie sur les perceptions et attitudes des consommateurs en Australie et en Nouvelle-Zélande a constaté préoccupation croissante que les formes actuelles de vérification de l'identité, tels que les codes PIN et les mots de passe, sont moins que des mesures de sécurité adéquates pour prévenir le vol d'identité. Dans un sondage en ligne de plus de 400 consommateurs, *callcentres.net*[9] a révélé que 67% des répondants étaient «très ou principalement préoccupé" par le vol d'identité par rapport à 63% il ya un an. Et tandis que les violations de sécurité de haut niveau et après la couverture médiatique peut alimenter cette préoccupation du public, l'étude, parrainée par Salmat VeCommerce, a révélé que 37% des Australiens (et 22% des Néo-Zélandais) ont vécu directement, ou connaissent un ami ou un membre de la famille qui ont connu, l'usurpation d'identité ou de vol. L'étude a révélé la biométrie vocale comme méthode privilégiée des processus d'identification (45%), suivi par code PIN (21%), mot de passe (18%) et les données personnelles ou des questions d'histoire (16%).

[9] http://callcentres.net/callcentres/Live/me.get?SITE.HOME

La biométrie vocale est considérée par la plupart des utilisateurs comme l'une des technologies biométriques la plus naturelle: elle n'est pas intrusive, n'exige aucun contact physique avec le système. De plus, elle correspond à une manière usuelle pour chaque individu de reconnaitre l'un de ses proches. Un autre avantage de la biométrie vocale est qu'elle est souvent la technologie la plus adaptée pour de nombreuses applications telles que la sécurisation de transaction bancaire téléphonique, etc. Le tableau 1.1 illustre les avantages de la biométrie vocale à savoir : simplicité d'utilisation, la précision, le coût et surtout l'acceptabilité (la non-intrusivité) ? Chose qui est rédhibitoire pour certaines technologies même si elles sont plus fiables à l'heure actuelle (ex. l'iris).

Tableau 1.1 Comparaison des technologies biométriques actuelles.

Biométrie	Empreintes digitales	Géométrie de la main	Rétine	Iris	Géométrie du Visage	Signature dynamique	Voix
Simplicité d'utilisation	H	H	L	M	M	H	H
Précision	H	H	VH	VH	H	H	M
Coût	M	VH	M	VH	M	M	L
Acceptabilité	M	M	M	M	M	M	H
Niveau de sécurité	H	M	H	VH	M	M	M
Stabilité long terme	H	M	H	H	M	M	M

L: Low; M: Medium; H: High et VH: Very High

En ce qui concerne le potentiel du marché de la biométrie vocale, *Opus Research*[10] prévoit que les dépenses mondiales sur ce type de technologie biométrique atteindront à peu près 260 millions de dollars d'ici 2014 avec un taux de croissance annuel avoisinant les 16%.

[10] Voice Biometrics Conference New York May 4-5, 2010

1.10 Bibliographie

(Cunedo et al., 1997) Cunado D., Nixon M. S., et Carter J. N., Using gait as a biometric, via phase-weighted magnitude spectra. Springer. Vol. 26, 1997.

(Israel et al., 2005) Israel S. A., Irvine J. M., Cheng A., Wiederhold M. D., et Wiederhold B. K., ECG to identify individuals. Pattern Recognition 38(1), 133–142. 27, 2005.

(Jain et al., 1999) Jain D. A. K., Bolle R., et Pankanti S., Biometrics: Personal Identification in Networked Society. Kluwer Academic Publishers. 25, 1999.

(Liu et Silverman, 2001) Liu S. et Silverman M., A practical guide to biometric security technology, IT Professional Journal, Vol. 3(1), January 2001.

(Marcel et Millan, 2007) Marcel S.et J. Millan del R., Person Authentication using Brainwaves (EEG) and Maximum A Posteriori Model Adaptation. IEEE transactions on Pattern Analysis and Machine intelligence 29(4), 743–748. 27, 2007.

(Monrose et Rubin, 2000) Monrose F. et Rubin A. D., Keystroke dynamics as a biometric for authentication. Future Gener Comput Syst 16(4), 351–359. 26, 2000.

(Oeezr Abowd, 2000) Orr R. J. et Abowd G. D., The smart floor : a mechanism for natural user identification and tracking. Dans les actes de Conference on Human Factors in Computing Systems, New York (USA), 275–276. ACM. 26, 2000.

(Rodriguez et al., 2008) Rodríguez R. V., Lewis R. P., Mason J. S., et Evans N. W., Footstep recognition for a smart home environment. International Journal of Smart Home 2(2), 95–110. 26, 2008.

(Rodriguez et al., 2007) Rodriguez R. V., Evans N., Lewis R., Fauve B., et Mason J., An experimental study on the feasibility of footsteps as a biometric. Dans les actes de European Signal and Image Processing Conference (EUSIPCO). 26, 2007.

Chapitre

2 La biométrie vocale vs Reconnaissance Automatique du Locuteur (RAL)

Sommaire

Résumé

Ce chapitre est consacré aux principes de la reconnaissance automatique du locuteur. Il présente tout d'abord les mécanismes de production de la parole et les principales sources de variabilités pour comprendre comment un individu peut être reconnu par sa voix. Il présente aussi les différentes tâches liées à la RAL. Nous exposons quelques approches utilisées pour les systèmes de RAL et présentons, enfin, les méthodes d'évaluation des performances des systèmes de RAL.

2.1 Introduction

La voix est porteuse d'informations variées. Émission de sons structurée, la parole humaine est essentiellement un vecteur de communication. À ce titre, un signal de parole est généralement porteur d'un message à destination d'une autre personne. La parole peut cependant contenir de nombreuses informations telles que la langue parlée par le locuteur, son identité ou même des indications sur son âge ou son état émotionnel.

Les systèmes de reconnaissance automatique de la parole sont utilisés pour transcrire le message porté par le signal vocal. Il peut s'agir de reconnaître un lexique limité et déterminé à l'avance (Furui, 1986) ou de transcrire un message au vocabulaire plus large (Aubert, 2002). Ce message est prononcé dans un contexte qui, s'il est connu, peut apporter une information sur le message porté par le signal de parole. Les systèmes de reconnaissance de la parole, cherchent à extraire du signal acoustique une information, a priori, indépendante du locuteur. Ils sont souvent perturbés par les variations inter-locuteurs.

La reconnaissance de la parole requiert la plupart du temps une étape préalable de reconnaissance de la langue parlée. Des approches automatiques comme celles décrites dans (Zissman et Singer, 1994 ; Singer et al., 2003; Rouas et al., 2005) permettent d'identifier la langue dans laquelle s'exprime un locuteur parmi un panel de langues connues par ces systèmes.

La reconnaissance des émotions, constitue également, depuis quelques années, un domaine de recherche en plein essor. L'analyse du signal de parole peut apporter des informations sur la volonté ou les émotions ressenties par le locuteur (Adami, 2007). L'âge d'un locuteur peut également être estimé d'après sa voix (Minematsu et al., 2003). De la même façon, certaines pathologies influent sur les organes de production de la parole et peuvent être détectées par des systèmes automatiques (Sáenz-Lechón et al., 2006; Pouchoulin et al., 2007).

Enfin, l'identité est l'une des informations les plus communément extraites du signal de parole. L'être humain est naturellement capable de reconnaître la voix d'une personne qui lui est proche. Il apparaît donc intuitif, dans un contexte de sécurisation des communications et des données, que des systèmes automatiques soient utilisés afin de reconnaître l'identité d'un locuteur d'après sa voix.

Le signal sonore, tel qu'il est utilisé par les systèmes automatiques, est aussi porteur d'une information relative au matériel qui compose la chaîne d'enregistrement et de transmission ou à l'environnement du locuteur. Ces informations sont généralement perçues comme nuisibles, car elles dégradent fortement les performances dans les différentes tâches des systèmes automatiques.

2.2 Les variabilités du signal de parole

Les informations transmises par le signal de parole sont multiples (Doddington, 1985). La variabilité du signal de parole entre locuteurs est majoritairement utilisée en reconnaissance du locuteur pour reconnaître les individus. Il s'agit de la variabilité

inter-locuteurs. La capacité des systèmes de RAL à authentifier une personne repose essentiellement sur la capacité à discriminer les individus grâce à cette variabilité. Mais d'autres facteurs de variations altèrent la parole. Le signal de parole est par exemple considéré comme non reproductible par son locuteur. Il existe une variabilité propre au locuteur dépendante de son état physique mais aussi psychologique. Il s'agit de la variabilité intra-locuteur. De plus les conditions d'environnement influent sur l'onde de parole ; les bruits additifs ambiants ou les bruits de convolution engendrés par la prise de son modifient le signal de parole.

2.2.1 La variabilité inter locuteur

Le signal de parole permet la communication entre les individus. Il véhicule un message linguistique mais aussi quantités d'informations extra linguistiques et des informations liées au locuteur. La personnalité, au sens large, du locuteur influence la production du signal de parole. Ceci permet notamment de discriminer les locuteurs.

Le signal de parole est un signal très complexe où se mêlent différents types d'informations classées par leur « niveau » de représentation. Les informations dites «bas niveau» sont facilement utilisables à partir de l'analyse numérique du signal de parole. Elles regroupent des informations liées principalement à des traits physiques de l'individu (facteurs morphologiques et physiologiques). Les informations de «haut niveau», comme la linguistique ou l'état émotionnel du locuteur sont beaucoup plus complexes à caractériser. Ces informations sont relatives aux facteurs socio-culturels de l'individu.

Ben (Ben, 2004) propose une hiérarchie composée sur 6 niveaux d'informations différents:

- ***le niveau acoustique*** : les paramètres sont liés à l'analyse de l'enveloppe spectrale du signal;
- ***le niveau prosodique*** : désigne la «mélodie» de l'énoncé de parole;
- ***le niveau phonétique*** : la distinction des différents sons identifiables d'une langue;
- ***le niveau idiolectal*** : se rapporte aux particularités langagières propres à un individu;
- ***le niveau dialogal*** : définit la façon de communiquer d'un individu, comme ses temps de parole dans une conversation;
- ***le niveau sémantique*** : caractérise la signification du discours.

Le niveau acoustique est le plus utilisé en RAL où l'influence de l'anatomie du locuteur sur l'émission du signal de parole est retenue (Furui, 1986 ; Rosenberg et Sambur, 1975). Ces approches se basent sur une représentation numérique de l'enveloppe du signal, définie comme une suite de paramètres, les cepstres. Ces informations, présentes au niveau de l'enveloppe du signal, sont facilement extraites. Les niveaux prosodique et phonétique sont basés sur la représentation du signal de parole à un niveau supérieur.

La prosodie caractérise le style d'élocution du locuteur. Les méthodes pour analyser cette information sont plus complexes bien que souvent basées sur l'analyse numérique du signal. Le niveau phonétique implique l'utilisation d'une segmentation en phonèmes, le plus souvent réalisée par un reconnaisseur de parole. Ces approches sont de plus en plus employées en RAL, en combinaison avec les approches acoustiques.

Enfin les niveaux idiolectal, dialogal et sémantique ne sont, à notre connaissance, pas utilisés en RAL.

2.2.2 La variabilité intra-locuteur

Il est impossible pour un même individu de reproduire exactement le même signal de parole. Les facteurs de variabilités pour un même individu sont multiples. Ils peuvent être liés à la nature physiologique de l'individu. Dans ce cas cette variabilité intra-locuteur est induite par l'évolution naturelle (volontaire ou non) de la voix d'une personne. L'état pathologique est un exemple de variation de la voix involontaire d'une personne. De plus, une altération de la voix due à l'âge est présente chez tous les individus. Cette variabilité est une difficulté majeure en RAL.

2.2.3 Les facteurs « extérieurs »

La variabilité inter-session (entre sessions d'enregistrements) fait apparaître l'influence de facteurs extérieurs sur le signal de parole. A la sortie du conduit vocal humain, l'onde de parole est considérée comme idéale, car aucune déformation/distorsion de l'environnement extérieur ne l'a modifiée. L'environnement sonore lors de l'enregistrement, le matériel d'acquisition ou le canal de transmission utilisé vont ensuite déformer l'onde sonore originelle. Le canal de transmission, par exemple, agit comme un filtre en fréquence sur l'onde sonore. Ces facteurs rendent complexe la comparaison entre plusieurs échantillons d'un même individu. De nombreux travaux expérimentaux ont montré que des variations de matériel entre les phases d'apprentissage et de test sont à l'origine de graves dégradations des performances (Van Vuuren, 1996). Par exemple, l'acquisition d'un signal de parole sur le réseau GSM introduit les dégradations suivantes sur le signal de parole :

- l'ajout du bruit de l'environnement,
- le sous-échantillonnage à 8kHz du signal,
- le filtrage sur la bande de fréquence [300-3400]Hz,
- le codage à bas débit de la parole,
- l'ajout du bruit de quantification des paramètres émis,
- la transmission sur un lien sans-fil avec pertes.

2.3 Extraction d'information du signal de parole

De par sa complexité, l'information portée par le signal de parole ne peut, quelle que soit la tâche considérée, être utilisée dans sa totalité. C'est pourquoi l'utilisation d'un système automatique nécessite une sélection préalable des informations à exploiter pour la tâche qui lui est confiée.

En reconnaissance automatique du locuteur, seules les informations présentant une forte variabilité inter-locuteurs permettent de discriminer les différents individus. À l'inverse, les informations dont la variabilité intra-locuteur est élevée rendent la tâche de RAL plus complexe.

Les informations les plus utilisées en RAL, du fait de leur fort potentiel discriminant, sont des informations acoustiques obtenues périodiquement par une analyse fréquentielle ou temporelle du signal. D'autres informations présentes dans le signal de parole et citées précédemment peuvent s'avérer discriminantes dans le cadre de la reconnaissance du locuteur. Des paramètres tels que la prosodie ou la fréquence fondamentale, par exemple, contiennent une information spécifique du locuteur (Ferrer et al., 2007; Sönmez et al., 1997).

Les systèmes de reconnaissance du locuteur utilisent des représentations du signal de parole dans lesquelles le bruit et la redondance ont été réduits afin de ne conserver que les informations considérées comme utiles à la tâche spécifiée. Ce traitement est appelé *paramétrisation*. Cette paramétrisation constitue l'objectif principal de cette thèse.

2.4 Fonctionnement d'un système RAL

La tâche de reconnaissance de personnes se décompose en deux phases. La première étape consiste à obtenir une représentation de l'utilisateur. Elle est appelée «enrôlement». Cette étape joue un rôle essentiel dans le processus de reconnaissance. Lors de cette phase, le système construit sa représentation de l'individu, représentation qui sera utilisée par la suite pour autoriser ou l'authentification. Cette représentation de l'utilisateur doit être unique et permanente.

La deuxième étape est l'étape de test. Des données provenant d'un utilisateur souhaitant être authentifié sont soumises au système. Cet utilisateur annonce, de plus, une identité connue du système. Le test consiste à mesurer la ressemblance entre les données fournies par l'utilisateur et le modèle existant correspondant à l'identité annoncée. Les étapes d'enrôlement et de test sont décrites dans la suite, sous la forme d'une succession de modules fonctionnels. Nous présentons enfin les éléments permettant d'apprécier les performances des systèmes de d'authentification.

2.4.1 Structure de la phase d'enrôlement

La phase d'enrôlement peut être décrite comme l'utilisation de deux modules: le module de paramétrisation et celui de modélisation, comme représentés (Figure 2.1).

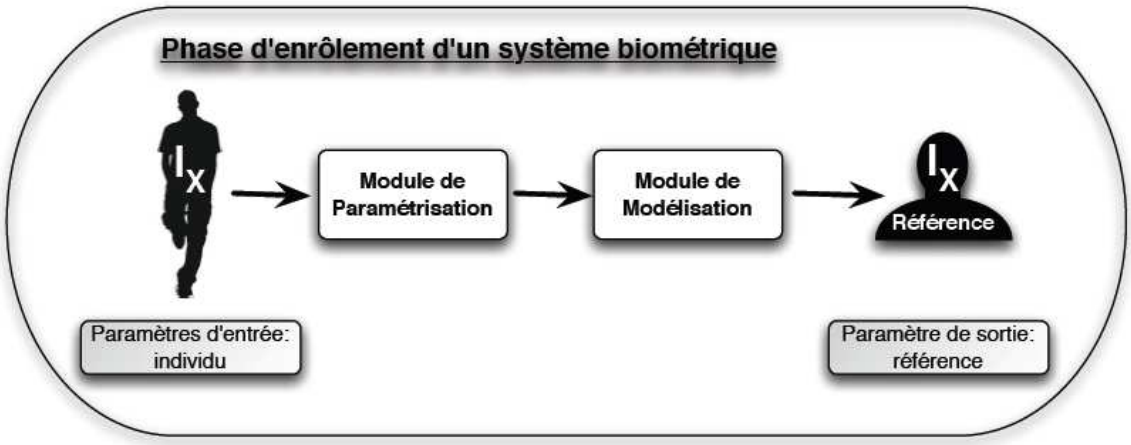


Figure 2.1 Schéma de principe de la phase d'enrôlement d'un système RAL.

2.4.1.1 Le module de paramétrisation

La variation de la nature du signal acoustique rend le traitement des données brutes issues de ce dernier très difficile. En effet, ces données contiennent des informations complexes, souvent redondantes et mélangées à du bruit.

Le module de paramétrisation, qui traite le signal acoustique reçu, doit remplir plusieurs objectifs :

- séparer le signal du bruit ;
- extraire l'information utile à la reconnaissance;
- convertir les données brutes à un format directement exploitable par le système.

Chacune de ces tâches pose des problèmes complexes et influe fortement sur les résultats des systèmes automatiques de reconnaissance.

2.4.1.2 Le module de modélisation

Le module de modélisation exploite les données fournies par le module de paramétrisation afin de créer la représentation d'un individu qui servira, par la suite, à l'authentifier. Le modèle utilisé est généralement une représentation statistique des données acquises lors de la phase d'enrôlement.

Quelle que soit la nature du modèle créé, il doit cependant respecter certaines contraintes :

- il doit être le plus précis possible afin de limiter l'ambiguïté inter-individus ;
- il doit prendre en compte la variabilité intra-individus afin de représenter l'individu au cours du temps.

2.4.2 Structure de la phase de test

La structure de la phase de test est représentée par la figure 2.2 et comprend (Jain et al., 2004):

- un module de paramétrisation ;
- un module de reconnaissance ;
- un module de décision.

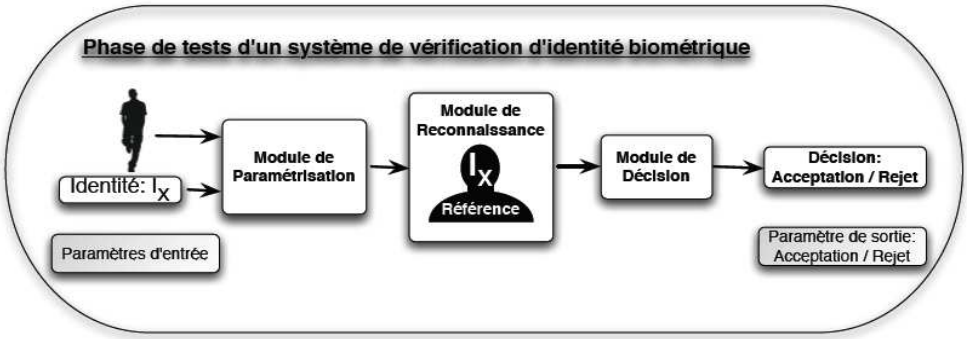


Figure 2.2 Schéma de principe de la phase de test d'un système RAL.

2.4.2.1 *Le module de paramétrisation*

Le module de paramétrisation est le même que celui utilisé durant la phase d'enrôlement. Il rempli, lors de la phase de test, les mêmes tâches que durant la phase de paramétrisation : séparation signal/bruit, extraction de l'information utile et, éventuellement, normalisation des données. Il est important de conserver la même paramétrisation lors des phases d'enrôlement et de test afin de fournir au système automatique des informations comparables et de même nature.

2.4.2.2 *Le module de reconnaissance*

Le module de reconnaissance a un rôle central dans le système RAL. Il compare les paramètres extraits du signal brut à un modèle d'individu calculé lors de la phase d'enrôlement. Lors de cette comparaison, le module de reconnaissance calcule une mesure de similarité entre les données d'entrée et le modèle testé. C'est une valeur numérique, aussi appelée score. Cette mesure de similarité est ensuite transmise au module de décision.

2.4.2.3 *Le module de décision*

Comme le décrit la partie précédente, un module de reconnaissance fournit en sortie un score. La nature de ce score varie selon les modules de reconnaissance utilisés. Il s'agit la plupart du temps d'une distance, d'une probabilité ou d'une vraisemblance. Le module de décision doit, à partir de ce score, fournir une décision qui constituera la réponse finale du système RAL.

2.5 Classification des systèmes RAL

2.5.1 Classification par tache

La reconnaissance du locuteur est un terme générique qui répond à plusieurs définitions selon le scénario applicatif envisagé. Les scénarios applicatifs sont regroupés en cinq catégories principales :

- l'identification de locuteurs,
- la vérification du locuteur,
- l'indexation des documents audio en locuteurs,
- la segmentation en locuteur,
- le suivi de locuteurs.

Chacune de ces catégories propose son protocole de reconnaissance selon que l'identité du locuteur à reconnaître soit proclamée, ou que les locuteurs à reconnaître soient connus ou non du système de RAL. Le système de RAL peut valider une identité pour la vérification du locuteur, proposer une identité à partir d'un ensemble de locuteurs, déterminer les durées de parole d'un locuteur, compter le nombre de locuteurs présents dans un signal.

2.5.1.1 Identification automatique du locuteur

L'identification du locuteur consiste à déterminer l'identité d'un individu parmi une population de personnes connues. A partir d'un échantillon de voix enregistré, il faut déterminer quel locuteur de la base a parlé. Les données sont alors comparées à une référence caractéristique de chaque utilisateur I_X connu par le système. Le résultat de chaque comparaison est un score, fonction de la similitude observée par le système entre les données S du locuteur et la référence considérée. Le score le plus élevé correspond à la référence la plus proche des données de test et l'identité du locuteur correspondant à cette référence est retournée par le système.

Il existe deux modes d'identification automatique, en milieu ouvert ou fermé.

- ***En milieu fermé*** (Figure 2.3), le système RAL décide de l'identité la plus probable parmi les utilisateurs connus (dont il possède une référence). Ce mode de fonctionnement tend à considérer que seules des personnes référencées peuvent accéder au système. Un tel système ne doit alors être utilisé que dans un environnement au sein duquel tous les individus sont connus.

  L'identité I_Y retournée, correspondant à la référence Y est obtenue par :

$$\mathbf{Y = argmax\ f(X|S)} \qquad \mathbf{(2.1)}$$

où f (X|S) est le score calculé lors de la comparaison des données S au modèle de l'individu I_X.

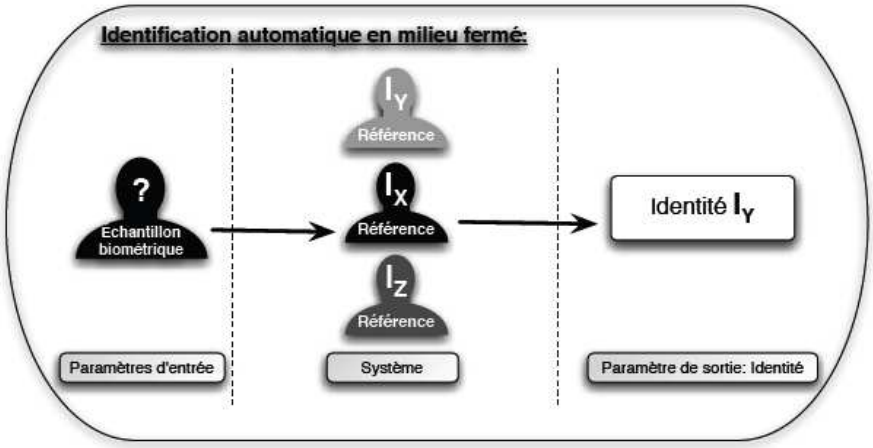


Figure 2.3 Schéma de principe de la tâche d'identification automatique en milieu fermé.

- ***En milieu ouvert*** (Figure 2.4), le système RAL a la possibilité de rejeter le locuteur dont il teste les données si elles ne correspondent à aucune des identités répertoriées. Ce locuteur est alors considéré comme inconnu du système et est rejeté. Pour ce faire, les données S sont comparées à chaque référence X connue par le système. Chaque comparaison fournit un score f(X|S). Le score le plus élevé est alors comparé à un seuil Ω, fixé au préalable. Si le score est supérieur à ce seuil, le système décide qu'il s'agit de la personne correspondant à la référence sélectionnée. Si le score est inférieur à ce seuil, le système décide qu'il ne s'agit pas d'une personne «connue».

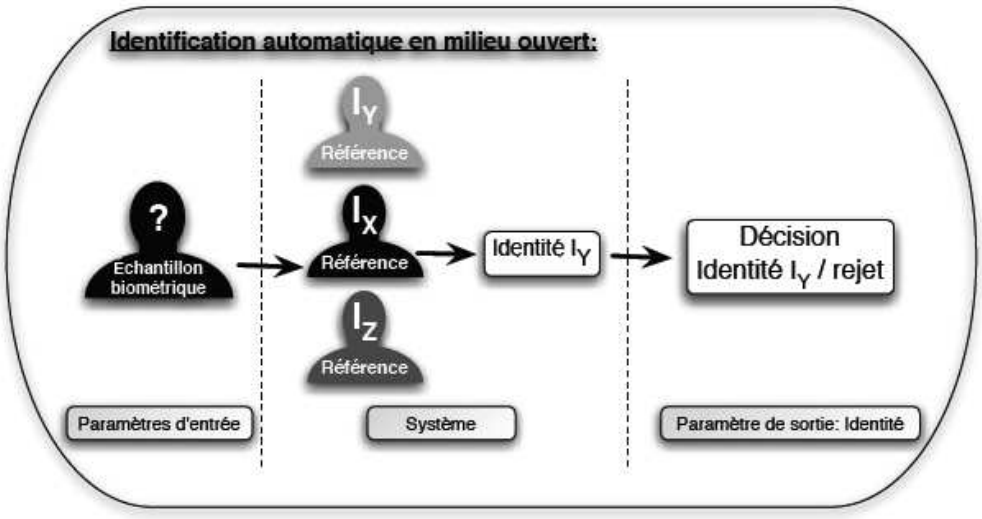


Figure 2.4 Schéma de principe de la tâche d'identification automatique en milieu ouvert.

Pour résumer, en identification en milieu ouvert, le système répond à deux interrogations: «Quelle est l'identité la plus probable ?» et «Les données biométriques analysées correspondent-elles à cette identité ?», alors qu'en milieu fermé il ne répond qu'à la première.

L'identité la plus probable est obtenue, comme dans le cas de l'identification en milieu fermé, par :

$$\mathbf{Y = argmax\ f(X|S)} \tag{2.2}$$

Et la réponse du système est donnée par :

$$\mathbf{f(Y|S)}(< ou >)\Omega \begin{cases} \boldsymbol{identité\ non\ reconnue} \\ \boldsymbol{identité\ est\ celle\ du\ client\ I_x} \end{cases} \tag{2.3}$$

où Ω est le seuil de décision fixé au préalable.

2.5.1.2 Vérification automatique du locuteur

La vérification du locuteur consiste à déterminer si l'identité proclamée d'un message vocal correspond à la véritable identité du locuteur. En pratique la réponse est binaire, acceptation ou rejet. Les éléments mis en jeu sont donc une identité proclamée, I_X, et la référence associée à un échantillon connu de l'identité proclamée. Une mesure de similarité entre le signal à vérifier et cette référence est calculée. Cette mesure est comparée à un seuil de vérification. Dans le cas où la mesure de similarité est supérieure au seuil, l'individu est accepté. Dans le cas contraire, l'individu est considéré comme un imposteur et rejeté.

De la comparaison de ces données est issue une mesure de similarité f(X|S) qui est comparée à un seuil Ω. Si la mesure de similarité est supérieure à ce seuil, l'utilisateur est accepté, si elle est inférieure il est rejeté.

La décision du système est exprimée par :

$$\mathbf{f(X|S)}(< ou >)\Omega \begin{cases} \boldsymbol{utilisateur\ rejeté} \\ \boldsymbol{utilisateur\ accepté} \end{cases} \tag{2.4}$$

La figure 2.5 illustre le schéma de principe de la vérification du locuteur. Notons toutefois que la tâche de vérification équivaut à la deuxième étape de l'identification en milieu ouvert et que la principale différence entre ces deux tâches est le nombre de références auxquelles doivent être comparées les données de test. Lors de la vérification, les données sont comparées au modèle correspondant à l'identité clamée, alors que pour l'identification en milieu ouvert, ce même échantillon de données est comparé au modèle de chacun des individus référencés par le système.

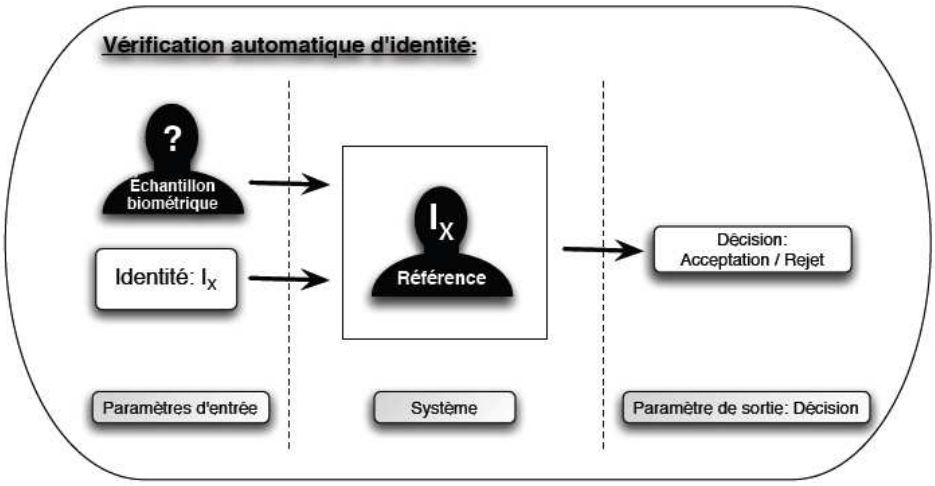


Figure 2.5 Schéma de principe de la tâche de vérification automatique d'identité.

2.5.1.3 Indexation automatique des documents audio en locuteurs

Grâce au développement des technologies numériques, les besoins en outils d'indexations se font cruellement ressentir. Il devient donc indispensable d'indexer automatiquement les documents audio pour être exploitables. La clé d'indexation qui nous intéresse ici est l'identité du locuteur : nous voudrions savoir *qui parle et quand*. En pratique, on dispose des documents audio représentés par leurs modèles respectifs. La phase de recherche du système d'indexation consiste, généralement, à évaluer des mesures de similarité entre la requête et ces différents modèles. Par ailleurs, le système d'indexation par locuteur peut servir également comme étape préliminaire pour des tâches de transcription ou pour le suivi de locuteurs.

L'indexation en locuteurs permet de déterminer les temps de parole des individus dans un signal audio. La spécificité de cette tâche réside dans le fait que le système ne détienne pas de référence pour les locuteurs présents dans le signal audio. Un mécanisme d'apprentissage aveugle et adaptatif est alors mis en place (Meignier, 2002). Il est possible de segmenter un signal audio par prise de parole des intervenants, étiqueter des données audio pour permettre des recherches de documents audio par locuteurs ou, enfin, identifier le nombre de locuteurs présents dans le signal.

2.5.1.4 Segmentation en locuteurs

La segmentation en locuteur consiste à découper le flux sonore avec ou sans modèle explicite des locuteurs ? Dans le cas où il n'y a pas de modèle, le processus procède généralement en deux phases : la détection de changements de locuteurs et le regroupement de locuteurs segments (appartenant au même locuteur). La première phase repose sur le principe de calcul d'une distance entre deux portions de signal consécutives. Si elle excède un seuil, on décide qu'il y a eu changement de locuteur. La deuxième phase consiste à regrouper les segments de parole par locuteur selon un critère de distance. Une bonne segmentation fournit les changements de locuteurs corrects et des segments ne contenant qu'un seul locuteur.

2.5.1.5 *La poursuite du locuteur*

La poursuite du locuteur se fait avec un modèle du locuteur, contrairement à la segmentation qui peut se faire sans modèle. Elle consiste à déterminer quand une personne parle dans une conversation. Elle est similaire à l'indexation en locuteur, à ceci près que les locuteurs présents dans le signal sont connus par le système de RAL. Il s'agit donc d'une simplification de la tâche d'indexation en locuteur mais qui reste néanmoins, une tâche très complexe.

2.5.2 Scénarios

Ces différentes tâches de reconnaissance du locuteur permettent de mettre en place de nombreux scénarios applicatifs distincts. La détermination du scénario s'effectue en fonction des besoins et contraintes exprimés :

- ***la tâche*** : vérifier une identité, segmenter les tours de parole des locuteurs, trouver une identité,
- ***le mode de dépendance au texte*** : prompté, fixé au préalable, libre ;
- ***les références connues des locuteurs*** : population inconnue, taille de la population connue, le sexe de la population, la ou les langue(s) parlée(s) par la population, la durée des références, la qualité d'enregistrement ;
- ***le signal audio disponible pour la tâche*** : sa taille, un locuteur présent ou plusieurs, sa qualité d'enregistrement ;
- ***le matériel disponible*** : la puissance de calcul et de stockage. Le traitement est-il différé ou en en temps réel ?

Une telle définition du scénario est nécessaire pour considérer certains principes généraux de RAL :

- selon le niveau de dépendance au texte les méthodes mises en oeuvre sont différentes,
- la paramétrisation doit être adaptée à la langue et au genre des locuteurs, comme par exemple avec le retrait de l'échelle de Mel pour les signaux féminins (Mason et Thompson, 1993) ou l'utilisation de la prosodie pour les langues tonales (Bin et Meng, 2004 ; Auckenthaler et al., 2001 ; Kleynhans et Barnard, 2005),
- la qualité et la quantité des enregistrements d'apprentissage et de test sont déterminantes,
- l'augmentation du nombre de locuteurs à reconnaître diminue les performances en identification du locuteur (Furui, 1978).

Les méthodes de RAL mises en œuvre peuvent être très différentes selon le scénario ainsi défini. La serrure vocale est un exemple d'utilisation de l'identification du locuteur en milieu ouvert. Les locuteurs ayant accès au bâtiment protégé sont connus. De manière coopérative ils ont enregistré un message (fixé ou libre) pour servir de référence au système de RAL. La vérification du locuteur peut, quant à elle, remplacer l'utilisation des mots de passe pour sécuriser les transactions bancaires, ou les accès aux systèmes informatiques. Sur les systèmes informatiques, l'utilisateur doit

proclamer son identité par un « login» et s'authentifier avec son mot de passe. La vérification du locuteur peut alors authentifier l'utilisateur en comparant un échantillon de sa voix avec une référence, enregistrée au préalable et associée au «login». Les systèmes d'indexation du locuteur sont particulièrement utiles pour le traitement des bases de données audio. Une recherche des documents audio dans lesquels un locuteur est intervenu devient possible.

2.5.3 Classification par dépendance au texte

Une seconde classification à l'intérieur de ces catégories repose sur le niveau de dépendance au texte. La reconnaissance peut être indépendante du texte ou dépendante du texte. En mode dépendant du texte la reconnaissance bénéficie de la connaissance du contenu linguistique prononcé (fixe ou prompté). L'estimation des paramètres caractéristiques du locuteur est alors plus robuste. En mode indépendant du texte, le système de reconnaissance n'a aucune connaissance sur le message linguistique prononcé par la personne. Les niveaux de dépendance au texte sont classés suivant les applications :

- ***Systèmes à texte libre (ou free-text)*** : le locuteur est libre de prononcer ce qu'il veut. Dans ce mode, les phrases d'apprentissage et de test sont différentes.
- ***Systèmes à texte suggéré (ou text-prompted)*** : un texte, différent à chaque session et pour chaque personne, est imposé au locuteur et déterminé par la machine. Les phrases d'apprentissage et de test peuvent être différentes.
- ***Système dépendant du vocabulaire (ou vocabulary-dependent)*** : le locuteur prononce une séquence de mots issus d'un vocabulaire limité ? Dans ce mode, l'apprentissage et le test sont réalisés sur des textes constitués à partir du même vocabulaire.
- ***Systèmes personnalisés dépendant du texte (ou user-specific text dependent)*** : chaque locuteur a son propre mot de passe. Dans ce mode, l'apprentissage et le test se font réalisés sur le même texte.

D'évidence, la connaissance à priori du message vocal rend la tâche des systèmes RAL plus facile et les performances sont meilleures. La reconnaissance en mode indépendant du texte nécessite plus de durée de parole que le mode dépendant du texte.

2.6 Evaluation d'un système de RAL

Un système de RAL peut être confronté à deux types de tests :

- ***les tests clients*** : lors desquels l'échantillon biométrique présenté au système correspond à l'identité clamée ;
- ***test imposteur*** : lors desquels l'échantillon biométrique présenté au système provient d'un individu inconnu du système.

Le système automatique doit répondre à chaque tentative d'authentification auquel il fait face par une décision binaire. Il peut donc engendrer deux types d'erreurs :

- ***Faux rejet (FR)*** : erreur commise lorsque le système rejette, à tort, un client légitime (i.e. erreur commise lors d'un test client) ;
- ***Fausse acceptation (FA)*** : erreur commise lorsqu'un imposteur est malencontreusement accepté en tant qu'utilisateur légitime (i.e. erreur commise lors d'un test imposteur).

Ces deux types d'erreurs n'ont pas toujours la même incidence en terme de sécurité et de qualité de service. La fausse acceptation peut être très pénalisante dans le cas d'une application requérant un niveau de sécurité élevé. Il n'est pas tolérable par exemple que n'importe qui puisse accéder à des informations personnelles, bancaires ou même de type secret défense. Le faux rejet peut également pénaliser des applications où l'utilisateur ne peut se permettre de perdre du temps en tentant de s'authentifier à plusieurs reprises. C'est le cas, par exemple, pour des services de secours d'urgence. Un utilisateur du système doit pouvoir être reconnu par le système dans les meilleurs délais.

Nous verrons dans la suite que les taux de fausses acceptations et de faux rejets sont liés et que le réglage d'un système de vérification d'identité doit tenir compte du coût de chaque type d'erreurs dans le cadre de l'application visée.

En fonction du type d'application souhaitée, le seuil de vérification peut être choisi pour minimiser le taux de fausses acceptations : application de sécurité, ou minimiser le taux de faux rejets pour augmenter l'ergonomie d'utilisation. Il n'est pas possible de minimiser conjointement ces deux taux (Figure 2.6).

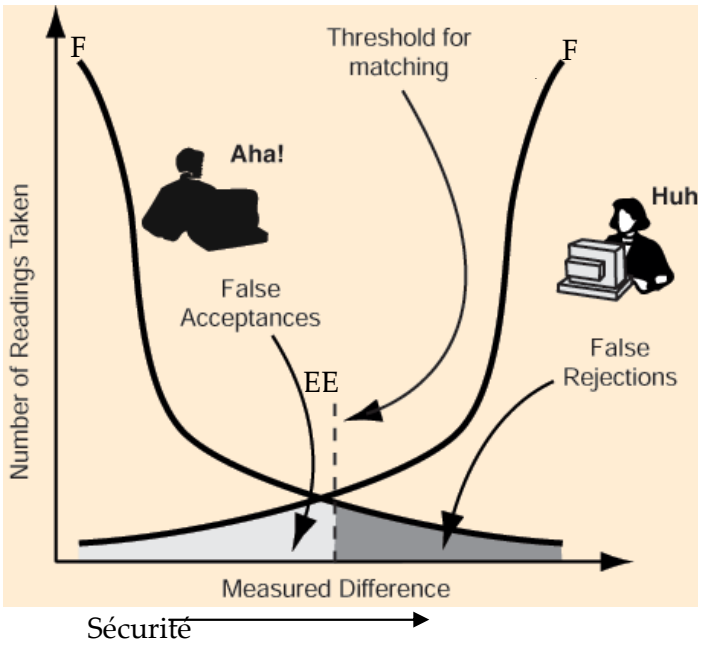


Figure 2.6 Evolution des taux FA et FR.

2.6.1 Point de fonctionnement et représentation des performances

Le module de décision décrit dans le paragraphe 2.4.2.3 reçoit, en entrée et pour chaque test, un score. Celui-ci résulte de la comparaison entre les caractéristiques biométriques de l'utilisateur testé et la référence apprise lors de la phase d'enrôlement. Un score élevé signifiera que la probabilité pour que l'utilisateur testé corresponde à l'identité qu'il annonce est élevée et un score faible signifiera que cette probabilité est faible. La décision binaire qui constitue la sortie du module résulte de la comparaison de ce score avec un seuil défini à l'avance. Si le score est supérieur au seuil, l'utilisateur est accepté et s'il est inférieur au seuil, l'utilisateur est rejeté.

Le choix d'un seuil a une incidence directe sur les performances du système. Pour un système idéal, les scores obtenus par les clients seront tous plus élevés que les scores obtenus par les imposteurs. Dans ce cas, le seuil à fixer se situe entre le score imposteur le plus élevé et le score client le plus faible, assurant ainsi une authentification parfaite (Figure 2.7).

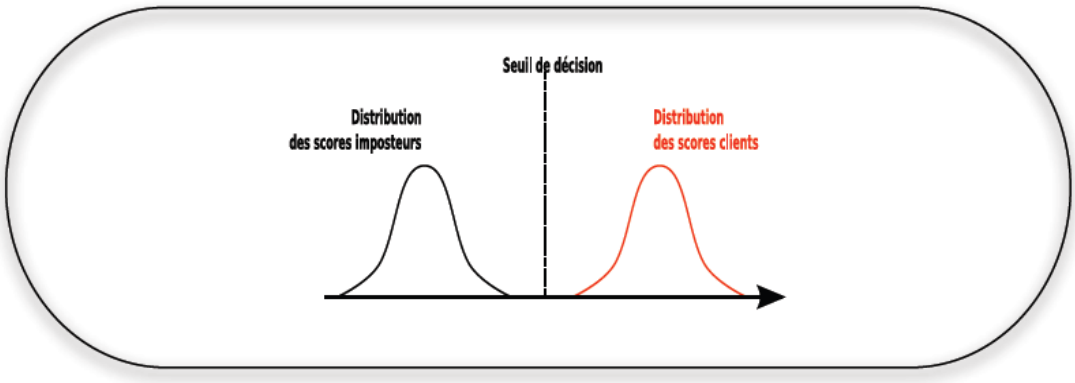


Figure 2.7 Répartition des scores clients et imposteurs et seuil de décision d'un système parfait.

En pratique, les distributions des scores clients et imposteurs se superposent partiellement. Ce cas ne permet pas une authentification parfaite et des erreurs de type faux rejets et fausses acceptations apparaissent. Le choix du seuil influe sur le taux de faux rejets et de fausses acceptations. Cet effet est illustré par la figure 2.8.

2.6.2 Les courbes DET

Pour un seuil de décision fixé les taux de faux rejet p(FR) et de fausse acceptation p(FA) que l'utilisation de ce seuil occasionne peuvent être calculés a posteriori.

$$p(FA) = \frac{nombre\ de\ tests\ dont\ résulte\ une\ fausse\ acceptation}{nombre\ de\ tests\ imposteurs} \quad (2.5)$$

$$p(FR) = \frac{nombre\ de\ tests\ dont\ résulte\ un\ faux\ rejet}{nombre\ de\ tests\ clients} \quad (2.6)$$

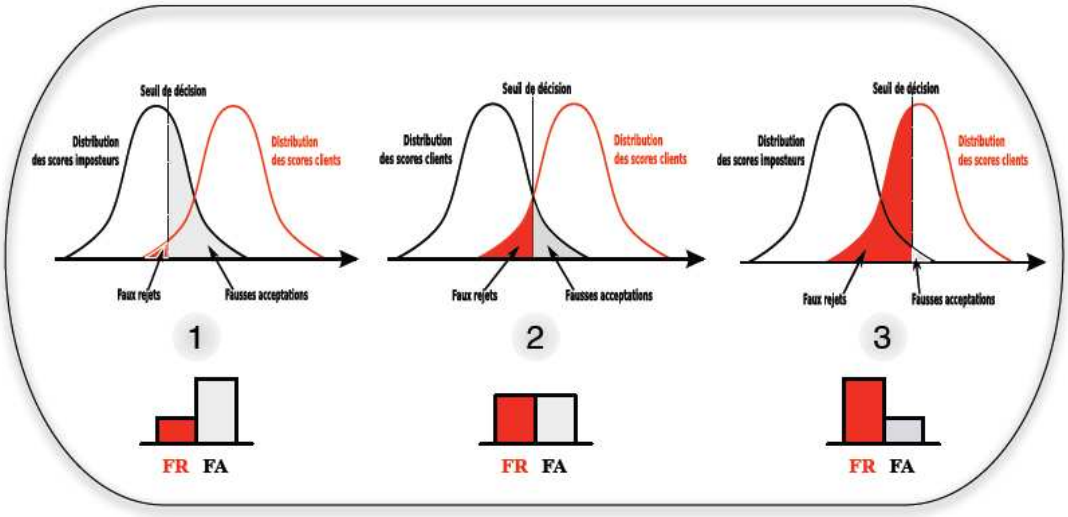


Figure 2.8 Influence du seuil de décision sur les erreurs d'un système de reconnaissance biométrique.

1 - Seuil de décision choisi dans le but de réduire le nombre de faux rejets
2 - Seuil de décision choisi pour obtenir autant de faux rejets que de fausses acceptations (EER)
3 - Seuil de décision choisi dans le but de réduire le nombre de fausses acceptations

À chaque valeur de seuil est associé un couple (p(FA), p(FR)) et l'ensemble des couples obtenus peut être représenté sous la forme d'une courbe ROC (Receiver Operating Characteristic) (Oglesby, 1995) ou d'une courbe DET (Detection Error Tradeoff) qui est la représentation la plus communément utilisée pour évaluer la pertinence du seuil de décision en fonction de ces deux taux d'erreurs. La courbe DET diffère principalement de la courbe ROC par l'échelle basée sur une distribution normale qui se substitue à l'échelle linéaire (Martin et al., 1997). Les échelles des axes suivent la répartition d'une loi normale contrairement à leurs prédécesseurs, les courbes ROC, qui utilisent une échelle linéaire. La figure 2.9 illustre un exemple de courbe DET.

D'autres solutions ont été proposées pour la représentation des performances d'un système de RAL :

- la courbe EPC (Expected Performance Curve) (Mariethoz , 2006),
- la courbe APE (Applied Probability of Error) (Van Leeuwen et Brummer, 2007).

Pour comparer les systèmes de RAL deux points de fonctionnement sont extraits pour caractériser plus simplement ces courbes. Le taux d'erreurs égales ou EER (Equal Error Rate) défini comme le point de fonctionnement où FA = FR. A ce point de fonctionnement aucune priorité n'est donnée à la minimisation des FA ou de FR. Cette mesure est très utilisée pour comparer les performances des systèmes de RAL.

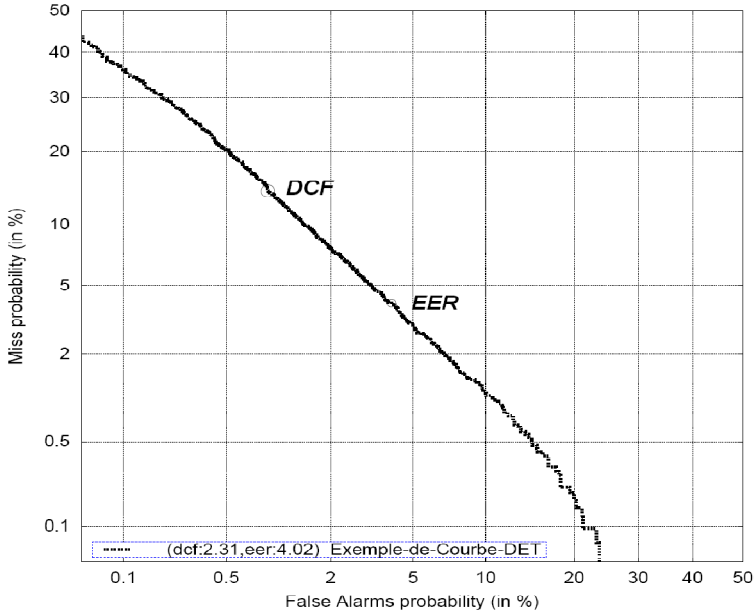


Figure 2.9 Exemple de courbe DET (False alarms : FA, Miss probability : FR).

Pour introduire une pondération pour chacun de ces taux, en fonction du contexte applicatif, une fonction de coût de détection (DCF, Decision Cost Function) peut être appliquée. Cette DCF s'exprime sous la forme :

$$\mathbf{DCF = C_{FA}\tau_{FA}P_{false} + C_{FR}\tau_{FR}P_{true}} \quad \textbf{(2.7)}$$

où :

τ_{FA} est le taux de fausses acceptations ;
τ_{FR} est le taux de faux rejets ;
C_{FA} est le coût associé à une fausse acceptation ;
C_{FR} est le coût associé à un faux rejet ;
P_{true} est la probabilité a priori d'un accès client ;
P_{false} la probabilité d'une imposture.

Une autre mesure, dénommée HTER ou Half Total Error Rate, est définie comme la distribution du taux d'erreur moyen pour chaque seuil de décision (Bengio et Mariethoz, 2004).

$$\mathbf{HTER = \frac{1}{2(FA+FR)}} \quad \textbf{(2.8)}$$

Les taux d'erreurs sont liés au point de fonctionnement d'utilisation. Le réglage du seuil de décision est effectué sur une population de tests, à priori. La calibration de ce seuil est très importante. Une variation du seuil entre la phase de calibration et de fonctionnement éloigne le système du point de fonctionnement optimal souhaité.

Le point de fonctionnement réel peut être déterminé a posteriori. C'est notamment le cas lors de campagnes d'évaluations des systèmes de VAL. Le point de fonctionnement optimal qui minimise le critère DCF est comparé au point de fonctionnement fixé a priori. Cette mesure, nommée minDCF, permet d'évaluer l'erreur de calibration du seuil de décision.

En général, pour comparer les performances des systèmes de RAL, le pourcentage relatif de gain/perte, pour les mesures DCF et EER, est utilisé :

$$\% \textbf{ relatif} = \frac{V_1 - V_2}{V_1} \tag{2.9}$$

où V peut être la mesure EER ou DCF.

2.6.3 Normalisation

La plupart des systèmes de vérification du locuteur état-de-l'art intègrent une étape de normalisation avant la prise de décision. Cette étape permet de prendre en compte la variabilité des scores obtenus lors des différents tests. La variabilité provient principalement des différences de locuteurs, contenus phonétiques ou durées d'enregistrement d'un test à l'autre. La variabilité intra-locuteur doit également être prise en compte afin de fixer le seuil de décision du système automatique. La plupart des approches état-del'art reposent sur une normalisation des distributions de scores imposteurs. Nous inciterons cependant le lecteur à se rapporter aux travaux de (Higgins et al., 1991; Li et Porter, 1998; Matsui et Furui, 1993; Reynolds, 1996; Auckenthaler et al., 2000; Fredouille et al., 1999).

2.7 Les approches classiques pour la RAL

Différentes méthodologies sont utilisées en RAL pour réaliser les références de locuteurs. Les approches génératives regroupent des méthodes qui utilisent les données d'apprentissage pour modéliser les densités de probabilité de chaque classe, par une famille de fonctions paramétriques. L'approche générative dominante pour représenter la référence du locuteur, en RAL indépendante du texte, est le modèle de mélanges de Gaussiennes (GMM, Gaussian Mixture Model). Elle a été introduite par (Reynolds et Rose, 1995; Reynolds et al., 2000) et constitue l'état de l'art des systèmes de RAL.

Il existe d'autres approches génératives comme les modèles de Markov cachés (HMM, Hidden Markov Model). Les HMM sont très employés en RAL dépendante du texte car ils sont capables de capturer les dépendances temporelles entre différentes variables aléatoires. Dans le cas de la RAL dépendante du texte, la modélisation des variations temporelles, des distributions des paramètres acoustiques, permet de très bonnes performances (Rosenberg et Soong, 1992).

Les approches à base de quantification vectorielle ont été utilisées en RAL. Elles proposent une représentation minimale d'une classe de paramètres observés : un représentant (dans un dictionnaire) pour chaque classe (Soong et al., 1985). Chaque classe de paramètres est déterminée par un algorithme de classification du type K-moyennes. Cette représentation est choisie en minimisant la distance entre le centroïde et les paramètres de la population observée. Ces approches ne sont plus très employées depuis l'apparition des GMM en RAL.

L'approche discriminante la plus employée en RAL sont les Support Vector Machine (SVM) (Wan et Campbell, 2000). A l'origine, ils ont été conçus comme une fonction discriminante permettant de séparer au mieux des régions complexes dans des problèmes de classification à 2 classes. Ils démontrent aujourd'hui des performances similaires à l'approche GMM. Ces deux méthodes sont aussi combinées dans un nouveau formalisme, le GMM/SVM Super-Vecteur (Campbell et al., 2006; Matrouf et al., 2008) qui profite des capacités génératives du GMM et discriminantes du SVM.

2.8 Conclusion

Dans ce chapitre, on a présenté un état de l'art et les principes aspects et concepts utilisés en RAL. On a présenté la structure générale d'un système en RAL et ses composants modulaires. Pour chaque module, on a donné les différentes techniques utilisées, souvent en citant leurs avantages et leurs faiblesses. Particulièrement, les paramètres MFCC et les modèles GMM sont utilisés dans la majorité des systèmes en reconnaissance du locuteur. Cependant, ces derniers sont reconnus pour leurs sensibilités aux bruits et aux distorsions introduites par les canaux de transmissions. De plus, les modèles ne tiennent pas compte de l'influence de la séquence temporelle des unités phonétiques alors que les paramètres négligent en grande partie la contribution liée à la cavité supra-glottique pendant le processus de phonation.

2.9 Bibliographie

(Adami, 2007) A. G. Adami, 2007. Modeling prosodic differences for speaker recognition. Speech Communication 49(4), 277–291.

(Aubert, 2002) X. Aubert, 2002. An overview of decoding techniques for large vocabulary continuous speech recognition. Computer speech & language 16(1), 89–114.

(Auckenthaler et al., 2000) R. Auckenthaler, M. Carey, et H. Lloyd-Thomas, 2000. Score Normalization for Text-Independent Speaker Verification System. Digital Signal Processing 1(10), 42–54.

(Auckenthaler et al., 2001) R. Auckenthaler, M. Carey, et J. Mason, 2001. Language dependency in text-independent speaker verification. Dans les actes de ICASSP.

(Ben, 2004) M. Ben, 2004. Approches robustes pour la vérification du locuteur par normalisation et adaptation hiérarchique. Thèse de Doctorat, Université de Rennes 1.

(Bengio et Mariethoz, 2004) S. Bengio et J. Mariethoz, 2004. The expected performance curve : a new assessment measure for person authentication. Dans les actes de 2001 : A Speaker Odyssey - The Speaker Recognition Workshop.

(Bin et Meng, 2004) M. Bin et H. Meng, 2004. English-chinese bilingual textindependent speaker verification. Dans les actes de ICASSP.

(Campbell et al., 2006) W. M. Campbell, D. E. Sturim, D. E. Sturim, D. A. Reynolds, et D. A. Reynolds, 2006. Support vector machines usingGMMsupervectors for speaker verification. Signal Processing Letters, IEEE 13(5), 308–311.

(Doddington, 1985) Doddington G., Speaker recognition - Identifying people by their voices. Proceedings of the IEEE, November, 73(11): 1651, 1985.

(Ferrer et al., 2007) L. Ferrer, E. Shriberg, S. Kajarekar, et K. Sonrnez, 2007. Parameterization of prosodic feature distributions for SVM modeling in speaker recognition. Dans les actes de IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Volume 4.

(Fredouille et al., 1999) C. Fredouille, J.-F. Bonastre, et T. Merlin, 1999. Similarity Normalization Method Based on World Model and a Posteriori Probability for Speaker Verification. Dans les actes de European Conference on Speech Communication and Technology (Eurospeech), Volume 2, Budapest (Hungary), 983–986.

(Furui, 1986) S. Furui, 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Transactions on Acoustics, Speech and Signal Processing 34(1), 52–59.

(Furui, 1978) S. Furui, 1978. Research on Individuality Information in Speech Waves. Thèse de Doctorat, Tokyo University

(Kleynhans et Barnard, 2005) N. Kleynhans et E. Barnard, 2005. Language dependence in multilingual speaker verification. Dans les actes de Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa, 117–121.

(Higgins et al., 1991) A. Higgins, L. Bahler, et J. Porter, 1991. Speaker verification using randomized phrase prompting. Dans les actes de Digital Signal Processing, Volume 1, 89–106. 40, 114

(Li et Porter, 1998) K.-P. Li et J. E. Porter, 1998. Normalizations and selection of speech segments for speaker recognition scoring. Dans les actes de IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Volume 1, New York (USA), 595–598.

(Martin et al., 1997) A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, et M. A. Przybocki, 1997. The DET Curve in Assessment of Detection Task Performance. Dans les actes de European Conference on Speech Communication and Technology (Eurospeech). ISCA.

(Mariethoz, 2006) J. Mariethoz, 2006. Algorithmes d'apprentissage discriminants en vérification du locuteur. Thèse de Doctorat, Lyon II Lumière.

(Mason et Thompson, 1993) J. S. Mason et J. Thompson, 1993. Gender effects in speaker recognition. Dans les actes de ICSLP, 733–736.

(Matrouf et al., 2008) D. Matrouf, J.-F. Bonastre, C. Fredouille, A. Larcher, S. Mezaache, M. McLaren, et F. Huenupan, 2008. LIA GMM-SVM system description : NIST SRE08. Dans les actes de NIST Speaker Recognition Evaluation Workshop, Montreal (Canada). 29, 60, 61

(Matsui et Furui, 1993) T. Matsui et S. Furui, 1993. Concatenated phoneme models for text-variable speaker recognition. Dans les actes de IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Volume 2, Minneapolis (USA), 391–394. 40, 63, 65

(Meignier, 2002) S. Meignier. Indexation en locuteurs de documents sonores: Segmentation d'un document et Appariement d'une collection. PhD thesis, LIA Avignon, France, 2002.

(Minematsu et al., 2003) N. Minematsu, K. Yamauchi, et K. Hirose, 2003. Automatic estimation of perceptual age using speaker modeling techniques. Dans les actes de European Conference on Speech Communication and Technology (Eurospeech). ISCA.

(Oglesby, 1995) J. Oglesby, 1995. What's in a number ? Moving beyond the equal error rate. Speech Communication 17(1-2), 193–208.

(Pouchoulin et al., 2007) G. Pouchoulin, C. Fredouille, J.-F. Bonastre, A. Ghio, et J. Revis, 2007. Characterization of the pathological voices (dysphonia) in the frequency space. Dans les actes de Proceedings of International Congress of Phonetic Sciences (ICPhS), Volume 16, 6–10.

(Reynolds et Rose, 1995) Douglas A. Reynolds et Richard C. Rose (1995). _ Robust textindependent speaker identi_cation using gaussian mixture speaker models _. IEEE Transactions on Acoustics, Speech and Signal Processing, 3(1):72_83.

(Reynolds, 1996) D. A. Reynolds, 1996. The effects of handset variability on speaker recognitionperformance : experiments on the Switchboard corpus. Dans les actes de IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, Volume 1.

(Reynolds et al., 2000) D. A. Reynolds, T. F. Quatieri, et R. B. Dunn, 2000. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10, 19–41. 112, 116

(Rosenberg et Sambur, 1975) A. Rosenberg et M. Sambur, 1975. New techniques for speaker verification. IEEE transactions on Acoustics, Speech and Signal Processing 23(2), 169–176.

(Rosenberg et Soong, 1992) A. E. Rosenberg et F. K. Soong, 1992. Advances in Speech Signal Processing, Chapter Recent Research in Automatic Speaker Recognition, 701 – 738. Marcel Dekker.

(Rouas et al., 2005) J.-L. Rouas, J. Farinas, F. Pellegrino, et R. AndrÈ-Obrecht, 2005. Rhythmic unit extraction and modelling for automatic language identification . Speech Communication 47(4), 436–456. ?OLDEditeur(Speech Communication, Elsevier).

(Sáenz-Lechón et al., 2006) N. Sáenz-Lechón, J. Godino-Llorente, V. Osma-Ruiz, et P. Gómez-Vilda, 2006. Methodological issues in the development of automatic systems for voice pathology detection. Biomedical Signal Processing and Control 1(2), 120–128.

(Singer et al., 2003) E. Singer, P. Torres-Carrasquillo, T. Gleason, W. Campbell, et D. Reynolds, 2003. Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification. Dans les actes de European Conference on Speech Communication and Technology (Eurospeech), 1345–1348. ISCA.

(Soong et al., 1985) F. Soong, A. Rosenberg, L. Rabiner, et B. Juang, 1985. A vector quantization approach to speaker recognition. Dans les actes de ICASSP, Volume 10.

(Sönmez et al., 1997) M. Sönmez, L. Heck, M. Weintraub, et E. Shriberg, 1997. A lognormal tied mixture model of pitch for prosody based speaker recognition. Dans les actes de European Conference on Speech Communication and Technology (Eurospeech). ISCA.

(Van Leeuwen et Brummer, 2007) D. van Leeuwen et N. Brummer, 2007. An introduction to application independent evaluation of speaker recognition systems. Dans les actes de Speaker Classification (1), 330–353.

(Van Vuuren 1996) Van Vuuren S., Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch, dans Proceedings of International Conference on Spoken Language Processing (ICSLP 96), pages 1788–1791, 1996.

(Wan et Campbell, 2000) V.Wan etW. M. Campbell, 2000. Support vector machines for speaker verification and identification. Dans les actes de Neural Networks for Signal Processing, Volume 2, 775–784.

(Zissman et Singer, 1994) M. A. Zissman et E. Singer, 1994. Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling. Dans les actes de IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, Volume 1.

Chapitre

3 Production et analyse numérique du signal parole

Sommaire

Résumé

Une bonne dose de théorie de la production de la parole ne ferait pas de mal. Cette introduction ne peut être complète qu'avec l'introduction des différents aspects et concepts liés au traitement numérique du signal parole tel que la modélisation par prédiction linéaire ou par banc de filtres.

3.1 Introduction

Comme nous l'avons vu dans l'introduction à la biométrie (Chapitre 1), il existe de nombreuses façons de décrire et reconnaître un individu, différents types d'informations et différentes modalités. Aux vues des contraintes inhérentes aux différentes biométries et des performances de chacune, il apparaît que l'usage de la parole présente de nombreux avantages.

La parole est un vecteur de communication naturelle aux êtres humains. Son utilisation en biométrie est relativement bien acceptée du fait de son caractère peu intrusif, et les contraintes ergonomiques imposées aux utilisateurs de systèmes vocaux sont mineures.

Les éléments nécessaires au déploiement d'interfaces matérielles (microphones, chaînes de traitement, etc.) sont de plus très répandus et peu coûteux. Comme nous le détaillerons dans la suite de cette partie, les performances des biométries liées à la parole fournissent un niveau de sécurité assez élevé.

L'usage de la parole en tant que modalité biométrique permet de tirer parti simultanément de trois informations.

- La parole résulte d'un processus complexe qui la rend porteuse d'une information sur les différents organes qui participent à sa production.
- Les signaux de parole contiennent aussi une information sur le vécu ou l'environnement de l'individu car la parole est un phénomène qui résulte d'un apprentissage.
- La parole est également vecteur de communication. Le message qu'elle transmet peut par exemple être utilisé comme mot de passe pour vérifier l'identité du locuteur.

Les informations liées directement aux organes de production de la parole présentent deux avantages majeurs : elles sont difficilement falsifiables et elles font partie intégrante de l'individu.

3.2 La production de la parole

La production de la parole fait intervenir différents organes. La source de la parole provient des poumons qui émettent un flux d'air. Ce flux d'air va traverser le larynx pour faire vibrer ou non les cordes vocales. Il va ensuite traverser le conduit vocal (cavité nasale et buccale) et les articulateurs tels que les lèvres et la langue (Figure 3.1).

La production de la parole est un processus complexe dans lequel sont impliqués de nombreux organes dont les principaux sont représentés sur la figure 3.1. Le contenu fréquentiel du signal acoustique de parole produit par un locuteur est fortement dépendant des caractéristiques morphologiques de son appareil phonatoire. Celui-ci peut être divisé en quatre parties : le générateur, le vibrateur, le résonateur et les modulateurs (Haton et al., 2006).

- ***Le générateur*** **:** c'est l'air expulsé des poumons qui est le moteur de la parole. Cet air traverse l'appareil phonatoire comme un instrument à vent et crée la pression nécessaire à la génération d'un signal acoustique.
- ***Le vibrateur*** **:** L'air expulsé des poumons traverse la trachée pour arriver dans le larynx où se trouvent les cordes vocales. Les cordes vocales sont une paire de muscles dont la longueur moyenne se situe entre 20 et 25 millimètres. Cette longueur varie cependant d'un individu à l'autre. L'air traversant le larynx met en vibration les cordes vocales. La fréquence de vibration des cordes vocales est modulée en fonction de leur degré de contraction. Le locuteur peut ainsi moduler la hauteur des sons qu'il émet.
- ***Le résonateur*** **:** Ces vibrations sont modifiées par le passage de l'air dans les différentes cavités qui composent le pharynx mais aussi dans les fosses nasales, la bouche et le larynx avec qui il communique. Ces résonateurs influent sur le son en atténuant certaines fréquences et en amplifiant d'autres. La forme et le volume de ces cavités, spécifiques au locuteur, modifient fortement le son produit.
- ***Les modulateurs*** **:** Enfin les organes modulateurs que sont la langue, les lèvres et la mâchoire sculptent le son pour produire les phonèmes qui composent la parole. La position de ces différents organes est le mécanisme final qui permet la production de parole articulée.

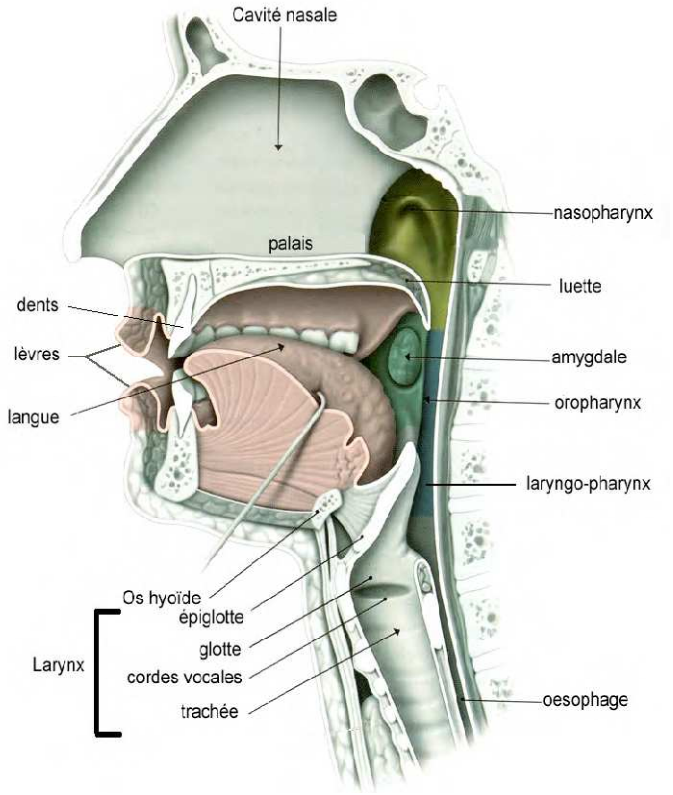


Figure 3.1 Vue schématique de l'appareil vocal[11], dans le plan sagittal médian.

[11] Illustration extraite de http://lecerveau.mcgill.ca

L'ensemble des organes agit comme un filtre, considéré comme linéaire, dont la réponse impulsionnelle comporte des fréquences de résonance caractérisées par des pics, appelés formants, dans le spectre du signal de sortie. Le signal résultant est globalement non stationnaire mais peut être considéré comme stationnaire sur de très courtes périodes, de l'ordre de 20ms (signal pseudo-stationnaire). Sur un segment de parole de cette longueur la voix est habituellement et schématiquement séparée en deux classes distinctes :

1. ***voisée*** lorsqu'il y a vibration des cordes vocales, le signal est alors quasi-périodique,

2. ***non voisée*** dans le cas d'un simple soufflement, le signal est alors considéré comme aléatoire.

Dans le premier cas, la source d'excitation est modélisée par un train d'impulsions périodique, de fréquence dite de voisement F_0, qui correspond à la fréquence de vibration des cordes vocales, la fréquence fondamentale ou pitch ;

$$\mathbf{u(n) = \sum_k \delta(n - k.P)} \qquad \mathbf{(3.1)}$$

où P représente la période de la fréquence fondamentale (en anglais, « pitch »).

Dans le second cas, la source est modélisée par un bruit blanc. Cette représentation binaire de la production de la parole a été introduite par (Fant, 1960). Elle est reprise sur la figure 3.2.

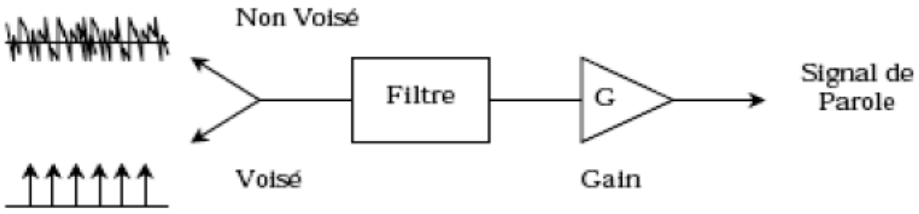


Figure 3.2 Modèle de production de la parole.

Certains organes impliqués dans la production de la parole se trouvent être en partie visibles pour un observateur extérieur (lèvres, langue, position des mâchoires). La configuration des organes modulateurs et leurs mouvements fournissent donc une information visuelle caractéristique de la parole produite (Luettin et Thacker, 1997 ; Mohamadi, 1993). Ainsi la parole est un phénomène temporel bi-modal structuré, dont les manifestations acoustiques et visuelles peuvent être observées au cours du temps. L'utilisation de la parole en biométrie rend possible l'acquisition simultanée du signal acoustique de parole et du signal vidéo correspondant. Cette acquisition est d'autant plus facile à réaliser que le matériel nécessaire (microphone et caméra) est bon marché et présent sur un grand nombre d'appareils commercialisés tels que les téléphones cellulaires, PDA, ordinateurs de bureau ou portables...

3.3 Analyse numérique du signal de parole

Les techniques d'analyse du signal de parole décrites dans ce chapitre sont désormais éprouvées et offrent une base solide aux techniques d'authentification. Ainsi la reconnaissance du locuteur a bénéficié des nombreux efforts de recherche en traitement du signal, originellement destinés au codage, en synthèse ou en reconnaissance de la parole. Les mécanismes de production ainsi que les paramètres caractéristiques du signal présentés ici en sont directement issus.

3.3.1 Du signal analogique à la représentation numérique

Les traitements effectués sur la parole sont aujourd'hui réalisés dans le domaine numérique. Au-dessus de 8kHz l'information vocale est négligeable, la bande de fréquence généralement utilisée est [0-8000Hz]. Un échantillonnage du signal de parole à 16kHz convient pour conserver la quasi-totalité de l'information (théorème d'échantillonnage de Nyquist/Shannon). L'amplitude est alors quantifiée généralement sur 16bits afin d'obtenir une bonne qualité. Pour un codage bas-débit l'échantillonnage est réalisé à 8kHz, ce qui permet de conserver la bande téléphonique (300-3400Hz). Le signal est représenté dans le domaine fréquentiel par l'utilisation des transformées de Fourier, ou encore sous une forme pouvant regrouper les informations de temps et de fréquence : le spectrogramme (Figure 3.3).

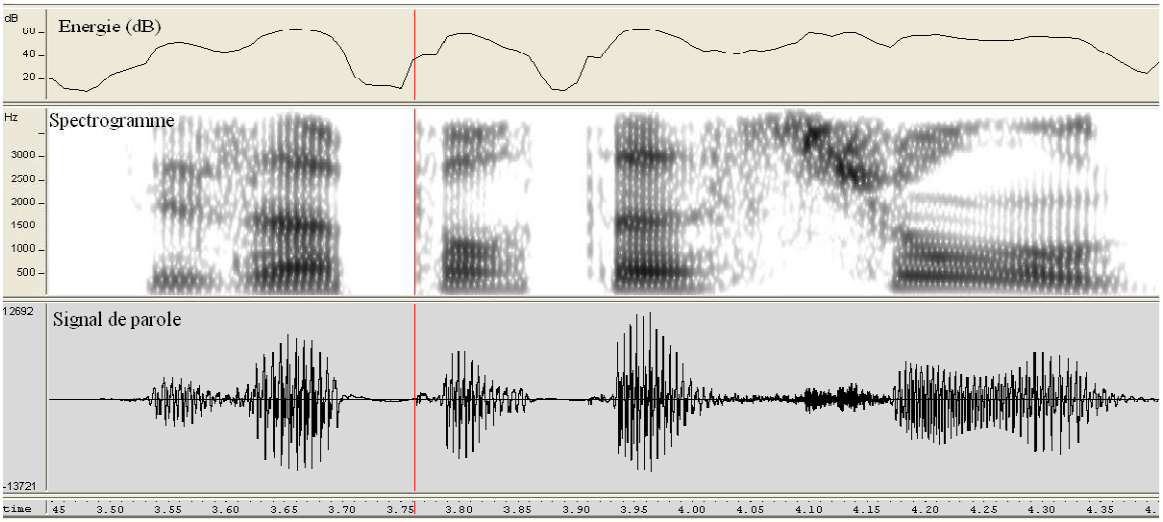


Figure 3.3 Représentation d'un signal de parole, de son spectrogramme et de son énergie.

Du fait de sa quasi-stationnarité sur de courtes périodes, le signal de parole est généralement analysé sur des trames découpées par une fenêtre de pondération de 20 à 30ms avec un taux de recouvrement de 50% à 75%, puis représenté dans le domaine spectral (Figure 3.4.b). Dans le cas d'un signal échantillonné à 8kHz, une fenêtre d'analyse de 256 points correspond à une longueur de 32ms (Figure 3.4.a). Une fenêtre classiquement utilisée est celle de Hamming.

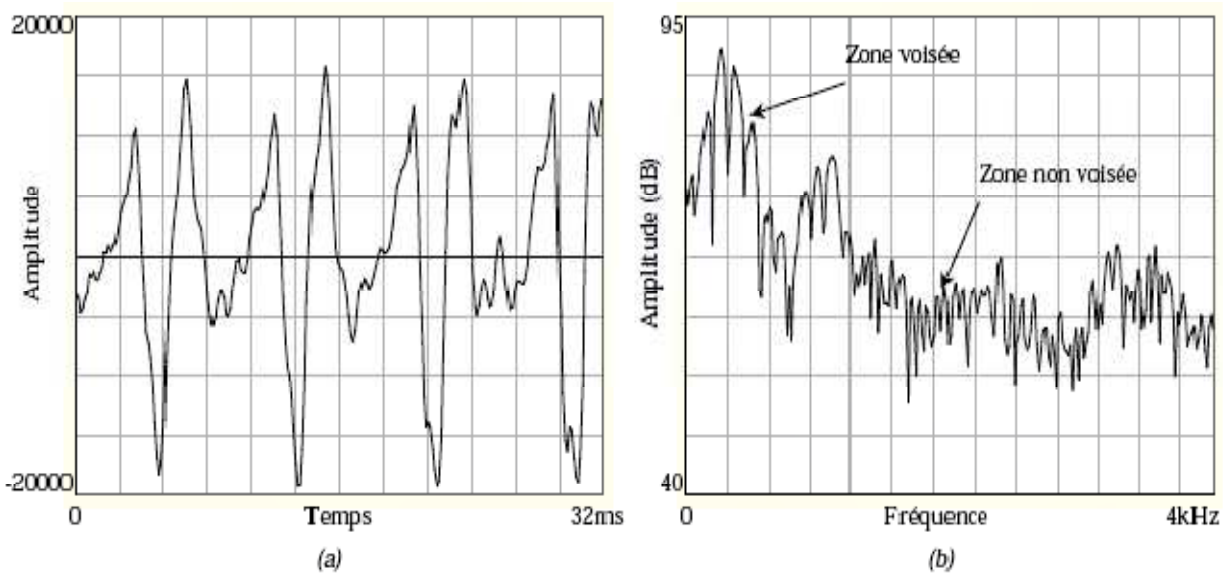


Figure 3.4 Représentation temporelle (a) et spectrale (b) d'un signal de parole voisé et non voisé.

De par son mécanisme de production, le signal de parole présente une corrélation à court terme, induite principalement par la cavité buccale, et une corrélation à long terme, qui découle directement de la structure périodique du signal. Spectralement ces caractéristiques se traduisent par une structure formantique de l'enveloppe du signal, pour la corrélation à court terme, et par une structure fine en peigne, dite harmonique, pour la corrélation à long terme. La première corrélation conduit à une dépendance des échantillons en fonction des précédents, cette propriété peut être exploitée par l'utilisation d'un filtre de prédiction linéaire. Les sons de parole sont produits par une source d'excitation G(w) qui passe à travers le filtre linéaire qui est la fonction de transfert du conduit vocal H(w).

$$\mathbf{s(t) = \int_0^t g(\tau).\, h(t-\tau).\, d\tau = g(t) * h(t)} \qquad \mathbf{(3.2)}$$

$$\mathbf{S(W) = G(W) \cdot H(W)} \qquad \mathbf{(3.3)}$$

La corrélation à long terme conduit à une périodicité dans le signal, elle est définie par la détection de la fréquence fondamentale et n'existe que dans le cas d'un son voisé. Les figures 3.4.a et 3.4.b présentent le signal temporel et le spectre de deux segments de parole, l'un voisé et l'autre non voisé. Le signal non voisé ne présente pas les mêmes caractéristiques que le signal voisé : la structure harmonique n'existe pas, l'enveloppe spectrale présente une structure formantique moins marquée. De plus, le niveau d'énergie d'un signal non voisé est généralement plus faible que pour un signal voisé.

3.4 Analyse par prédiction linéaire

L'analyse par prédiction linéaire ou analyse LPC (Linear Predictive Coding), utilise cette propriété. Le signal est alors remplacé par une source, un train d'impulsions périodiques pour les sons voisés ou bruit blanc pour les sons non voisés (Oppenheim et Schaffer, 1975; Atal, 1974). Le filtre qui représente la fonction de transfert du conduit vocal est un filtre tout pôle variant dans le temps.

$$\hat{s}(n) = \sum_{k=1}^{P} \hat{a}_k s_{n-k} \tag{3.4}$$

$$H(z) = \frac{\sigma}{A(z)} \tag{3.5}$$

où le polynôme A(Z) sera noté:

$$A(z) = 1 + \sum_{k=1}^{P} a_k . z^{-k} \tag{3.6}$$

Un réajustement périodique des paramètres tous les 10 à 30 ms est suffisant pour obtenir une bonne précision de la variation du conduit vocal. Le filtre linéaire est donc complètement défini par un facteur d'échelles (le gain) et les p coefficients regroupés dans un vecteur donné par :

$$a = [a_1, a_2, \dots, a_p]^T \tag{3.7}$$

Le filtre linéaire possède p pôles réels ou complexes conjugués. Le nombre p de paramètres nécessaires à la bonne représentation du signal de parole dépend de plusieurs paramètres. Dans les applications de codage à bas débit où la fréquence d'échantillonnage est de 8 kHz, la qualité subjective du signal synthétisé implique de choisir l'ordre du filtre environ égal à 10.

En résumé, les p coefficients a_k, le facteur d'échelle et les caractéristiques de l'excitation (telles que la période de la fréquence fondamentale et la décision de voisement) fournissent toute l'information nécessaire à la synthèse du signal de parole pendant le temps où la forme du conduit vocal peut être considérée comme stationnaire. Le signal de parole sera donc décomposé en une séquence de tranches d'échantillons s(n) , appelées segments acoustiques, dont la durée varie de 10 à 30 ms.

L'estimée du signal x est ainsi représentée par une somme pondérée des échantillons précédents x_{n-i}. Les coefficients de pondération utilisés sont les paramètres du filtre (α_i). La réponse en fréquence du filtre LPC suit les pics du spectre du signal de parole. Cette analyse est donc naturellement utilisée pour déterminer les formants. L'analyse LPC permet de représenter l'enveloppe spectrale du signal à partir des coefficients de pondérations (α_i) et les pics du spectre représentent les formants.

La figure 3.5 présente le spectre en fréquence issu de l'analyse LPC d'un signal de parole.

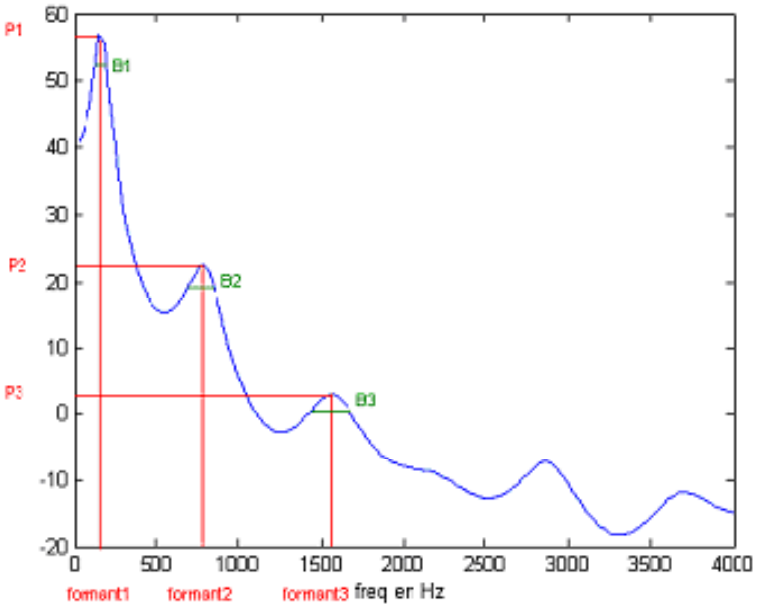


Figure 3.5 Spectre de l'analyse LPC à l'ordre 10.

3.4.1 Mise en équation du modèle LP et estimation des paramètres

Des équations (3.1), (3.5) et (3.6), la transformée du signal peut s'écrire :

$$\mathbf{S(z) = \frac{\sigma}{A(z)} . U(z)} \qquad \mathbf{(3.8)}$$

Le passage de l'excitation dans le filtre H(z) produit le signal de parole S(z). H(z) est appelé filtre de synthèse. Dans ce modèle autorégressif, encore appelé modèle LP, la sortie du signal, à l'instant n, est fournie par la récurrence suivante:

$$\mathbf{s(n) = \sum_{k=1}^{P} a_k s(n-k) + \sigma . u(n)} \qquad \mathbf{(3.9)}$$

où u(n) est le $n^{ème}$ échantillon de l'excitation.

Cette récurrence linéaire exprime qu'un échantillon quelconque s(n) est une combinaison linéaire des p échantillons qui le précèdent, à un facteur d'excitation près. C'est pourquoi ces paramètres a_k (k = 1,..., P) sont appelés coefficients de prédiction linéaire ou coefficients LPC (de l'anglais, « Linear Prediction Coefficients »).

Pour estimer les paramètres du modèle autorégressif, nous ne disposons que d'une information partielle. Elle se compose uniquement de l'observation du signal s(n) et non de son excitation. Dès lors, nous considérons que ce signal peut être prédit par la

récurrence suivante (eq. 3.4). Les coefficients $\hat{a}_k$ (k = 1, ..., P) sont les estimations des paramètres a_k du modèle autorégressif. Comme nous le verrons par la suite, la détermination des coefficients $\hat{a}_k$ est basée sur la résolution d'un système d'équations linéaires dont les différents coefficients sont dérivés de la fonction d'auto-corrélation du signal. Si le modèle est excité par un bruit blanc ou par un train périodique d'impulsions, nous pouvons démontrer que la fonction d'auto-corrélation du signal original s(n) et du signal estimé $\hat{s}(n)$ sont identiques. Il en résulte que :

$$\mathbf{a_k = \hat{a}_k\ ;\ \forall\ k = 1, 2, \dots, P} \qquad \textbf{(3.10)}$$

L'erreur commise, appelée erreur de prédiction ou signal résiduel, est donnée par :

$$\begin{aligned} \mathbf{e(n) = s(n) - \hat{s}(n)} \\ \mathbf{e(n) = \sigma . u(n)} \\ \mathbf{avec\ a_0 = 1} \end{aligned} \qquad \textbf{(3.11)}$$

Il est intéressant de montrer que cette erreur est engendrée par le passage du signal vocal s(n) dans le filtre de transmittance A(z). Ce filtre est appelé filtre inverse car sa transmittance est l'inverse de H(z) à une constante près.

L'estimation $\hat{a}_k$ est basée sur la minimalisation de la variance de l'erreur de prédiction par rapport aux paramètres a_k.

$$\mathbf{\sigma^2 = E^2 = \sum_n \big(s(n) - \hat{s}(n)\big)^2 = \sum_{i,j} a_i . a_j . \emptyset(i - j)} \qquad \textbf{(3.12)}$$

Ce critère de minimisation donne lieu à la résolution du système d'équations linéaires suivantes :

$$\mathbf{\sum_{j=1}^{P} \emptyset(i - j) . a_j = -\emptyset(i)\ ,\forall\ i = 1, 2, \dots, P} \qquad \textbf{(3.13)}$$

où la fonction d'auto-corrélation ϕ du signal est définie par :

$$\mathbf{\emptyset(i) = \emptyset(-i) = \sum_n s(n) . s(n + i)} \qquad \textbf{(3.14)}$$

L'algorithme de Levinson-Durbin permet de résoudre efficacement le système (3.14) par une récurrence sur l'ordre de prédiction. Les relations de récurrence sont les suivantes :

$$\mathbf{E(0) = \emptyset(0)}$$
$$\mathbf{k_i = \frac{\emptyset(i) + \sum_{m=1}^{i-1} a_m^{(i-1)} . \emptyset(i-m)}{E(i-1)}, i = 1, 2, \dots, P}$$
$$\mathbf{a_i^{(i)} = k_i} \qquad \textbf{(3.15)}$$
$$\mathbf{a_j^{(i)} = a_j^{(i-1)} + k_i . a_{j-1}^{(i-1)}, j = 1, 2, \dots, i-1}$$
$$\mathbf{E(i) = \left(1 - k_i^2\right) . E(i-1)}$$

Les paramètres $a_j^{(i)}$, j = 1, …, i, sont les coefficients d'un filtre du i^{ème} ordre. Ainsi, les coefficients du filtre d'ordre P sont donnés par :

$$\mathbf{a_k = a_k^{(P)}, k = 1, 2, \dots, P} \qquad \textbf{(3.16)}$$

3.4.2 Les Coefficients PARCOR

Si le signal suit la récurrence (3.15), il est immédiat de voir que le facteur d'échelle, appelé le gain du modèle, est égal à la variance optimalisée de l'erreur de prédiction. Les coefficients k_i sont appelés coefficients de réflexions ou coefficients PARCOR (de l'anglais, « PARtial CORrelation »). Dans (Boite, 1987), il est montré que le filtre inverse peut être mis sous une forme particulière dans laquelle les paramètres PARCOR sont directement utilisables. Cette structure est appelée filtre en treillis. De par cette structure, ces coefficients PARCOR jouent un rôle essentiel dans l'analyse de prédiction linéaire. On peut montrer que ces paramètres jouissent de propriétés intéressantes :

- La stabilité du modèle autorégressif est assurée si tous les coefficients PARCOR sont en module inférieurs à l'unité:

$$\mathbf{|k_i| < 1, \forall\, i = 1, 2, \dots, P} \qquad \textbf{(3.17)}$$

- L'étendue bornée des coefficients PARCOR permet une meilleure quantification.
- Quel que soit l'ordre P choisi du filtre en treillis, le calcul en cascade des coefficients PARCOR conserve la valeur des coefficients PARCOR d'ordre inférieur.

Les coefficients PARCOR et les coefficients LPC sont liés par des relations biunivoques :

PARCOR vers LPC :

$$a_i^{(i)} = k_i$$

$$a_i^{(i)} = a_i^{(i-1)} + k_i . a_{i-j}^{(i-1)}, i = 1, 2, \dots, P \; ; \; j = 1, 2, \dots, i-1 \qquad (3.18)$$

LPC vers PARCOR :

$$k_i = a_i^{(i)}$$

$$a_i^{(i-1)} = \frac{a_i^{(i)} + a_i^{(i)} . a_{i-j}^{(i)}}{1 - k_i^2}, i = 1, 2, \dots, P \; ; \; j = 1, 2, \dots, i-1 \qquad (3.19)$$

Les paramètres PARCOR sont généralement préférés aux coefficients de prédiction linéaire pour représenter l'information contenue dans le filtre autorégressif. Néanmoins, la forte sensibilité aux erreurs de quantification des paramètres PARCOR a poussé à chercher des nouvelles formes dérivées. En effet, une erreur de quantification de ces paramètres peut rapidement conduire à une instabilité du filtre de synthèse.

3.4.3 Les pôles du filtre de synthèse

Les pôles du filtre de synthèse sont les zéros de la fonction de transfert A(z). L'interprétation physique que l'on peut donner de ces paramètres constitue leur principal avantage. La présence des zéros proches du cercle unité correspond à des sommets dans le spectre à court terme du signal de parole.

Ces sommets, appelés formants, représentent les bandes spectrales les plus énergétiques. Malheureusement, ces pôles, provenant d'un supplément d'opérations arithmétiques, ne forment pas un jeu ordonné de paramètres.

3.4.4 Les paramètres transformés des coefficients PARCOR

Comme mentionné précédemment, les coefficients PARCOR bénéficient de propriétés intéressantes. De plus, ils forment un jeu ordonné de paramètres dont les valeurs codées dans la plage de variation (-1, +1) assurent la stabilité du filtre de synthèse. Leur ordonnancement peut être exploité de façon à encoder les premiers coefficients PARCOR avec plus de précision. D'autres recherches ont montré que l'application d'une transformation non-linéaire aux coefficients PARCOR fournit des nouveaux paramètres mieux appropriés à la quantification.

Les paramètres LAR (de l'anglais, « Log Area Ratio ») (Makhoul, 1975) ont été utilisés dans beaucoup d'applications de reconnaissance. Ces paramètres sont définis par la relation suivante:

$$LAR_i = \log_{10} \frac{1+k_i}{1-k_i}, 1 \leq i \leq P \qquad (3.20)$$

3.4.5 Les paramètres LSP (de l'anglais, « Line Spectrum Pairs »)

L'analyse prédictive linéaire conduit à une modélisation du signal de parole par la sortie d'un filtre tous-pôles H(z) dont la forme mathématique est rappelée ci-après :

$$\mathbf{H(z) = \frac{1}{A(z)} = \frac{1}{1+\sum_i^P a_i.z^{-i}}} \quad \textbf{(3.21)}$$

où a_i (i=1,...,P) sont les coefficients de prédiction.

Nous avons vu que le système PARCOR fournit une représentation équivalente:

$$\mathbf{A_i(z) = A_{i-1}(z) + k_i.z^{-i}.A_{i-1}(z^{-1}), i = 1, 2, \dots, P} \quad \textbf{(3.22)}$$

$$\mathbf{A_0(z) = 1}$$

où k_i (i =1, ..., P) sont les coefficients PARCOR

Les paires de raies spectrales proviennent de la décomposition du polynôme $A_P(z)$ en deux polynômes dont l'un est symétrique et l'autre antisymétrique. Pour ce faire, nous prenons la somme et la différence entre $A_P(z)$ et son polynôme conjugué. k_{P+1} prend respectivement la valeur +1 et -1.

$$\mathbf{P(z) = A_p(z) + z^{-(p+1)}.A_p(z^{-1}), (k_{p+1} = +1)} \quad \textbf{(3.23)}$$

$$\mathbf{Q(z) = A_p(z) + z^{-(p+1)}.A_p(z^{-1}), (k_{p+1} = -1)} \quad \textbf{(3.24)}$$

Les paires de raies spectrales sont les arguments des zéros (paires complexes conjugués) des polynômes P(z) et Q(z).

Ils ont les propriétés suivantes:

- tous les zéros de P(z) et Q(z) se trouvent sur le cercle unité ;
- les zéros de P(z) et Q(z) sont alternés sur ce même cercle unité. Ils forment donc un jeu de paramètres ordonnés ($0 \leq w_1 < w_2 < w_3 < \dots < w_P \leq \pi$) ;
- Les paires de raies spectrales sont regroupées dans un vecteur défini comme suit :

$$\mathbf{[w_1, w_2, \dots, w_p]^T (en\ radians)} \quad \textbf{(3.25)}$$

Ce vecteur caractérise bien le modèle autorégressif vu qu'il existe toujours une relation biunivoque entre le vecteur LSP et le vecteur des coefficients de prédiction.

$$\mathbf{A_p(z) = \frac{P(z)+Q(z)}{2}} \quad \textbf{(3.26)}$$

La comparaison terme à terme entre ces deux polynômes dans (3.26) permet de déduire cette relation bi-univoque. Ils bénéficient d'une interprétation physique liée aux fréquences formantiques du signal. En effet, l'étude du carré de la fonction de transfert H(Z) en fonction des paramètres LSP montre que le gain de la fonction de transfert est d'autant plus grand que deux composantes successives du vecteur LSP sont proches. Enfin, l'étude de l'influence de l'enveloppe spectrale par rapport à la variation d'un des paramètres LSP montre une plus faible sensibilité. La complexité de calcul des paramètres LSP est peut-être le seul désavantage.

3.4.6 Les coefficients cepstraux et leurs formes dérivées

Les coefficients cepstraux (en anglais, « Cesptral coefficients ») proviennent de l'analyse homomorphique du signal de parole. Cette technique de traitement non-linéaire du signal est appropriée au signal de parole. En effet, la parole est la convolution temporelle de la réponse impulsionnelle du conduit vocal et de la fonction d'excitation. Cette convolution devient une multiplication dans le domaine fréquentiel. Si le logarithme du spectre est considéré, cette multiplication devient une addition. Etant donné que l'oreille humaine est pratiquement insensible à la phase relative entre deux composantes sonores, on peut seulement se limiter au module de la transformée. On a donc :

$$\log\left(\left|S(e^{jw})\right|\right) = \log\left(\left|H(e^{jw})\right|\right) + \log\left(\left|U(e^{jw})\right|\right) \qquad (3.27)$$

où $S(e^{jw})$ est le spectre de la parole, $H(e^{jw})$ est le spectre du conduit vocal et $U(e^{jw})$ est le spectre de l'excitation.

La transformée de Fourier inverse du logarithme de la norme du spectre du signal de parole définit le cepstre réel (Bogert, 1963). Ce cepstre se compose des coefficients, appelés coefficients cepstraux. La référence (Furui, 1989) montre qu'il existe des relations récursives entre ces coefficients et les coefficients de prédiction :

$$\begin{aligned} c_0 &= E(0) = \emptyset(0) \\ c_1 &= -a_1 \\ c_i &= -a_i - \sum_{k=1}^{i-1} \frac{i-k}{i} . c_{i-k} . a_k \,, \quad 1 < i \leq P \\ c_i &= -\sum_{k=1}^{P} \frac{i-k}{k} . c_{i-k} . a_k \,, \qquad i > P \end{aligned} \qquad (3.28)$$

Le premier coefficient, c_0, représente l'énergie de la tranche de parole analysée. Ces coefficients cepstraux présentent plusieurs propriétés. Ils permettent de calculer la distorsion spectrale entre deux filtres autorégressifs par une simple distance euclidienne. De plus, les coefficients CMS (de l'anglais, « Cesptral Mean Substraction ») dérivés des coefficients cepstraux sont insensibles aux distorsions linéaires engendrées par le canal de transmission ou par le microphone (Mammone, 1996).

Les coefficients CMS sont les coefficients cepstraux centrés sur leurs moyennes respectives :

$$c_i^{cms} = c_i - E(c_i) \quad (3.29)$$

Les coefficients cepstraux, qu'ils soient calculés à partir de la transformée de Fourier du signal ou de l'analyse LPC, représentent l'influence du conduit vocal et de la source sur le signal de parole émis. Le nombre de coefficients cepstraux calculés détermine le niveau de lissage de l'estimation de l'enveloppe spectrale. Les coefficients cepstraux d'ordre faible sont très utilisés en RAL. Ils caractérisent un trait anatomique de l'individu, principalement le conduit vocal. Une autre propriété intéressante des coefficients cepstraux est leur faible corrélation, introduite par la transformation DCT. Enfin, le calcul des coefficients cepstraux est indépendant de l'énergie du signal d'entrée, ce qui permet de réduire en conséquence la variabilité du signal.

3.5 Paramètres issus de l'analyse par banc de filtres

L'analyse par banc de filtres est une technique qui a été initialement utilisée pour le codage du signal de parole (vocodeur à canaux). Dans ce cadre elle fait partie des techniques de codage par analyse-synthèse à partir du modèle source/filtre. Elle consiste à filtrer le signal par un ensemble de filtres passe-bande. L'énergie en sortie de chaque filtre est attribuée à sa fréquence centrale. Pour simuler le fonctionnement du système auditif humain, les fréquences centrales sont réparties uniformément sur une échelle perceptive. Plus la fréquence centrale du filtre est élevée, plus sa bande passante est large. Cela permet d'augmenter la résolution dans les basses fréquences, zone qui contient le plus d'information utile dans le signal de parole. Les échelles perceptives les plus utilisées sont le Mel ou le Bark. Du point de vue des performances des systèmes de reconnaissance, ces deux échelles sont quasiment identiques.

3.5.1 Coefficients MFCC (en anglais « Mel Frequency Cepstrum Coefficients)

Le traitement décrit dans le paragraphe précédent permet d'obtenir une estimation de l'enveloppe spectrale (densité spectrale lissée). Il est possible d'utiliser les sorties des bancs de filtres comme entrées dans le système de reconnaissance. Cependant, d'autres coefficients dérivés des sorties d'un banc de filtres, sont plus discriminant, plus robustes au bruit ambiant et moins corrélés entre eux. Il s'agit des coefficients cepstraux dérivés des sorties du banc de filtres répartis linéairement sur l'échelle Mel, ce sont les paramètres MFCC (Davis, 1980). Les étapes d'une analyse MFCC sont présentées dans la figure 3.6.

Une autre modification consiste à utiliser des paramètres variationnels (dits Δ-cepstraux) qui caractérisent les variations des paramètres cepstraux dans les fenêtres proches de la fenêtre à court-terme courante (Furui, 1981a; Soong, 1988). Ces paramètres variationnels sont nettement moins efficaces que les paramètres cepstraux

dans le cadre de la reconnaissance du locuteur lorsqu'ils sont utilisés seuls. Par contre, utilisés avec les paramètres cepstraux "instantanés" ils conduisent à une amélioration substantielle des performances. Les paramètres variationnels présentent en outre l'intérêt d'être insensibles à des variations linéaires du canal de transmission entre deux enregistrements. Toutefois, les paramètres variationnels ne sont réellement significatifs que dans les cas où il est possible de les comparer dans un même contexte, c'est à dire dans les applications ou le texte prononcé par chaque locuteur est fixé (Soong, 1988). En mode indépendant du texte l'utilisation de paramètres, Δ-cepstraux, en plus des paramètres cepstraux, ne semble pas améliorer sensiblement les performances de reconnaissance (et ce même dans les cas où le vocabulaire est fortement contraint comme dans (Tseng 1992)).

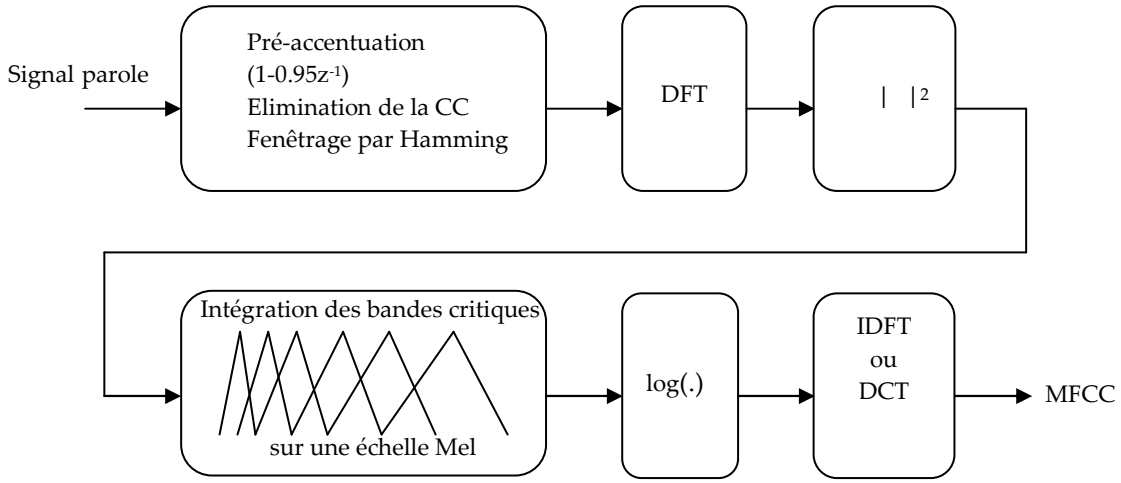


Figure 3.6 Calcul des coefficients MFCC.

Un autre point important est le fait que tous les coefficients cepstraux ne présentent pas des variations intra-locuteur de même ampleur : la variance des coefficients cepstraux décroît avec leur ordre. Pour tenir compte de cette propriété, on utilise généralement la distance dite de Mahalanobis, qui se réduit ici à une distance euclidienne pondérée du fait de la décorrélation des coefficients cepstraux. Cette distance est beaucoup plus efficace dans le cadre de la reconnaissance du locuteur car elle permet de réduire l'influence des coefficients qui présentent une forte variabilité intra-locuteur (Soong, 1988). Compte tenu des propriétés du signal de parole, les poids à appliquer aux coefficients cepstraux pour le calcul de la distance pondérée peuvent être approximés par une progression linéaire en fonction de l'ordre des coefficients cepstraux (on obtient alors la distance dite root powers sum) (Soong, 1988).

3.5.2 Coefficients PLP (en anglais « Perceptual Linear Predictive »)

La technique conventionnelle de la prédiction linéaire (LPC) est fondée sur un modèle tout-pôles autorégressif qu'on estime en utilisant une analyse par prédiction linéaire. Ce modèle constitue une approximation de l'enveloppe de la densité spectrale

à court terme. Cette analyse est en accord avec les aspects de la production de la parole qu'avec les aspects de la perception de la parole. L'analyse PLP (Hermansky, 1990) est une alternative à l'analyse LPC qui tient compte de trois aspects de la perception (Figure 3.7).

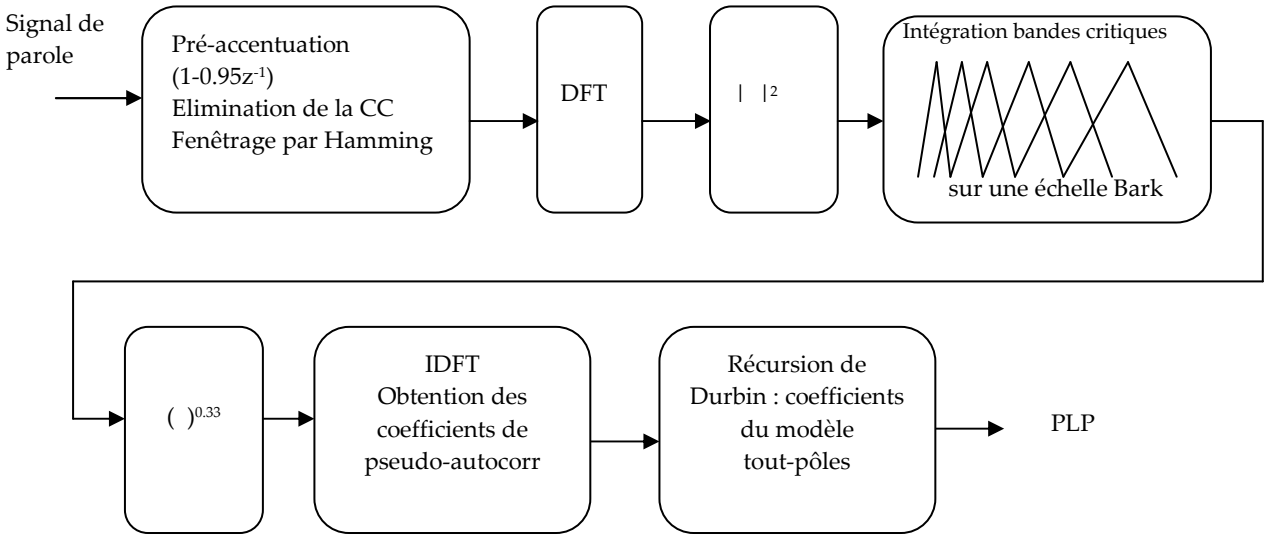


Figure 3.7 Calcul des coefficients PLP.

- **Intégration des bandes critiques :** l'analyse LPC donne la même approximation de l'enveloppe de la densité spectrale pour toute la zone de fréquences utiles, ceci n'est pas en accord avec le fonctionnement de l'oreille où chaque fréquence porte une force sonore résultant de l'intégration de l'information sur une zone de fréquences appelée bande critique. Les bandes critiques sont réparties suivant l'échelle fréquentielle de Bark sur toute la zone de fréquence utilisée. Un Bark correspond à une augmentation de fréquence de quantité égale à une bande critique. Pour simuler ce fonctionnement dans le cadre de l'analyse PLP, la densité spectrale est reportée dans l'échelle de Bark, puis convoluée avec une fonction qui simule une bande critique. Cette fonction est la même pour toute les fréquences dans l'échelle de Bark (à une translation près). Ce qui réduit largement la résolution, surtout dans les hautes fréquences dans l'échelle Herz. Généralement la nouvelle densité spectrale est échantillonnée à chaque intervalle de 1 Bark.

- ***Préaccentuation par courbe d'isosonie :*** des expériences de psychoacoustique ont montré que l'oreille possède des caractéristiques non linéaires. En effet, des expériences réalisées par (Fletcher, 1933) ont mis en évidence que l'intensité perçue, lorsqu'on écoute un son pur d'intensité acoustique constante, varie avec la fréquence de ce son pur. Pour simplifier ce phénomène dans le cadre de l'analyse PLP, on multiplie la densité spectrale résultante de l'étape précédente par une fonction de pondération. Il

est possible d'estimer cette fonction en reportant sur une abaque des lignes isosoniques le long desquelles un son pur donne à l'oreille une égale sensation d'intensité. Ceci est à l'origine de ce qu'on appelle la sonie considérée comme l'intensité subjective des sons. Dans nos expériences, nous remplaçons cette opération par l'opération de préaccentuation habituelle, c'est-à-dire l'application dans le domaine temporel du filtre passe-haut $(1–0,95.z^{-1})$.

- ***Loi de Stevens :*** les deux traitements précédents ne sont pas suffisants pour avoir une correspondance entre l'intensité mesurée et l'intensité subjective (la sonie). La loi de Stevens (Stevens, 1957) affirme qu'après avoir réalisé l'intégration des bandes critiques et la préaccentuation, la relation entre l'intensité et la sonie devient :

$$\textbf{Sonnie} = (\textbf{intensité})^{0.33} \qquad \textbf{(3.30)}$$

L'oreille humaine est également plus sensible aux variations relatives de la valeur d'un signal acoustique qu'à ses variations absolues. L'intégration de cette nouvelle propriété dans la technique PLP a conduit à l'élaboration des paramètres RASTA-PLP (en anglais « RelAtive SpecTrAl ») (Hermansky, 1991). Elle consiste à un filtrage passe-haut des sorties d'un banc de filtres dans le domaine du logarithme du spectre afin de supprimer les variations lentes du signal, puis à appliquer un opérateur exponentiel pour retourner dans le domaine du spectre de puissance. Ces paramètres bénéficient de très bonnes propriétés de robustesse qui permettent de compenser des variations de microphone, mais ne luttent pas contre les perturbations provoquées par un bruit. Pour cela, (Hermansky, 1993) a développé un filtrage passe-bande d'une fonction du spectre, égale à log(1+Jx), approximativement logarithmique lorsque l'amplitude x du spectre est élevée, et approximativement linéaire lorsque l'amplitude du spectre est faible. Cette méthode appelée J-RASTA, permet de compenser principalement les effets du bruit lorsque x est faible, et les effets d'un filtrage linéaire lorsque x est élevé.

3.6 Paramètres prosodiques

Le terme « paramètres prosodiques » réunit l'énergie, la durée et la fréquence fondamentale (ou pitch) (Atal, 1972). Ces paramètres s'avèrent cependant fragiles en pratique et ne permettent pas, à eux seuls, de discriminer les locuteurs. En conséquence, ils sont souvent associés aux paramètres de l'analyse spectrale.

3.6.1 L'énergie

L'énergie du signal est un indice qui peut par exemple contribuer à la détection du voisement d'un segment de parole. L'énergie totale E_0 est calculée directement dans le domaine temporel sur une trame de signal s(n) avec n entre 0 et N-1comme :

$$\mathbf{E_0 = \frac{1}{N}\sum_{n=0}^{N-1}(s(n))^2} \qquad \mathbf{(3.31)}$$

L'énergie ainsi obtenue est sensible au niveau d'enregistrement; on choisit en général de la normaliser, et d'exprimer sa valeur en décibels par rapport à un niveau de référence.

3.6.2 Fréquence fondamentale (ou pitch)

Le pitch est un paramètre très important pour l'étude acoustique et la synthèse de la parole, l'oreille est très sensible à ses variations lesquelles constituent la prosodie, l'évolution de la fréquence en fonction du temps au niveau du phonème constitue la micromélodie, par contre son évolution au niveau des groupes syntaxiques de la phrase est la macromélodie, l'intonation du message est directement liée au pitch. Plusieurs techniques en vue de l'extraction du fondamental peuvent être employées parmi celles-ci on peut citer la méthode d'AMDF (en anglais « Autocorrelation Modified Difference Function ») qui est donnée par :

$$\mathbf{AMDF(k) = \frac{1}{N}\sum_{n=0}^{N-1}|a(n) - a(n-k)|} \qquad \mathbf{(3.32)}$$

cette fonction présente un minimum aux multiples de la période fondamentale.

D'une manière générale, les méthodes d'estimation de pitch comportent trois étapes :

- Le prétraitement du signal parole pour l'adaptation du signal (filtrage, fenêtrage, pré-accentuation, …),
- Le traitement pour l'extraction de la fréquence fondamentale,
- Le post-traitement pour corriger les erreurs de calcul surtout pour les transitions voisées/non-voisées.

3.6.3 D'autres paramètres

Une palette de paramètres a été définie et utilisée dans le domaine de la reconnaissance du locuteur tel que le taux de voisement, le taux de passage par zéro, les durées des segments, les pauses, …etc. Le but de ce chapitre n'était pas de donner toute les paramétrisations disponible dans la littérature mais plutôt une introduction à ce domaine avec en citation quelques représentations des plus répandues.

3.7 Conclusion

Ce chapitre clore la première partie de cette thèse qui a été consacrée principalement à fournir au lecteur : 1) une bonne entrée en matière dans le domaine des technologies de la biométrie pour le guider petit à petit à la biométrie vocale ou reconnaissance automatique du locuteur, en lui donnant les notions de base, les concepts fondamentaux, les différentes tâches, les difficultés et les outils d'évaluation, et 2) les concepts théoriques de la production de la parole et les outils mathématiques nécessaires pour son analyse et son traitement numérique. Cette première partie permet au lecteur de bien s'armer pour attaquer le chapitre suivant consacré à l'état de l'art du module de paramétrisation, appelé front end dans la littérature, dont tout système RAL est fortement dépendant. Ce front-end ne peut être étudié sans la maitrise des phénomènes de la production et la perception de la parole et des outils mathématiques pour les modélisés.

3.8 Bibliographie

(Atal, 1972) B. S. Atal. Automatic speaker recognition based on pitch contours. The Journal of the acoustical society of America, no. 52, pp. 1687-1697, 1972.

(Atal, 1974) B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. JASA, vol. 55, pp. 1304-1312, Jun. 1974.

(Bogert, 1963) B. Bogert & al. The quefrency analysis of time series for echoes. Proceedings of Sumposium on Time Series Analysis, 1963.

(Boite, 1987) R. Boite & M. Kunt. Traitement de la parole. Presses Polytechniques Romandes, 1987.

(Davis, 1980) S. P. Davis & P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, pp. 357-366, Aug. 1980.

(Fant, 1960) G. Fant. Acoustic theory of speech production. s-Gavenhage, Mouton, 1960.

(Fletcher, 1933) H. F. Fletcher & W. A. Munson. Loudness, its definition, measurement and calculation. Journal of the Acoustical Society of America, vol. 5, 1933.

(Furui, 1989) S. Furui. Digital speech processing, synthesis and recognition. Marcel Dekker, N-Y, 1989.

(Furui, 1981) S. Furui, 1981. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing] 29(2), 254–272. 50, 51, 52

(Haton et al., 2006) J. P. Haton, C. Cerisara, D. Fohr, Y. Laprie, et K. Smaili, 2006. Reconnaissance automatique de la parole, du signal ‡ son inteprÈtation. Dunod. 44

(Hermansky, 1993) H. Hermansky & al. Recognition of speech in additive and convolutional noise based on RASTA spectral processing. Proceedings of the IEEE, ICASSP, vol. 2, pp. 83-86, 1993.

(Hermansky, 1991) H. Hermansky & al. Compensation for the effect of the communication channel in auditory-like analysis of speech (Rasta-PLP). Proceedings of EUROSPEECH, pp. 1367-1370, 1991.

(Hermansky, 1990) H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. Journal of Acoustical Society of America, vol. 4, no. 87, 1990.

(Luettin et Thacker, 1997) J. Luettin et N. A. Thacker, Speechreading using probabilistic models. Computer Vision and Image Understanding 65(2), 163–178, 1997.

(Makhoul, 1975) J. Makhoul. Linear prediction : a tutorial review. Proceedings of the IEEE, vol. 63, pp. 561-580, 1975.

(Mammone, 1996) R. J. Mammone & al. Robust speaker recognition. IEEE Signal Processing Magazine, pp. 58-71, Sept. 1996.

(Mohamadi, 1993) T. MOHAMADI , Synthèse à partir du texte de visages parlants : réalisation d'un prototype et mesures d'intelligibilité bimodale, Thèse 3ème cycle, INPG, 1993.

(Oppenheim et Schaffer, 1975) Digital Signal Processing. Englewood Cliffs, New Jersey: Prentice-Hall, 1975.

(Soong, 1988) F. K. Soong & A. E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 36, no. 6, pp. 871-879, 1988.

(Stevens, 1957) S. S. Stevens. On the psychological law. Psychological Review 64, 1957.

(Tseng 1992) B. L. Tseng & al. Continuous probabilistic acoustic map for speaker recognition. Proceedings of the IEEE, ICASSP, vol. 2, pp. 161-164, 1992.

Chapitre

4 Extraction des traits du locuteur: Etat de l'art et tendances actuelles

Sommaire

Résumé

Ce chapitre donne une revue de la technologie de reconnaissance automatique de locuteur. Nous donnons une revue des méthodes classiques et de l'état de l'art. Nous commençons par les notions fondamentales de la reconnaissance du locuteur, surtout concernant l'extraction des paramètres. Nous élaborons les techniques avancées de calcul pour adresser la robustesse et la variation inter-session, ainsi que les progrès récents qui ouvre une nouvelle ère d'exploration.

4.1 Introduction

Avant de faire une revue sur les travaux à propos de la paramétrisation, on propose de rappeler la définition des paramètres variant, les paramètres à temps-invariant (indépendant du temps) et les paramètres prosodiques définis dans plusieurs domaines par la fréquence fondamentale (vibration des cordes vocales), l'intensité de la voix (ou énergie) et la durée successive des segments syllabiques.

Les paramètres à temps-variant sont estimés sur une fenêtre d'analyse de courte durée alors que les paramètres à temps-invariant sont ceux estimés sur une fenêtre d'analyse de longue durée. Ces derniers correspondent à des caractéristiques fixes et stables dans la parole. Ils sont considérés indépendants du contexte et sont en général utilisés dans les systèmes de reconnaissance indépendants du texte (Furui et al, 1972; Atal, 1976). Pourtant, ils ne tiennent pas compte de la variabilité intra-locuteur ce qu'il les rends facile à imiter par les imposteurs. Les paramètres à temps variants sont dépendants du texte et sont souvent utilisés avec un grand succès dans les systèmes de vérification du locuteur et dans les systèmes dépendants du texte.

Les paramètres prosodiques quant à eux prennent une importance particulière pour donner aux systèmes de synthèse une meilleure intelligibilité tout en permettant aux systèmes de reconnaissance d'effectuer une analyse ou segmentation par ordre d'unité phonétique. La variation dans le temps de ces paramètres (intonation) véhicule divers indices caractéristiques de l'individu que ce soit au niveau de son état physique (âge, sexe, physiologie), de son état émotionnel ou de son accent régional. En dépit de l'importance de la prosodie, on trouve que les systèmes en traitement automatique de la parole (reconnaissance de parole ou identification/vérification du locuteur) se basent particulièrement sur les paramètres en codage destinés à caractériser la contribution et la dynamique du conduit vocal. Cependant, les paramètres prosodiques ne sont utilisés en général que pour faire rehausser légèrement les performances de ces systèmes.

4.2 Classification des caractéristiques

Le nombre de caractéristiques doit être relativement faible. Les modèles statistiques traditionnels tel que le modèle de mélange de gaussienne (Reynolds et al., 2000 ; Reynolds et Rose, 1995) ne peut pas géré des données de grande dimension. Le nombre des échantillons d'apprentissage pour l'estimation d'une densité fiable croit exponentiellement avec le nombre des caractéristiques. Ce problème est connu sous le nom de la malédiction de la dimensionnalité (Jain et al., 2000). La réduction et la préservation des calculs sont évidentes avec des caractéristiques de faible dimension.

Il y'a différentes façons de catégoriser les caractéristiques du locuteur (Figure 4.1). D'un point de vue de leur interprétation physique, nous pouvons les diviser en:

- Caractéristiques spectrales à court-terme ;
- Caractéristiques de la source vocale ;
- Caractéristiques spectro-temporelles ;
- Caractéristiques prosodiques ;
- Caractéristiques de haut-niveau

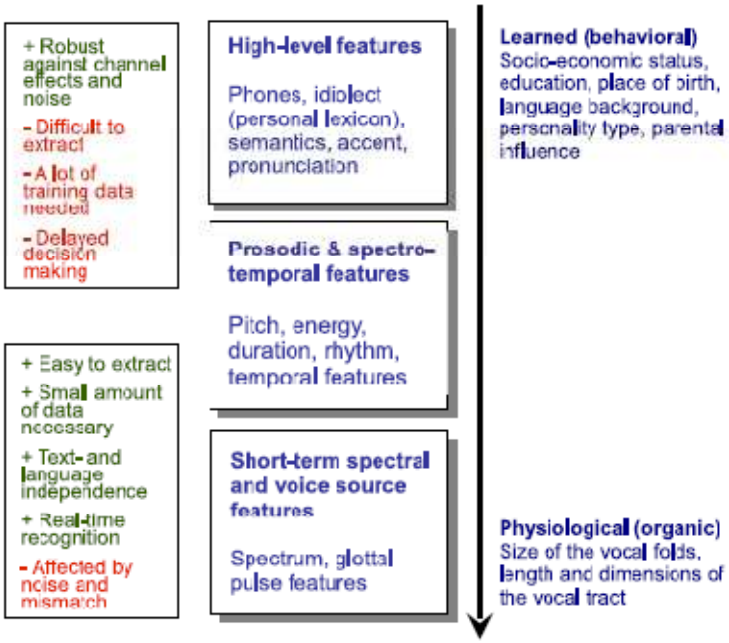


Figure 4.1 Résumé des caractéristiques du point de vue de leur interprétation physique.

Les caractéristiques spectrales à court-terme, comme son nom l'indique, sont calculées à partir des trames courtes d'une durée d'environ 20-30ms. Elles sont habituellement des descripteurs de l'enveloppe spectrale à court-terme qui est correlée avec le timbre, c'est-à-dire « la couleur » du son ainsi qu'avec les propriétés de résonance du conduit vocal supralaryngal. Les caractéristiques de la source vocale, à son tour, caractérisent la source de la voix (le débit glottique). Les caractéristiques spectro-temporelles et prosodiques s'étendent sur des dizaines ou des centaines de millisecondes incluant l'intonation et le rythme par exemple. Enfin, les caractéristiques de haut-niveau tentent de saisir les caractéristiques du locuteur au niveau conversationnel, telle que l'utilisation caractéristiques de mots (« uh-huh », « you know », « oh yeah », etc.) (Doddington, 2001). Quelles sont les caractéristiques à utiliser ? Cela dépend de l'application envisagée, des ressources de calcul, de la quantité des données vocales disponibles (pour le développement et d'utilisation en temps réel) et si les locuteurs sont coopératifs ou non.

Pour ceux qui débutent leur recherche dans la reconnaissance de locuteur, (Reynolds et al., 2003) recommandent de commencer avec les caractéristiques spectrales à court-terme car elles sont faciles à calculer et donnent de bonnes performances. Les caractéristiques prosodiques et les caractéristiques de haut-niveau sont censées être plus robuste, mais moins discriminantes et facile à imiter; par exemple, il est relativement bien connu que les imitateurs professionnels ont tendance à modifier le contour global du pitch du locuteur imité (Kitamura, 2008 ; Ashour et Gath, 1999). Les caractéristiques de haut-niveau ont également besoin d'un front-end plus complexe, nécessitant par exemple un système de reconnaissance de la parole. Pour conclure, globalement il n'existe pas de « best caractéristiques » mais le choix et un compromis entre la discrimination des locuteurs, la robustesse et la faisabilité.

4.2.1 Les caractéristiques spectrales court-terme

Le signal de parole présente des changements continus qui sont dus aux mouvements articulatoires, et par conséquent, le signal doit être décomposé en trames courtes d'une durée de 20-30ms. Dans cet intervalle, le signal est considéré comme stationnaire et un vecteur de caractéristiques est extrait de chaque trame.

Habituellement, la trame est pré-accentuée et multipliée par une fonction fenêtre avant de procéder à de nouveaux traitements. La pré-accentuation rehausse les fréquences hautes dont l'intensité serait par ailleurs très faible en raison du spectre déscendant causé par la souce vocale glottique (Harrington et Cassidy, 1999). La fonction de fenêtre, généralement celle de Hamming, est nécessaire en raison des effets de longueur finie de la transformée de fourrier (DFT) ; pour plus de détail, se reporter aux travaux de (Deller et al., 2000 ; Oppenheim et al., 1999 ; Harris, 1978). Dans la pratique, le choix de la fenêtre n'est pas critique. Bien que la longueur de la trame est généralement fixe, l'analyse synchrone du pitch a également été étudiée (Gong et al. 2008 ; Zilca et al., 2006 ; Nakasone et al., 2004). Les expériences dans (Zilca et al., 2006 ; Nakasone et al., 2004) indiquent que le taux de reconnaissance baisse avec cette technique, alors que (Gong et al., 2008) ont obtenu des améliorations en milieu bruité. Les modèles de locuteur dépendant du pitch ont également été étudiés (Arcienega et al., 2001; Ezzaidi et al., 2001).

La bien connue transformée de Fourrier rapide (FFT), une mise en œuvre rapide de la DFT, décompose un signal en ses composantes fréquentielles (Oppenheim et al, 1999). Une alternative à la décomposition du signal basée sur la FFT, tels que des bases non-harmoniques, les fonctions apériodiques et les bases guidées par les données issues de l'analyse en composantes indépendantes (ICA) ont été étudiées dans la littérature (Jang et al., 2002 ; Gopalan et al., 1999 ; Imperl et al., 1997). Cependant, la DFT est toujours utilisée dans la pratique en raison de sa simplicité et de son efficacité. Habituellement, seulement le spectre d'amplitude est retenu, basé sur l'idée que la phase a peu d'importance perceptuelle. Toutefois, (Paliwal et Alsteris, 2003) ont prouvé le contraire. Quant à (Hedge et al., 2004), ils décrivent une technique qui utilise l'information de la phase.

La forme globale du spectre d'amplitude de la DFT (Figure 4.2), connue sous le nom d'enveloppe spectrale, contient des informations sur les propriétés de résonance du conduit vocal et est considérée comme la partie la plus instructive du spectre dans la reconnaissance du locuteur. Un modèle simple de l'enveloppe spectrale utilise un ensemble de filtres passe-bande pour réaliser l'intégration de l'énergie dans les bandes de fréquences voisines. Motivé par des études psycho-acoustiques, la gamme de fréquence basse est généralement représentée avec une meilleure résolution en allouant d'avantages de filtres avec des bandes passantes étroites (Harrington et Cassidy, 1999).

Bien que les valeurs des énergies des sous-bandes ont été utilisées directement comme caractéristiques (Damper et Higgins, 2003 ; Sivakumaran et al., 2003 ; Besacier et Bonastre, 2000 ;), habituellement la dimension est encore réduite utilisant d'autres transformations. Les coefficients dites Mel-Frequency Cepstral Coefficient (MFCC) (Davis et Mermelstein, 1980) sont les caractéristiques les plus populaires dans le traitement de la parole et du son. Ils ont été introduits au début des années 1980 pour

la reconnaissance vocale, puis adopté en reconnaissance du locuteur. Même si les caractéristiques de diverses variantes, telles que les centroides des sous-bandes spectrales (SSC) (Kinnunen et al., 2007 ; Thian et al., 2004) ont été étudié, les MFCC semblent être difficiles à battre dans la pratique.

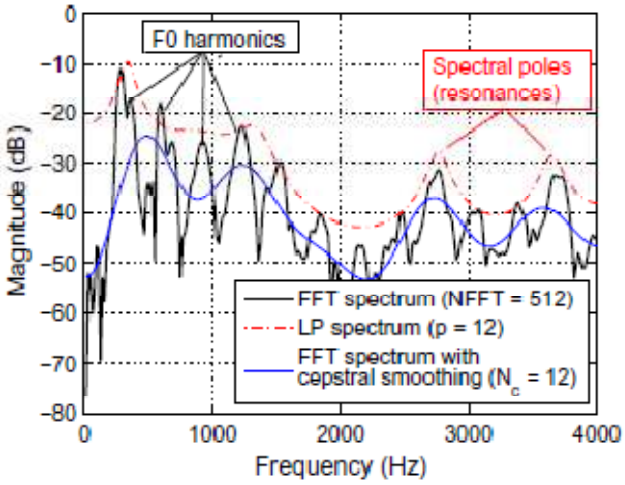


Figure 4.2 Enveloppe spectrale issue d'une analyse LP.

Les MFCC sont calculé à l'aide des bancs de filtre motivés par les études psycoacoustiques, suivie par une compression logarithmique et d'une transformation cosine discrete (DCT). Notons les sorties de M banc de filtres Y(m), m = 1, 2, … , M, les MFCC sont obtenus comme suit :

$$\mathbf{c_n = -\sum_{m=1}^{M}[\log Y(m)].\cos\left[\frac{\pi n}{M}\left(m-\frac{1}{2}\right)\right]} \qquad \mathbf{(4.1)}$$

n est l'index du coefficient cepstral.

Le vecteur MFCC final est obtenu en retenant entre 12 et 15 coefficients DCT d'ordre faible. Plus de détails concernant les MFCCs peuvent être trouvés dans (Huang et al., 2001 ; Deller et al., 2000). Des caractéristiques alternatives qui accentuent les spécificités du locuteur ont été étudiées dans (Charbouillet et al., 2006 ; Miyajima et al., 2001 ; Orman et Arslam, 2001). Pour l'étude de l'information discriminante dans le spectre du locuteur, se référer à (Lu et Dang, 2007). Enfin, certaines nouvelles tendances en matière d'extraction des paramètres peuvent être trouvées dans (Ambikairajah, 2007).

La prédiction linéaire (LP) (Mammone et al., 1996 ; Makhoul, 1975) est une méthode alternative pour l'estimation du spectre de DFT qui a une bonne interprétation intuitive à la fois dans le domaine temporel (échantillons adjacents sont corrélés) et le domaine fréquentiel (tous les pôles du spectre correspondent à la structure de résonance). Dans le domaine temporel, l'équation de prédiction LP est définie comme étant :

$$\hat{s}(n) = -\sum_{k=1}^{P} a_k s(n-k) \quad (4.2)$$

$\hat{s}(n)$ est le signal observé, a_k sont les coefficients de prédiction et s(n) est le signal prédit. Le signal d'erreur de prédiction, ou résiduel, est défini comme ($e(n)=\hat{s}(n)-s(n)$), est illustré dans le panneau du milieu de la figure 4.3.

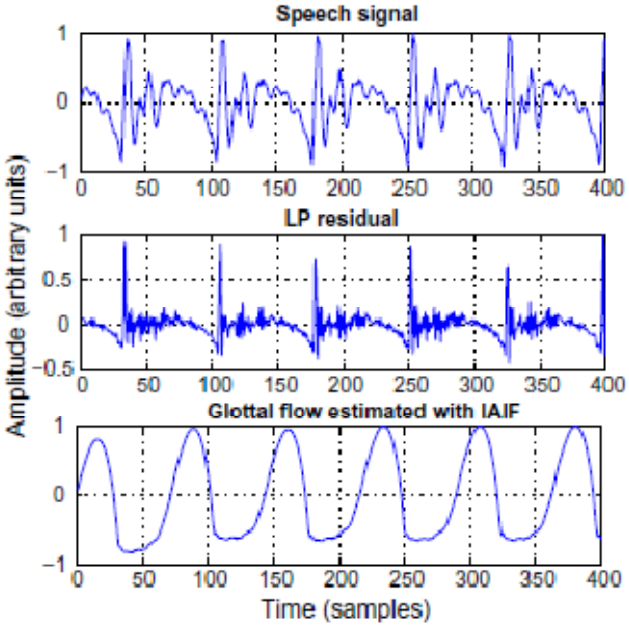


Figure 4.3 Extraction des caractéristiques de la source glottale.

Les coefficients a_k sont habituellement déterminés en minimisant l'énergie résiduelle en utilisant l'algorithme de Levinson-Durbin (Huang et al., 2001 ; Harrington et Cassidy, 1999 ; Rabiner et Juang, 1993). Le modèle spectral est défini comme étant :

$$H(z) = \frac{1}{1-\sum_{k=1}^{P} a_k z^{-k}} \quad (4.3)$$

et il se compose des pics spectraux ou pôles seulement (ligne en tirets et points sur la figure 4.2).

Les coefficients de prédiction $\{a_k\}$ sont rarement utilisés comme des caractéristiques mais ils sont transformés en caractéristiques robustes et moins corrélées tels que les coefficients cepstraux issus de la prédiction linéaire (LPCC) (Huang et al., 2001), les fréquences de raies spectrales (LSF) (Huang et al., 2001), et les coefficients de la prédiction linéaire perceptelle (PLP) (Hermansky, 1990). D'autres caractéristiques, qui ont eu moins de succès, notamment les coefficients de corrélation partielle (PARCOR), les coefficients de surface log (LAR) et les fréquences des formants et les bandes passantes (Rabiner et Juang, 1993).

Compte tenu de toutes les caractéristiques spectrales, lesquelles doivent être choisies pour la reconnaissance du locuteur et comment les paramètres (ex. le number de coefficients) doivent être sélectionnés? Des comparaisons peuvent être trouvées dans (Kinnunen et al., 2004 ; Reynolds et Rose, 1995 ; Atal, 1974), et il a été observé qu'en général les méthodes de compensation de canal sont beaucoup plus importantes que le choix de la base des paramètres (Reynolds et Rose, 1995). Toutefois, différentes caractéristiques spectrales sont complémentaires et peuvent être combinées pour améliorer la précision (Brummer et al., 2007 ; Campbell et al., 2006). En résumé, pour une utilisation pratique, nous recommandons l'une des caractéristiques suivantes : MFCC, LPCC, LSF, PLP.

4.2.2 Caractéristiques de la source vocale

Les caractéristiques de la source vocale caractérisent le signal d'excitation glottique des sons voisés comme la forme de l'impulsion glottale et la fréquence fondamentale, et il est raisonnable de supposer qu'elles transportent l'information spécifique au locuteur. La fréquence fondamentale, la vitesse de vibration des cordes vocales, est populaire et sera discuté dans la section 4.2.4. D'autres paramètres sont liés à la forme de l'impulsion glottique telle que le degré d'ouverture des cordes vocales et la durée de la phase de fermeture. Ils contribuent à la qualité vocale qui peut être décrite par exemple, en tant que soufflée, grinçante ou pressée (Espy-Wilson et al., 2006).

Les caractéristiques de la glotte ne sont pas directement mesurables en raison de l'effet du conduit vocal. En supposant que la source glottique et le conduit vocal sont indépendants les uns des autres, les paramètres de l'appareil vocal peuvent être estimés d'abord en utilisant, par exemple, le modèle de prédiction linéaire, suivi par filtrage inverse du signal d'origine pour obtenir une estimation du signal source (Kinnunen et Alku, 2009 ; Zheng et al , 2007 ; Murty et Yegnanarayana, 2006 ; Prasanna et al., 2006). Une autre méthode utilise une analyse de covariance phase-fermée lorsque les cordes vocales sont fermée (Gudnason et Brookes, 2008 ; Slyh et al., 2004 ; Plumpe et al, 1999). Ceci conduit à l'amélioration de l'estimé du conduit vocal, mais la détection précise de la phase fermée est nécessaire ce qui est difficile dans un environnement bruyant. A titre d'exemple, la figure 4.3 montre un signal de parole avec son signal résiduel LP et le débit glottique avec une méthode de filtrage inverse simple (Alku et al., 1999).

Les caractéristiques du signal inverse filtré peuvent être extraites, par exemple, en utilisant un réseau de neurones auto-associatif (Prasanna et al., 2006). D'autres approches ont utilisé les paramètres du modèle du flux glottique (Plumpe et al., 1999), l'analyse en ondelettes (Zheng et al., 2007), la phase résiduelle (Murty et Yegnanarayana, 2006), les coeffiecients cepstraux (Chetouani et al., 2009 ; Kinnunen et Alku, 2009 ; Gudnason et Brookes, 2008) et des statistiques d'ordre élevé (Chetouani et al., 2009) pour n'en citer que quelques uns.

Basé sur la littérature, les caractéristiques de la source vocale ne sont pas aussi discriminant que les caractéristiques du conduit vocal, mais la fusion de ces deux aspects complémentaires peuvent améliorer la précision (Zheng et al., 2007 ; Murty et Yegnanarayana, 2006). Les expériences de (Chan et al., 2007 ; Prasanna et al., 2006)

suggèrent également que la quantité des données d'apprentissage et de test pour les caractéristiques de la source vocale peuvent être sensiblement inférieure par rapport à celle nécessaire pour les caractéristiques du conduit vocal (10s vs 40s dans (Prasanna et al., 2006)). Une explication possible est que les caractéristiques du conduit vocal dépendent du contenu phonétique et nécessitent donc une couverture phonétique suffisante pour les uttérances d'apprentissage et de test. Les caractéristiques de la source vocale, à leur tour, sont beaucoup moins dépendantes des facteurs phonétiques.

4.2.3 Caractéristiques spectro-teporelles

Il est raisonnable de supposer que les détails du signal spectro-temporel telles que les transitions des formants et les modulations d'énergie contiennent des informations utiles spécifiques au locuteur.

Une façon courante d'incorporer des informations temporelles à des caractéristiques se fait par l'estimation des dérivées premièr et deuxième ordre, connues sous le nom delta et delta-delta coefficients, respectivement (Huang et al., 2001 ; Soong et Rosenberg, 1988 ; Furui, 1981). Ils sont calculés comme la différence de temps entre les vecteurs adjacents de coefficients et en général ajoutés aux coefficients de base au niveau de la trame (par exemple 13 MFCC avec 13 Delta-MFCC et 13 Delta-Delta-MFCC, ce qui implique 39 coefficients par trame). Une alternative, potentiellement plus robuste, la méthode suit une droite de régression (Rabiner et Juang, 1993) ou un polynôme orthogonal (Furui, 1981) pour les trajectoires temporelles, mais dans la pratique une différentiation simple semble donner une performance égale ou meilleure (Kinnunen, 2004). Les composantes principales temps-fréquence (Magrin-Chagnolleau et al., 2002) et les données pilotées par les filtres temporels (Malayath et al., 2000) ont également été étudiées.

Dans (Kinnunen et al., 2008), ils proposent d'utiliser la modulation de fréquence (Atlas et Shamma, 2003 ; Hermansky, 1998) comme une caractéristique pour la reconnaissance du locuteur, comme illustré sur la Figure 4.4. La modulation de fréquence représente le contenu fréquentiel de l'enveloppe de l'amplitude de la sous-bande et elle contient potentiellement des informations sur le taux de parole et d'autres attributs stylistiques. Les fréquences de modulation pertinentes pour l'intelligibilité de la parole sont approximativement dans la gamme 1-20 Hz (Atlas et Shamma, 2003 ; Hermansky, 1998). Dans (Kinnunen, 2006), le meilleur taux de reconnaissance a été obtenu en utilisant une fenêtre temporelle de 300ms et en incluant les fréquences de modulation dans la gamme 0-20Hz. La dimension du vecteur de modulation de fréquence dépend du nombre de points de la FFT du spectrogramme et du nombre de trames couvrant le calcul de la FFT dans la direction temporelle. Pour la meilleure combinaison des paramètres, la dimension du vecteur de caractéristiques a été 3200 (Kinnunen, 2006).

Dans (Kinnunen et al., 2008), ils ont étudié les caractéristiques temporelles à dimension réduite. La méthode de la Transformée Cosinus Discrète Temporelle (TDCT), proposée dans (Kinnunen et al., 2006) et illustrée à la figure 4.5, applique la DCT sur les trajectoires temporelles des vecteurs cepstraux plutôt que sur les amplitudes du spectrogramme. L'utilisation de la DCT plutôt que l'amplitude de la

DFT dans ce cas a l'avantage de conserver les phases relatives des trajectoires des coefficients des caractéristiques, et donc, on peut préserver les informations phonétiques et les informations spécifiques au locuteur. Ceci, cependant, exige encore d'autres travaux de recherche pour confirmer cette préservation. Dans (Kinnunen et al., 2008), la DCT a été utilisée dans un rôle différent : la réduction de la dimensionnalité de la modulation de l'amplitude du spectre. Les meilleurs résultats dans (Kinnunen et al, 2008) ont été obtenus en utilisant un contexte temps de 300-330ms, nettement plus long par rapport à des contextes temps typiques des paramètres delta.

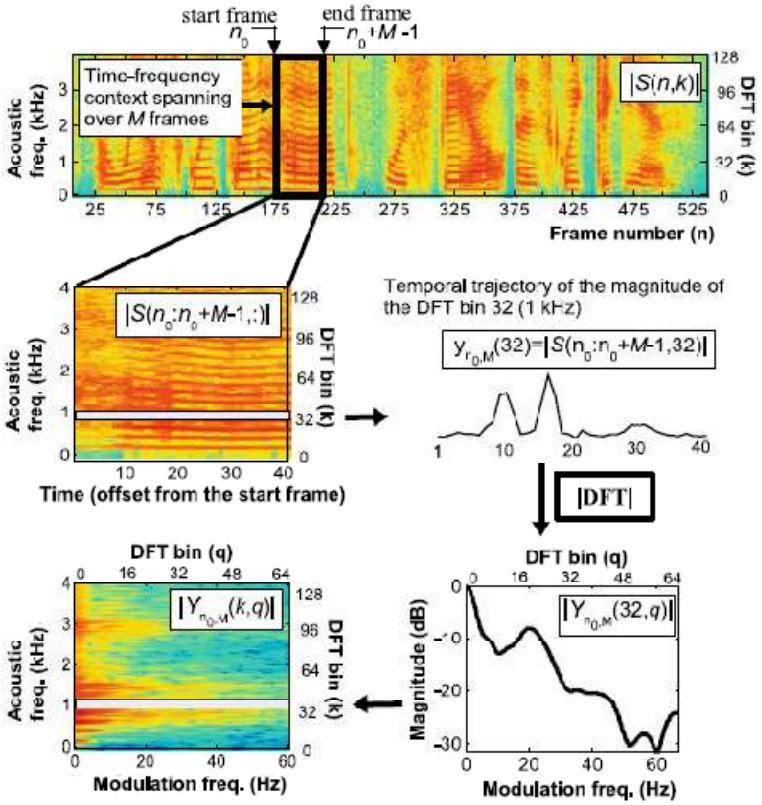


Figure 4.4 Extraction des caractéristiques de la modulation du spectrogramme.

Même s'ils ont obtenu une certaine amélioration par rapport aux systèmes cepstraux en fusionnant les scores des caractéristiques cepstrales et temporelles (Kinnunen, 2006 ; Kinnunen et al, 2006), le gain a été plutôt modeste et plus d'autres travaux de recherche sont nécessaires avant de recommander ces caractéristiques pour des applications pratiques. Un problème pourrait venir du fait qu'ils ont appliqué des techniques de modélisation du locuteur qui ont été conçues pour les caractéristiques court-termes. En raison du contexte temporel plus large, le nombre de vecteur d'apprentissage est généralement plus faible comparé avec les caractéristiques court-termes. En outre, les caractéristiques à court-terme et à long-terme ont des taux de trames différents, ils ne peuvent pas être facilement combinés au niveau de la trame. Peut-être une modélisation complètement différente et une technique de fusion est nécessaire pour ces caractéristiques.

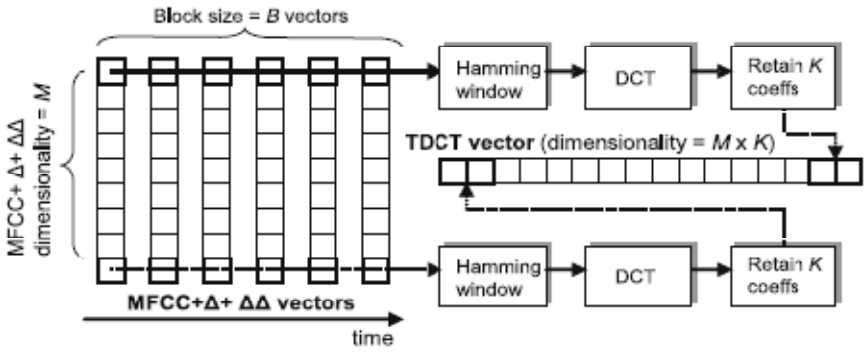


Figure 4.5 Transformée Cosine discrète temporelle (TDCT).

Une alternative aux méthodes basées sur les amplitudes estime les modulations de fréquence (FM) (Thiruvaran et al., 2008a). Dans les méthodes basées sur la FM, le signal d'entrée est d'abord divisé en signaux sous-bandes à l'aide d'un banc de filtres pass-bandes. Les composantes des fréquences dominantes (tels que les centres de gravités des fréquences) dans les sous-bandes captures les caractéristiques similaires à celles des formants. A titre d'exemple, la procédure décrite dans (Thiruvaran et al., 2008a) utilise l'analyse tout-pôles de second ordre pour détecter la fréquence dominante. Les caractéristiques FM sont ensuite obtenues en soustrayant la fréquence centrale de la sous-bande de la fréquence de pôle, ce qui donne une mesure de déviation de la fréquence par « défaut » du signal passe-bande. Cette caractéristique a été appliquée à la reconnaissance du locuteur dans (Thiruvaran et al., 2008b), montrant des résultats prometteurs en la fusionnant avec des MFCC classiques.

4.2.4 Les caractéristiques prosodiques

La prosodie fait référence à des aspects non-segmentaires de la parole, y compris le syllabe-stress par exemple, l'intonation, le taux et le rythme de la parole. Un aspect important de la prosodie est que, contrairement aux traditionnelles caractéristiques spectrales court-terme, elle s'étend sur des segments plus longs, comme les syllabes, les mots et les uttérances et reflète les différences dans le style de parler, les habitudes linguistiques, le type des sentences et l'émotion pour ne citer que quelques uns. Le défi dans la reconnaissance du locuteur indépendante du texte est la modélisation des différents niveaux d'informations prosodiques (instantanée, à long terme) pour capter les différences entre locuteurs. Dans le même temps, les caractéristiques doivent être exemptes d'effets que le locuteur peut contrôler volontairement.

Le paramètre le plus important de la prosodie est la fréquence fondamentale (ou F_0). La combinaison des caractéristiques liées à F_0 avec des caractéristiques spectrales s'est montrée plus efficace surtout dans un environnement bruyant. D'autres caractéristiques prosodiques pour la reconnaissance du locuteur incluent la durée (par exemple les statistiques des pauses, la durée de phonation), le taux de parole et la distribution/modélisation de l'énergie parmi d'autres (Shriberg et al., 2005 ; Reynolds et al, 2003 ; Adami et al., 2003 ; Bartkova et al., 2002). Le lecteur intéressé peut se référer à (Shriberg et al., 2005) pour plus de détails. Dans cette étude, il a été constaté, entre un

certain nombre d'autres observations, que les caractéristiques liées à F_0 donnent la meilleure précision, suivi pas les caractéristiques de l'énergie et la durée dans cet ordre. Depuis, F_0 est la caractéristique prosodique prédominante, que nous allons discuter plus en détail.

La détermination fiable de F_0 est une tâche difficile. Par exemple, dans les cas de la qualité téléphonique de la parole, F_0 est souvent à l'extérieur de la bande étroite du réseau téléphonique (0.3-3.4 kHz) et les algorithmes ne peuvent compter que sur l'information contenue dans les harmoniques supérieures pour la détection de F_0. Pour une discussion détaillée des approches classiques d'estimation de F_0, se reporter à (Hess, 1983). Une comparaison plus récente des suiveurs de F_0 peut être trouvée dans (DeCheveigne et Kawahara, 2001). Pour une utilisation pratique, des méthodes telles que la méthode YIN (DeCheveigne et Kawahara, 2002) et la méthode d'auto-corrélation mis en œuvre dans le logiciel Praat (Boersma et Weenink, 2009) peuvent être utilisées.

Pour la reconnaissance du locuteur, F_0 exprime à la fois des caractéristiques physiologiques et des caractéristiques liées aux habitudes linguistiques. Par exemple, la valeur moyenne de F_0 peut être considérée comme un corrélat acoustique de la taille du larynx (Rose, 2002), tandis que les variations temporelles de la hauteur sont liées à la manière de parler. En reconnaissance dépendante du texte, l'alignement temporel des contours du pitch ont été utilisés (Atal, 1972). Dans les études indépendantes du texte, les statiques long-terme de F_0 – en particulier la valeur moyenne – ont été largement étudiées (Kinnunen et Gonzalez-Hautamaki, 2005 ; Sonmez et al., 1998). La valeur moyenne combinée avec d'autres statistiques comme la variance et l'applatissement (Kurtosis) peuvent être utilisés comme modèle du locuteur (Kinnunen et Gonzalez-Hautamaki, 2005 ; Bartkova et al., 2002 ; Carey et al., 1996), même si les histogrammes (Kinnunen et Gonzalez-Hautamaki, 2005), l'analyse sémantique latente (Chen et al., 2004) et les SVM (Shriberg et al., 2005) donnent de meilleurs résultats. Il a été démontré par certain nombre d'expériences que le logF_0 est meilleure que F_0 lui-même (Kinnunen et Gonzalez-Hautamaki, 2005 ; Sonmez et al., 1997).

F_0 est une caractéristique unidimensionnelle, donc mathématiquement, elle ne devrait pas être très discriminante. Des caractéristiques multidimensionnelles liées au pitch – et au voisement – peuvent être extraites de la fonction d'auto-corrélation (Laskowski et Jin, 2009 ; Ma et al., 2006a ; Wildermoth et Paliwal, 2000) . Par exemple, en utilisant les variations locales et temporelles long-terme de F_0. La capture des dynamiques locales de F_0 peut être obtenue en ajoutant les paramètres delta de la valeur F_0 instantanée. Pour la modélisation à long terme, le contour de F_0 peut être segmenter et présenter par un ensemble de paramètres associés à chaque segment (Adami, 2007 ; Mary et Yegnanarayana, 2006 ; Shriberg et al., 2005 ; Adami et al., 2003 ; Sonmez et al., 1998). Les segments peuvent être obtenus en utilisant les syllabes de la raconnaissance automatique de la parole (ASR) (Shriberg et al., 2005). Une approche alternative, approche ASR-libre, est de diviser l'utterance en unités syllabiques utilisant, par exemple, les onsets voyelles (Mary et Yegnanarayana, 2008) ou les points d'inflexion F_0/Energie (Adami, 2007 ; Dehak et al., 2007) comme des limites pour les segments.

Pour le paramétrage des segments, les statistiques des caractéristiques prosodiques et les pentes temporelles locales (inclinaison) dans chaque segment sont souvent

utilisées. Dans (Adami et al., 2003 ; Sonmez et al., 1998), chaque segment a été paramétré par des morceaux de modèles linéaires dont les paramètres forment les caractéristiques. Dans (Shriberg et al., 2005), les auteurs ont utilisé les valeurs de N-gram des traits discrétisé comme caractéristiques d'un classificateur SVM avec des résultats prometteurs. Dans (Dehak et al., 2007), les caractéristiques prosodiques ont été extraites avec des fonctions à base polynomiale.

4.2.5 Caractéristiques haut-niveau

Les locuteurs diffèrent non seulement par leur timbre de la voix et l'accent/prononciation, mais aussi dans leur lexique – le genre de mots que les locuteurs ont tendance à utiliser dans leurs conversations. Les travaux sur ces conversationnelles caractéristiques haut-niveau ont été initiés par (Doddington, 2001), où le vocabulaire caractéristique d'un locuteur, l'idiolecte soi-disant, a été utilisé pour caractériser les locuteurs. L'idée d'une modélisation haut-niveau est de convertir chaque uttérance dans une séquence de jetons où les modèles de co-occurrence des jetons caractérisent les différences des locuteurs. L'information modélisée est donc une forme catégorique (discrète) plutôt que sous forme numérique (continue).

Les jetons considérés incluent des mots (Doddington, 2001), des phones (Campbell et al., 2004 ; Andrews et al., 2002), les gestes prosodiques (montée/descente pitch/énergie) (Shriberg et al., 2005 ; Chen et al., 2004 ; Adami et al., 2003), et même des jetons articulatoires (mode et lieu d'articulation) (Leung et al., 2006). Le top-1 scoring des indices des composantes des mélanges de gaussienne ont également été utilisés comme jetons (Ma et al., 2006b ; Xiang, 2003 ; Torres-Carrasquillo et al., 2002).

Parfois, plusieurs tokeniseurs parallèles sont utilisés (Ma et al., 2006b ; Campbell et al., 2004). Ceci est en partie motivé par le succès des reconnaisseurs parallèles de phones dans l'état de l'art de la reconnaissance des langues (Zissman, 1996). Cette orientation est motivée par l'espoir que les différents tokeniseurs (par exemple, les reconnaisseurs de phone entrainés sur différents langages ou avec différents modèles de phones) seraient capables de capter des aspects complémentaires de l'énoncé. A titre d'exemple, dans (Ma et al., 2006b) un ensemble de tokeniseurs GMM parallèles (Xiang, 2003 ; Torres-Carrasquillo et al., 2002) ont été utilisés. Chaque tokeniseur a été entrainé à partir d'un groupe différent de locuteurs obtenu par regroupement.

Le classificateur de base pour éléments du jeton est basé sur la modélisation N-gram. Notons la séquence symbolique de l'énoncé par $\{a_1, a_2, \ldots, a_T\}$, où a_t appartient à V est un vocabulaire fini. Un modèle N-gram est construit par l'estimation de la probabilité conjointe de N consécutifs jetons. Par exemple, N=2 donne le modèle bigramme où les probabilités de la paire des jetons (a_t, a_{t+1}) sont estimées. Un modèle trigramme est composé de triplets (a_t, a_{t+1}, a_{t+2}), et ainsi de suite. A titre d'exemple, les bigrammes du jeton de la séquence hello_world sont (h,e), (e,l), (l,l), (l,o), (o,_), (_,w), (w,o), (o,r), (r,l) et (l,d). La probabilité de chaque N-gram est estimé de la même manière que le N-gram dans le modèle statistique de langage en reconnaissance automatique du locuteur (Ney et al., 1997). L'estimation du N-gram c'est le maximum de vraisemblance (ML) ou maximum a postériori (MAP) dans le corpus d'apprentissage (Leung et al., 2006). Les statistiques N-gram ont été utilisées dans

l'espace vectoriel (Ma et al., 2006b ; Campbell et al., 2004) et avec des mesures d'entropie (Leung et al., 2006 ; Andrews et al., 2001) pour évaluer la similitude entre les locuteurs.

4.3 Sélection des caractéristiques

D'après (Rose, 2002 ; Wolf, 1972), les paramètres idéaux et efficaces pour identifier un locuteur devront être :

- facilement mesurable ;
- difficiles à imiter par les imposteurs ;
- robustes aux bruits et canaux de transmission ;
- être très fréquents dans le signal vocal.
- ayant une grande variabilité inter-locuteurs et une faible variabilité intra-locuteur ;
- ne pas être affectés par la santé du locuteur ou des variations à long terme dans la voix.

Plusieurs travaux ont été publiés pour comparer différentes techniques en paramétrisation. L'enjeu de ces travaux était surtout axé à cibler les meilleurs paramètres encodant de façon efficace les propriétés caractéristiques de chaque locuteur.

Atal (Atal, 1976 ; Atal, 1974 ; Atal, 1972) a expérimenté une série de paramètres censés représenter la contribution des différents organes mis en jeu lors du mécanisme de production (vibration des cordes vocales, volume d'air dans les poumons, résonance du conduit vocal, etc.). L'ensemble de ces paramètres proposés sont les coefficients par prédiction linéaire, les coefficients cepstraux calculés par la méthode de LPC (« Linear Predictive Coding ») et leurs dérivés (log-area, énergie spectrale, coefficients de réflexion) ainsi que l'évolution de l'énergie et du fondamentale dans le temps. Les meilleurs résultats ont été obtenus par les coefficients cepstraux (extraits par la méthode LPC) suivis par ceux de la prédiction linéaire.

Cependant, le succès de l'utilisation des coefficients cepstraux a été freiné par des facteurs liés à la variabilité entre les données d'entraînement et de tests, le bruit ainsi que les distorsions introduites par le canal de transmission.

Dans cette perspective, d'autres études et approches ont été proposées dans l'objectif de trouver des paramètres plus robustes et indépendants de conditions d'utilisation. Les articles suivants discutent d'une éventuelle paramétrisation dans le domaine spectrale (Ong et al, 1994 ; Reynolds, 1994). On trouve les méthodes PLP, RASTA, RASTA-PLP et MFCC (Hermansky et Morgan, 1994 ; Hermansky, 1992), qui se distinguent par l'incorporation de certaines propriétés perceptives. En effet, ces approches ont déjà été utilisées dans les systèmes de reconnaissance du vocabulaire et sont reconnues par leur robustesse à la variabilité intra-locuteur et au contexte d'environnement. Openshaw et al (Openshaw et al, 1993) ont comparé la robustesse de ces méthodes dans la situation où les données sont affectées (entraînement et/ou tests) par un bruit additif de type blanc gaussien. Les paramètres PLP-RASTA donnent les

meilleurs résultats que les paramètres MFCC, PLP, MFCC-RASTA, delta-PLP-RESTA, etc. Ils confirment que l'utilisation du vecteur différence semble intéressante dans la seule situation où le rapport signal bruit est faible.

La méthode du bispectre (statistique d'ordre supérieur) a été introduite par Wenndt et al (Wenndt et al, 1997). Dans la situation où les données d'apprentissage sont propres et les données testées sont bruitées, les paramètres de bispectre donnent de meilleurs résultats que ceux du cepstre. Par contre, cette méthode n'a pas obtenu de succès avec la parole téléphonique à cause des distorsions de la phase et l'absence du fondamental.

Imprel et al (Imprel et al., 1997) ont présenté une méthode basée sur la décomposition harmonique de Hildebrand-Prony qui consiste à modéliser les données par une combinaison linéaire d'exponentielles sans poser de contrainte sur les fréquences. Ils confirment que cette méthode est plus précise et offre une haute résolution comparativement à la procédure utilisant la FFT. Ils indiquent que les indices harmoniques estimés par Hildebrand-Prony donnent de meilleurs résultats comparativement aux coefficients cepstraux dérivés par LPC. Par contre, le problème majeur de cette approche est sa sensibilité au bruit et la non-stationnarité du signal vocal.

L'enveloppe spectrale du résiduel est considérée comme caractéristique du locuteur (Doddongton, 1985). En effet le résiduel (à spectre large) a été fortement négligé et peu d'auteurs lui ont accordé une importance. Peut-être l'estimation de ce dernier est accompagnée d'erreur et les outils disponibles en traitement de parole ne permettent pas d'extraire adéquatement le résiduel. Thevenaz et Hugli (Thevenaz et Hugli, 1995) ont étudié la contribution du résiduel. Les résultats obtenu démontrent l'utilité propre du résiduel même si les performances du résiduel apparaissent moindre que celles du filtre de synthèse. Particulièrement, le résiduel se montre utile quand sa composante est combinée à celles du filtre de synthèse.

Dans toutes les méthodes présentées, les paramètres caractérisant le locuteur renferment conjointement l'information liée au contexte phonétique et celle propre du locuteur. Malayath et al (Malayath et al, 1997) ont essayé d'extraire les paramètres propres du locuteur sans tenir compte du contexte phonétique. Ils ont procédé par une analyse par composante principale normée pour trouver les valeurs propres et les vecteurs propres de deux matrices de données. La première matrice correspond à la différence des segments de paroles ayant le même contexte phonétique du même locuteur. La deuxième matrice tient compte de la variabilité inter-locuteur. Sous l'hypothèse que le contexte phonétique domine les caractéristiques du locuteur, les vecteurs propres à faibles valeurs propres sont désignés comme un nouvel espace de représentation du locuteur. Le problème de cette méthode revient à la segmentation et sera évidemment difficile à utiliser dans les situations réelles.

Jankowski et al (Jankowski et al, 1995) ont proposé une approche capable d'extraire les variations fines du signal vocal à chaque impulsion glottale par modulation AM-FM et par analyse des lobes secondaires de l'enveloppe. Au début de l'analyse, ils calculent les coefficients de prédiction pour localiser les 3 premiers formants. Un filtre passe-bande est appliqué autour de la fréquence centrale des 3 formants, suivi par Teager (Teager, 1993) (opérateur non-linéaire) pour démoduler le signal. Les lobes secondaires

et le fondamental sont estimés à partir de l'enveloppe du signal. Les résultats obtenus avec les paramètres MFCC+AM-FM améliorent les performances en reconnaissance du locuteur de 1.2% (+8.2% pour les femmes et -2.2% pour les hommes). D'un autre côté, les paramètres MFCC+fondamental+lobes secondaires améliorent les performances chez les hommes de 4% sans pour autant affecter celle des femmes.

On peut trouver d'autres approches sur la modélisation de la source glottale et l'incorporation d'informations prosodiques destinées à la caractérisation du locuteur (Hansen et al, 2004 ; Plumpe et al, 1999).

Cependant, malgré la recherche intense pour trouver de nouveaux paramètres en caractérisation du locuteur, il reste que les coefficients MFCC demeurent le meilleur choix. En effet, ils sont supposés être très bien représentatifs de la forme du conduit vocal. Leurs distributions statistiques sont particulièrement bien modélisées par le modèle à mélanges de gaussiennes et les composantes du vecteur des coefficients sont convenablement décorrelées.

4.4 Conclusion

L'étude de l'efficacité de différents types d'analyse pour la RAL a fait l'objet de nombreuses publications. Dans la plupart de ces études, le système d'identification ou de vérification est fixé et les résultats obtenus pour des paramètres variés (LPC, LPCC, MFCC, Pitch, …) sont comparés.

Dans les méthodes utilisées pour ces comparaisons, la nature de la distance entre deux vecteurs de paramètres est essentielle. Une des caractéristiques essentielles d'une « bonne » distance est de tenir compte des variances des différents coefficients d'un même vecteur de paramètres. Les coefficients c_1, …, c_p d'un vecteur LPCC présentent par exemple des variances de plus en plus faibles. Une distance euclidienne entre deux vecteurs LPCC non normalisés est donc peu intéressante. Cette étape de normalisation est primordiale, et a été introduite initialement en RAL en 1974. A partir des résultats comparatifs publiés, on peut tirer un certain nombre de conclusions :

- Les paramètres dérivés d'une analyse temporelle, les LPCs, LARs, PARCORs, contiennent sensiblement le même type d'information. Les résultats obtenus (p.exp taux de reconnaissance) pour ces différents paramètres sont non seulement comparables en moyenne, mais les confusions interviennent sur les mêmes segments de parole.
- Les coefficients cepstraux LPCC conduisent à de meilleurs résultats que les coefficients précités (LPC, LAR, …).
- Le pitch et sa dynamique ne permettent pas de reconnaître efficacement l'identité d'un locuteur.
- Différents paramètres gagnent à être combinés car ils peuvent contenir des informations complémentaires (p.exp les LPCC et le pitch).
- La dynamique des paramètres (paramètres différentielles contiennent également une information différente des vecteurs instantanés. Ceci a été mis

en évidence par l'étude de la corrélation entre ces différentes informations. Les paramètres instantanés gagnent donc à être combinés.

- Les résultats généralement reportés montrent une légère supériorité des paramètres instantanés sur les paramètres différentiels (pour le pitch, les LPCC ou les MFCC), mais il ne s'agit pas d'un comportement systématique (des études ont montré la supériorité des paramètres différentiels sur des données bruitées).
- Cependant, malgré la recherche intense pour trouver de nouveaux paramètres en caractérisation du locuteur, il reste que les coefficients MFCC demeurent le meilleur choix. En effet, ils sont supposés être très bien représentatifs de la forme du conduit vocal. Leurs distributions statistiques sont particulièrement bien modélisées par le modèle à mélanges de Gaussiennes et les composantes du vecteur des coefficients sont convenablement décorrélées.
- A ce jour, l'information de haut niveau n'a pas été fréquemment mise en œuvre dans la reconnaissance du locuteur, cela est dû principalement à la difficulté de mesure automatique et quantitative de ce genre d'information. Néanmoins, récemment, les efforts conjoints de 10 instituts (MIT, IBM, OGI, et d'autres) ont été mis en place afin d'exploiter l'efficacité de l'information haut-niveau pour un système de reconnaissance de locuteur précis. Dans ce projet commun, une large panoplies d'approches utilisant des modèles de prononciation, de la dynamique prosodique, les caractéristiques du pitch et de la durée, les flux de phones, et les interactions conversationnelles ont été explorés et développées. Il a été démontré que ces nouvelles caractéristiques et classificateurs, en effet, fournissent des informations complémentaires et peuvent êtres fusionner pour réduire l'erreur de reconnaissance (Reynolds et al., 2003).

4.5 Bibliographie

(Adami, 2007) Adami, A., 2007. Modeling prosodic differences for speaker recognition. Speech Comm. 49 (4), 277–291.

(Adami et al., 2003) Adami, A., Mihaescu, R., Reynolds, D., Godfrey, J., 2003. Modeling prosodic dynamics for speaker recognition. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003), Hong Kong, China, April 2003, pp. 788–791.

(Alku et al., 1999) Alku, P., Tiitinen, H., Na¨a¨ta¨nen, R., 1999. A method for generating natural-sounding speech stimuli for cognitive brain research. Clin. Neurophysiol. 110 (8), 1329–1333.

(Ambikairajah, 2007) Ambikairajah, E., 2007. Emerging features for speaker recognition. In: Proc. Sixth Internat. IEEE Conf. on Information, Communications & Signal Processing, Singapore, December 2007, pp. 1–7.

(Andrews et al., 2002) Andrews, W., Kohler, M., Campbell, J., Godfrey, J., Hernandez-Cordero, J., 2002. Gender-dependent phonetic refraction for speaker recognition. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002), Vol. 1, Orlando, Florida, USA, May 2002, pp. 149–152.

(Arcienega et al., 2001) Arcienega, M., Drygajlo, A., 2001. Pitch-dependent GMMs for textindependent speaker recognition systems. In: Proc. Seventh European Conf. on Speech Communication and Technology (Eurospeech 2001), Aalborg, Denmark, September 2001, pp. 2821–2824.

(Ashour et Gath, 1999) Ashour, G., Gath, I., 1999. Characterization of speech during imitation. In: Proc. Sixth European Conf. on Speech Communication and Technology (Eurospeech 1999), Budapest, Hungary, September 1999, pp. 1187–1190.

(Atal, 1976) Atal B. S., Automatic recognition of speakers from their voices, In Proc. IEEE, pp. 460-475, Vol. 64, 1976.

(Atal, 1974) Atal, B., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Amer. 55 (6), 1304–1312.

(Atal, 1972) Atal, B., 1972. Automatic speaker recognition based on pitch contours. J. Acoust. Soc. Amer. 52 (6), 1687–1697.

(Atlas et Shamma, 2003) Atlas, L., Shamma, S., 2003. Joint acoustic and modulation frequency. EURASIP J. Appl. Signal Process. 7, 668–675.

(Bartkova et al., 2002) Bartkova, K., Gac, D.L., Charlet, D., Jouvet, D., 2002. Prosodic parameter for speaker identification. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP 2002), Denver, Colorado, USA, September 2002, pp. 1197–1200.

(Besacier et Bonastre, 2000) Besacier, L., Bonastre, J.-F., 2000. Subband architecture for automatic speaker recognition. Signal Process. 80, 1245–1259.

(Boersma et Weenink, 2009) Boersma, P., Weenink, D., 2009. Praat: doing phonetics by computer [computer program]. WWWpage, June 2009, <http://www.praat.org/>.

(Brummer et al., 2007) Bru¨mmer, N., Burget, L., C˘ ernocky´, J., Glembek, O., Gre´zl, F., Karafia´t, M., Leeuwen, D., Mate˘jka, P., Schwartz, P., Strasheim, A., 2007. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. IEEE Trans. Audio, Speech Language Process. 15 (7), 2072–2084.

(Campbell et al., 2006) Campbell, W., Campbell, J., Reynolds, D., Singer, E., Torres-Carrasquillo, P., 2006a. Support vector machines for speaker and language recognition. Comput. Speech Lang. 20 (2–3), 210–229.

(Campbell et al., 2004) Campbell, W., Campbell, J., Reynolds, D., Jones, D., Leek, T., 2004. Phonetic speaker recognition with support vector machines. In: Thrun, S., Saul, L., Schokopf, B. (Eds.), . In: Advances in Neural Information Processing Systems, Vol. 16. MIT Press, Cambridge, MA.

(Carey et al., 1996) Carey, M., Parris, E., Lloyd-Thomas, H., Bennett, S., 1996. Robust prosodic features for speaker identification. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP 1996), Philadelphia, Pennsylvania, USA, 1996, pp. 1800–1803.

(Chan et al., 2007) Chan, W., Zheng, N., Lee, T., 2007. Discrimination power of vocal source and vocal tract related features for speaker segmentation. IEEE Trans. Audio, Speech Language Process. 15 (6), 1884–1892.

(Charbouillet et al., 2006) Charbuillet, C., Gas, B., Chetouani, M., Zarader, J., 2006. Filter bank design for speaker diarization based on genetic algorithms. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006), Vol. 1, Toulouse, France, May 2006, pp. 673–676.

(Chen et al., 2004) Chen, Z.-H., Liao, Y.-F., Juang, Y.-T., 2004. Eigen-prosody analysis for robust speaker recognition under mismatch handset environment. In: Processing (ICSLP 2004), Jeju, South Korea, October 2004, pp. 1421–1424.

(Chetouani et al., 2009) Chetouani, M., Faundez-Zanuy, M., Gas, B., Zarader, J., 2009. Investigation on LP-residual presentations for speaker identification. Pattern Recognition 42 (3), 487–494.

(Damper et Higgins, 2003) Damper, R., Higgins, J., 2003. Improving speaker identification in noise by subband processing and decision fusion. Pattern Recognition Lett. 24, 2167–2173.

(Davis et Mermelstein, 1980) Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics, Speech, Signal Process. 28 (4), 357–366.

(Decheveigne et Kawahara, 2002) DeCheveigne, A., Kawahara, H., 2002. YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Amer. 111 (4), 1917–1930.

(DeCheveigne et Kawahara, 2001) Cheveigne´, A., Kawahara, H., 2001. Comparative evaluation of f0 estimation algorithms. In: Proc. Seventh European Conf. on Speech Communication and Technology (Eurospeech 2001), Aalborg, Denmark, September 2001, pp. 2451–2454.

(Dehak et al., 2007) Dehak, N., Kenny, P., Dumouchel, P., 2007. Modeling prosodic features with joint factor analysis for speaker verification. IEEE Trans. Audio, Speech Language Process. 15 (7), 2095–2103.

(Deller et al., 2000) Deller, J., Hansen, J., Proakis, J., 2000. Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York.

(Doddington, 2001) Doddington, G., 2001. Speaker recognition based on idiolectal differences between speakers. In: Proc. Seventh European Conf. on Speech Communication and Technology (Eurospeech 2001), Aalborg, Denmark, September 2001, pp. 2521–2524.

(Doddongton, 1985) Doddington G., Speaker recognition - Identifying people by their voices. Proceedings of the IEEE, November, 73(11): 1651, 1985.

(Epsy-Wilson et al., 2006) Espy-Wilson, C., Manocha, S., Vishnubhotla, S., 2006. A new set of features for text-independent speaker identification. In: Proc. Interspeech 2006 (ICSLP), Pittsburgh, Pennsylvania, USA, September 2006, pp. 1475–1478.

(Ezzaidi et al., 2001) Ezzaidi, H., Rouat, J., O'Shaughnessy, D., 2001. Towards combining pitch and MFCC for speaker identification systems. In: Proc. Seventh European Conf. on Speech Communication and Technology (Eurospeech 2001), Aalborg, Denmark, September 2001, pp. 2825–2828.

(Furui, 1981) Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoustics, Speech Signal Process. 29 (2), 254–272.

(Furui et al., 1972) Furui S., Itakura F. and Saito S., Talker recognition by long time averaged speech spectrum, In Elect. Commun., pp. 65-61, Vol. 55-A(10), 1972, Japan.

(Gong et al., 2008) Gong, W.-G., Yang, L.-P., Chen, D., 2008. Pitch synchronous based feature extraction for noise-robust speaker verification. In: Proc. Image and Signal Processing (CISP 2008), Vol. 5, (May 2008), pp. 295–298.

(Gopalan et al., 1999) Gopalan, K., Anderson, T., Cupples, E., 1999. A comparison of speaker identification results using features based on cepstrum and Fourier–Bessel expansion. IEEE Trans. Speech Audio Process. 7 (3), 289–294.

(Gudnason et Brookes, 2008) Gudnason, J., Brookes, M., 2008. Voice source cepstrum coefficients for speaker identification. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2008), Las Vegas, Nevada, March–April 2008, pp. 4821–4824.

(Hansen et al., 2004) Hansen, E., Slyh, R., Anderson, T., 2004. Speaker recognition using phoneme-specific GMMs. In: Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2004), Toledo, Spain, May 2004, pp. 179–184.

(Harrington et Cassidy, 1999) Harrington, J., Cassidy, S., 1999. Techniques in Speech Acoustics. Kluwer Academic Publishers, Dordrecht. Harris, F., 1978. On the use of windows for harmonic analysis with the discrete fourier transform. Proc. IEEE 66 (1), 51–84.

(Harris, 1978) Harris, F., 1978. On the use of windows for harmonic analysis with the discrete fourier transform. Proc. IEEE 66 (1), 51–84.

(Hedge et al., 2004) Hedge, R., Murthy, H., Rao, G., 2004. Application of the modified group delay function to speaker identification and discrimination. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2004), Vol. 1, Montreal, Canada, May 2004, pp. 517–520.

(Hermansky, 1998) Hermansky, H., 1998. Should recognizers have ears?. Speech Comm. 25 (1–3) 3–27.

(Hermansky et Morgan, 1994) Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589.

(Hermansky, 1992) Hermansky, H. 1992. RASTA extensions: Robustness to additive and convolutional noise, In Proceedings of the Workshop on Speech Processing in Adverse Conditions SPAC-1992, 115-118.

(Hermansky, 1990) Hermansky, H., 1990. Perceptual linear prediction (PLP) analysis for speech. J. Acoust. Soc. Amer. 87, 1738–1752.

(Hess, 1983) Hess, W., 1983. Pitch Determination of Speech Signals: Algorithms and Devices. Springer-Verlag, Berlin.

(Huang et al., 2001) Huang, X., Acero, A., Hon, H.-W., 2001. Spoken Language Processing: a Guide to Theory, Algorithm, and System Development. Prentice-Hall, New Jersey.

(Imprel et al., 1997) Imperl, B., Kacic, Z., Horvat, B., 1997. A study of harmonic features for the speaker recognition. Speech Comm. 22 (4), 385–402.

(Jain et al., 2000) Jain, A., Duin, R., Mao, J., 2000. Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Machine Intell. 22 (1), 4–37.

(Jang et al., 2002) Jang, G.-J., Lee, T.-W., Oh, Y.-H., 2002. Learning statistically efficient features for speaker recognition. Neurocomputing 49, 329–348.

(Jankowski et al, 1995) Jankowski, Quatieri, et al. Measuring Fine Structure in Speech: Application to. Speaker Identification," Proc. Intl. Conf. Acoust., Speech, and Sig. Proc. (1995), pp. 325-328.

(Kinnunen et Alku, 2009) Kinnunen, T., Alku, P., 2009. On separating glottal source and vocal tract information in telephony speaker verification. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2009), Taipei, Taiwan, April 2009, pp. 4545–4548.

(Kinnunen et al., 2008) Kinnunen, T., Lee, K.-A., Li, H. 2008. Dimension reduction of the modulation spectrogram for speaker verification. In: The Speaker and Language Recognition Workshop (Odyssey 2008), Stellenbosch, South Africa, January 2008.

(Kinnunen et al., 2007) Kinnunen, T., Zhang, B., Zhu, J., Wang, Y., 2007. Speaker verification with adaptive spectral subband centroids. In: Proc. Internat. Conf. on Biometrics (ICB 2007), Seoul, Korea, August 2007, pp. 58–66.

(Kinnunen, 2006) Kinnunen, T., 2006. Joint acoustic-modulation frequency for speaker recognition. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006), Vol. I, Toulouse, France, 2006, pp. 665–668.

(Kinnunen et al., 2006) Kinnunen, T., Koh, C., Wang, L., Li, H., Chng, E., 2006. Temporal discrete cosine transform: Towards longer term temporal features for speaker verification. In: Proc. Fifth Internat. Symposium on Chinese Spoken Language Processing (ISCSLP 2006), Singapore, December 2006, pp. 547–558.

(Kinnunen et Gonzalez-Hautamaki, 2005) Kinnunen, T., Gonza´lez-Hautama¨ki, R., 2005. Long-term f0 modeling for text-independent speaker recognition. In: Proc. 10th Internat. Conf. on Speech and Computer (SPECOM'2005), Patras, Greece, October 2005, pp. 567–570.

(Kinnunen, 2004) Kinnunen, T., 2004. Spectral Features for Automatic Text-Independent Speaker Recognition. Licentiate's Thesis, University of Joensuu, Department of Computer Science, Joensuu, Finland.

(Kinnunen et al., 2004) Kinnunen, T., Hautama¨ki, V., Fra¨nti, P., 2004. Fusion of spectral feature sets for accurate speaker identification. In: Proc. Ninth Internat. Conf. on Speech and Computer (SPECOM 2004), St. Petersburg, Russia, September 2004, pp. 361–365.

(Kitamura, 2008) Kitamura, T., 2008. Acoustic analysis of imitated voice produced by a professional impersonator. In: Proc. Interspeech 2008, September 2008, pp. 813–816.

(Laskowski et Jin, 2009) Laskowski, K., Jin, Q., 2009. Modeling instantaneous intonation for speaker identification using the fundamental frequency variation spectrum. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2009), Taipei, Taiwan, April 2009, pp. 4541–4544.

(Leung et al., 2006) Leung, K., Mak, M., Siu, M., Kung, S., 2006. Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification. Speech Comm. 48 (1), 71–84.

(Lu et Dang, 2007) Lu, X., Dang, J., 2007. An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. Speech Comm. 50 (4), 312–322.

(Ma et al., 2006a) Ma, B., Zhu, D., Tong, R., 2006. Chinese dialect identification using tone features based on pitch flux. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006), Vol. 1, Toulouse, France, May 2006, pp. 1029–1032.

(Ma et al., 2006b) Ma, B., Zhu, D., Tong, R., Li, H., 2006. Speaker cluster based GMM tokenization for speaker recognition. In: Proc. Interspeech 2006 (ICSLP), Pittsburgh, Pennsylvania, USA, September 2006, pp. 505–508.

(Magrin-Chagnolleau et al., 2002) Magrin-Chagnolleau, I., Durou, G., Bimbot, F., 2002. Application of time–frequency principal component analysis to text-independent speaker identification. IEEE Trans. Speech Audio Process. 10 (6), 371–378.

(Makhoul, 1975) Makhoul, J., 1975. Linear prediction: a tutorial review. Proc. IEEE 64 (4), 561–580.

(Malayath et al., 2000) Malayath, N., Hermansky, H., Kajarekar, S., Yegnanarayana, B., 2000. Data-driven temporal filters and alternatives to GMM in speaker verification. Digital Signal Process. 10 (1–3), 55–74.

(Malayath et al, 1997) Malayath, Narendranath / Hermansky, Hynek / Kain, Alexander (1997): "Towards decomposing the sources of variability in speech", In EUROSPEECH-1997, 497-500.

(Mammone et al., 1996) Mammone, R., Zhang, X., Ramachandran, R., 1996. Robust speaker recognition: a feature based approach. IEEE Signal Process. Mag. 13 (5), 58–71.

(Mary et Yegnanarayan, 2008) Mary, L., Yegnanarayana, B., 2008. Extraction and representation of prosodic features for language and speaker recognition. Speech Comm. 50 (10), 782–796.

(Mary et Yegnanarayan, 2006) Mason, M., Vogt, R., Baker, B., Sridharan, S., 2005. Data-driven clustering for blind feature mapping in speaker verification. In: Proc. Interspeech 2005, Lisboa, Portugal, September 2005, pp. 3109–3112.

(Miyajima et al., 2001) Miyajima, C., Watanabe, H., Tokuda, K., Kitamura, T., Katagiri, S., 2001. A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction. Speech Comm. 35, 203–218.

(Murty et Yegnanarayana, 2006) Murty, K., Yegnanarayana, B., 2006. Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Process. Lett. 13 (1), 52–55.

(Nakasone et al., 2004) Nakasone, H., Mimikopoulos, M., Beck, S., Mathur, S., 2004. Pitch synchronized speech processing (PSSP) for speaker recognition. In: Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2004), Toledo, Spain, May 2004, pp. 251–256.

(Ney et al., 1997) Ney, H., Martin, S., Wessel, F., 1997. Statistical language modeling using leaving-one-out. In: Young, S., Bloothooft, G. (Eds.), Corpus-based Methods in Language and Speech Processing. Kluwer Academic Publishers, pp. 174–207.

(Ong et al, 1994) Ong, S., Moody, M. P., Sridharan, Sridha (1994): "Confidence analysis for speaker identification: the effectiveness of various features", In ASRIV-1994, 91-94.

(Openshaw et al, 1993) J. P. Openshaw, Z. P. Sun, and J. S. Mason. A comparison of composite features under degraded speech in speaker recognition. In Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pages II-371-II-374, 1993.

(Oppenheim et al., 1999) Oppenheim, A., Schafer, R., Buck, J., 1999. Discrete-Time Signal Processing, second ed. Prentice-Hall, 1999.

(Orman et Arslam, 2001) Orman, D., Arslan, L., 2001. Frequency analysis of speaker identification. In: Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2001), Crete, Greece, June 2001, pp. 219–222.

(Paliwal et Alsteris, 2003) Paliwal, K., Alsteris, L., 2003. Usefulness of phase spectrum in human speech perception. In: Proc. Eighth European Conf. on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, September 2003, pp. 2117–2120.

(Plumpe et al., 1999) Plumpe, M., Quatieri, T., Reynolds, D., 1999. Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. Speech Audio Process. 7 (5), 569–586.

(Prasanna et al., 2006) Prasanna, S., Gupta, C., Yegnanarayana, B., 2006. Extraction of speakerspecific excitation information from linear prediction residual of speech. Speech Comm. 48, 1243–1261.

(Rabiner et Juang, 1993) Rabiner, L., Juang, B.-H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, New Jersey.

(Reynolds et al., 2003) Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D., Xiang, B., 2003. The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003). Hong Kong, China, April 2003, pp. 784–787.

(Reynolds et al., 2000) Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted gaussian mixture models. Digital Signal Process. 10 (1), 19–41.

(Reynolds et Rose, 1995) Reynolds, D., Rose, R., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. 3, 72–83.

(Reynolds, 1995) Reynolds, D., 1995. Speaker identification and verification using Gaussian mixture speaker models. Speech Comm. 17, 91–108.

(Reynolds, 1994) Reynolds, Douglas A. (1994): "Speaker identification and verification using Gaussian mixture speaker models", In ASRIV-1994, 27-30.

(Rose, 2002) Rose, P., 2002. Forensic Speaker Identification. Taylor & Francis, London.

(Shriberg et al., 2005) Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A., 2005. Modeling prosodic feature sequences for speaker recognition. Speech Comm. 46 (3–4), 455–472.

(Sivakumaran et al., 2003) Sivakumaran, P., Ariyaeeinia, A., Loomes, M., 2003a. Sub-band based text-dependent speaker verification. Speech Comm. 41, 485–509.

(Slyh et al., 2004) Slyh, R., Hansen, E., Anderson, T., 2004. Glottal modeling and closedphase analysis for speaker recognition. In: Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2004), Toledo, Spain, May 2004, pp. 315–322.

(Sonmez et al., 1998) So¨nmez, K., Shriberg, E., Heck, L., Weintraub, M., 1998. Modeling dynamic prosodic variation for speaker verification. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP 1998), Sydney, Australia, November 1998, pp. 3189–3192.

(Sonmez et al., 1997) So¨nmez, M., Heck, L., Weintraub, M., Shriberg, E., 1997. A lognormal tied mixture model of pitch for prosody-based speaker recognition. In: Proc. Fifth European Conf. on Speech Communication and Technology (Eurospeech 1997), Rhodos, Greece, September 1997, pp. 1391–1394.

(Soong et Rosenberg, 1988) Soong, F., Rosenberg, A., 1988. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Trans. Acoustics, Speech Signal Process. 36 (6), 871–879.

(Thevenaz et Hugli, 1995) The´venaz, P., Hu¨ gli, H., 1995. Usefulness of the LPC-residue in textindependent speaker verification. Speech Comm. 17 (1–2), 145–157.

(Thian et al., 2004) Thian, N., Sanderson, C., Bengio, S., 2004. Spectral subband centroids as complementary features for speaker authentication. In: Proc. First Internat. Conf. on Biometric Authentication (ICBA 2004), Hong Kong, China, July 2004, pp. 631–639.

(Thiruvaran et al., 2008a) Thiruvaran, T., Ambikairajah, E., Epps, J., 2008a. Extraction of FM components from speech signals using all-pole model. Electronics Lett. 44 (6).

(Thiruvaran et al, 2008b) Thiruvaran, T., Ambikairajah, E., Epps, J., 2008b. FM features for automatic forensic speaker recognition. In: Proc. Interspeech 2008, Brisbane, Australia, September 2008, pp. 1497–1500.

(Torres-Carrasquillo et al., 2002) Torres-Carrasquillo, P., Reynolds, D., Deller Jr., J.D., 2002. Language identification using Gaussian mixture model tokenization. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002), Vol. 1, Orlando, Florida, USA, May 2002, pp. 757–760.

(Wildermoth et Paliwal, 2000) Wildermoth, B., and Paliwal, K. 2000. Use of voicing and pitch information for speaker recognition. In: Proc. Eighth Australian Internat. Conf. on Speech Science and Technology, Canberra, December 2000, pp. 324–328.

(Wolf, 1972) Wolf, J., 1972. Efficient acoustic parameters for speaker recognition. J. Acoust. Soc. Amer. 51 (6), 2044–2056 (Part 2).

(Xiang, 2003) Xiang, B., 2003. Text-independent speaker verification with dynamic trajectory model. IEEE Signal Process. Lett. 10, 141–143.

(Zheng et al., 2007) Zheng, N., Lee, T., Ching, P., 2007. Integration of complementary acoustic features for speaker recognition. IEEE Signal Process. Lett. 14 (3), 181–184.

(Zilca et al., 2006) Zilca, R., Kingsbury, B., Navra´til, J., Ramaswamy, G., 2006. Pseudo pitch synchronous analysis of speech with applications to speaker recognition. IEEE Trans. Audio, Speech Language Process. 14 (2), 467–478.

(Zissman, 1996) Zissman, M., 1996. Comparison of four approaches to automatic language identification of telephone speech. IEEE Trans. Speech Audio Process. 4 (1), 31–44.

Chapitre

5 Sélection, Extraction et Fusion de données

Sommaire

Résumé

Ce chapitre donne un aperçu rapide des différentes méthodes de sélection, d'extraction et de fusion de données, telle que l'Analyse en Composante Principale (ACP), l'Analyse Linéaire Discriminante et la recherche par Sélection Séquentiel (SFS ou SBS), …etc. Les méthodes décrites sont utilisées dans ce qui suit en commencant par les premières expérimentations réalisées sur la base BDSONS.

5.1 Introduction

L'objectif de l'extraction des caractéristiques est d'estimer les paramètres représentants la spécificité du locuteur, qui est la combinaison des différences physiques du système vocal et habitudes linguistiques ainsi que du style de parler. En conséquence, les informations spécifiques au locuteur contenues dans le signal de parole peuvent être classées en deux catégories : 1) l'information de bas niveau, qui est liée à la structure anatomique de l'appareil vocal et 2) l'information de haut niveau, qui est liée aux habitudes linguistiques et aux styles. Le tableau 5.1 présente les caractéristiques hiérarchiques pour la reconnaissance du locuteur par l'homme et la machine.

Tableau 5.1 Caractéristiques hiérarchiques pour la reconnaissance du locuteur par l'homme et la machine.

Affiliation physique/Sociale	Indices perceptifs pour l'homme	Caractéristiques pour la RAL	Faisabilité dans la RAL
Status socio-économique, éducation, lieu de naissance, …etc.	Accent, diction, sémantique, idiosynchrasies, …etc.	Mots, phrase, syntaxe, …etc.	Caractéristiques de haut-niveau, la représentation effective est en attente
Type de personnalité, l'influence parentale, …etc.	Style de parler, prosodie, rythme, intonation, modulation du volume, débit de parole, …etc.	Contour de F_0, les fluctuations d'énergie, les pauses, les durées, …etc.	Caractéristiques de niveau modéré, elles sont utilisées pour compléter les caractéristiques bas-niveau
Structure anatomique de l'appareil vocal	Les aspects acoustiques de la parole, les nasales, profondeur, souffle, raideur, la dureté, …etc.	F_0, les harmoniques de l'enveloppe spectrale, énergie, …etc.	Caractéristiques bas-niveau, largement et effectivement dans les systèmes RAL actuels.

Les caractéristiques acoustiques représentants les informations bas-niveau ont été largement appliquées dans la reconnaissance de la parole et la reconnaissance du locuteur. Ces caractéristiques bas-niveau révèlent les configurations du conduit vocal liées à la parole/au locuteur. Les caractéristiques bas-niveau les plus répandues sont celles basées sur l'analyse cepstrale de la parole, tels que les Coefficients Cepstraux issus de la Prédiction Linéaire (LPCC) (Atal, 1974 ; Furui, 1981) et les Coefficients Cepstraux à Fréquence Mel (MFCC) (Davis et Mermelstein, 1980). D'autres caractéristiques qui visent à capter les informations liées aux vibrations des cordes vocales, telles que la fréquence fondamentale (F_0) (Atal, 1972 ; Harrag et al., 2005 ; Sonmez et al., 1998) et les informations de l'intensité des harmoniques (Imprel et al.,

1997) ont été utilisées. Contrairement à la reconnaissance de la parole, où l'on pense que les différences sont principalement liées à la structure des formants du système du conduit vocal, dans la reconnaissance du locuteur un certain nombre d'expérimentations ont montré que le style de vibration des cordes vocales comporte des informations riches en spécificité du locuteur et sont utiles pour sa reconnaissance. Cette thèse se concentre sur le développement de techniques efficaces pour extraire l'information spécifique au locuteur issue de l'excitation de la source vocale pour améliorer la performance du système de reconnaissance du locuteur conventionnel utilisant uniquement les caractéristiques du conduit vocal.

La sélection est la transformation des vecteurs de caractéristiques pour réduire la dimension tout en conservant les informations pertinentes. Ceci est particulièrement nécessaire pour des applications réelles où les données d'apprentissage disponibles sont généralement limitées. Une technique utile pour réduire la dimension du vecteur des caractéristiques est l'Analyse en Composantes Principales (ACP) (Jollife, 2002). Dans l'ACP, le vecteur de caractéristiques original est transformé dans un autre espace de représentation avec des coordonnées orthogonales. La sélection des caractéristiques est basée sur les vecteurs propres de la matrice de covariance des données fournies. En fait, les composantes dans l'espace orthogonal correspondent aux plus grandes valeurs propres restantes, tandis que celles correspondantes aux petites valeurs propres sont rejetées. Ainsi, les vecteurs de caractéristiques transformés retiennent les informations les plus importantes, donnant une représentation optimale des caractéristiques originales. En outre, l'orthogonalité entre les composantes des caractéristiques est particulièrement adaptée pour la modélisation des données par distribution gaussienne multi-variables à covariance diagonale, qui est une hypothèse nécessaire à la modélisation des données et à l'estimation des paramètres.

Une autre technique largement utilisée et l'Analyse Discriminante Linéaire (ADL)(Mclachlan, 1992). La sélection des caractéristiques par ADL est basée sur un critère discriminant. Seulement les composantes des caractéristiques avec les plus larges variations inter-classes et des faibles variations intra-classes sont retenues. Ces critères discriminants sont particulièrement compatibles avec la reconnaissance du locuteur, qui est un problème de discrimination plutôt qu'un problème de représentation. Un certain nombre d'articles ont démontré l'application de la méthode ACP et l'ADL pour la sélection des caractéristiques dans les domaines de la reconnaissance de la parole et la reconnaissance du locuteur (Jin et Waibel, 2000 ; Thyes et al., 2000).

5.2 Sélection et Extraction des caractéristiques

Les front-ends des systèmes de reconnaissance de la parole et du locuteur utilisent des caractéristiques spectrales court-terme, car non seulement elles sont porteuses de la distribution fréquentielle qui permet d'identifier les sons, mais aussi les informations liées à la source glottale et à la forme et la longueur du conduit vocal, qui sont des informations spécifiques au locuteur. Selon le front-end des caractéristiques concaténées, les vecteurs de caractéristiques qui en résultent peuvent avoir une dimension de 20 à 50 paramètres.

Dans des applications temps réelles utilisant des dispositifs à faibles ressources, par exemple, les services d'accès par téléphone portable ou des dispositifs embarqués avec de faibles tailles de stockage et de faibles capacités de calcul, un vecteur de caractéristiques avec 50 paramètres ne semble pas approprié, ce qui nécessite une réduction du jeu des paramètres.

Le problème de l'extraction des caractéristiques est parfois établi comme une transformation linéaire qui projette les vecteurs de caractéristiques sur le sous-espace transformé défini par les directions concernées. Etant donné un vecteur de caractéristiques X d'une dimension D, une matrice K*D est appliquée pour obtenir un vecteur Y des caractéristiques transformées de dimension K (K<D). La matrice est estimée de sorte que, du point de vue de la classification, la redondance est supprimée et les caractéristiques transformées ne retiennent que les informations pertinentes, ce qui a pour effet, en théorie, d'optimiser les performances pour les valeurs cibles de K, et devrait surpasser les performances des caractéristiques de base, vu qu'on a supprimé les éléments nuisibles ou qui prêtent à confusion et, plus probablement, de mieux estimer le modèle de paramètres (plus robuste). Plusieurs méthodes d'extraction sont discutées dans la littérature de la reconnaissance des formes, parmi elles:

5.2.1 La recherche exhaustive

La recherche exhaustive est une méthode optimale pour la sélection d'un sous-ensemble de caractéristiques d'une dimension k parmis l'ensemble de caractéristiques de dimension plus grande K. La recherche exhaustive considère toutes les combinaisons possibles de (k, K). Une implémentation de ce type de recherche nécessite une énorme quantité de calcul, à savoir :

$$\binom{K}{k} = \frac{K!}{k!(K-k)!} \textbf{ recherches} \qquad (5.1)$$

Par exemple, pour k = 20 et K = 50, le nombre de recherches est d'environ 4.712×10^{13}. Par conséquent, il est nécessaire de trouver d'autres procédures plus efficaces pour éviter ce type de recherche.

5.2.2 Meilleure Caractéristique Individuelle

Connue en anglais sous le nom de « Best Individual Feature » (BIF), la performance de classification de chaque caractéristique est calculée séparément, c'est-à-dire, sur une base individuelle, et les caractéristiques donnant lieu au plus haut taux de reconnaissance sont sélectionnées. Le meilleur sous-ensemble de caractéristiques k est composé des meilleurs éléments k considéré un à un. Cependant, un ensemble des meilleures caractéristiques k prise une à une n'est pas forcément le meilleur ensemble de caractéristiques k.

5.2.3 Algorithme de recherche séquentielle

Le célèbre algorithme connu en anglais sous le nom « Sequential Forward Search » (SFS) et son homomogue (SBS) (B pour backword) (Withney, 1997) sont des méthodes qui obtiennent une chaine de sous-ensembles de caractéristiques imbriquées d'une manière directe, soit par l'addition (soustraction dans le cas du SBS) de la meilleure (la mauvaise) caractéristique dans l'ensemble. Cet effet de nidification constitue un des principaux inconvénients de ces méthodes. Les deux algorithmes ne peuvent corriger des additions (soustractions) précédentes de caractéristiques.

Dans la méthode SFS, les caractéristiques sont sélectionnées successivement en ajoutant la meilleure caractéristique locale, la caractéristique qui offre la meilleure information discriminante incrémentale au sous-ensemble des caractéristiques existantes. La technique SFS agit comme la technique BIF en identifiant la première caractéristique ayant le plus haut pouvoir discriminant. Elle procède toutefois, en ajoutant successivement au sous-ensemble sélectionné, les éléments qui contribuent le plus à la performance de classification au-dessus de ceux déjà sélectionnés. Ainsi, à partir d'une caractéristique singulière BIF, le sous ensemble SFS passe une paire, puis un triplet et ainsi de suite.

5.2.4 Les algorithmes génétiques

L'Algorithme Génétique (AG)(Holland, 1975) constitue une autre et nouvelle approche pour la recherche des caractéristiques, car elle permet une recherche aléatoire guidée par une mesure de fitness. Les AGs sont une classe de méthodes de recherche profondément inspirée du processus naturel d'évolution. A chaque itération de l'algorithme (correspondant à une génération), un nombre fixe (population) de solutions possibles (chromosomes) est générée par l'application de certains opérateurs génétiques dans un processus stochastique guidé par une mesure de fitness. Les plus importants opérateurs génétiques et couramment utilisés sont la recombinaison, le croisement et la mutation. Le résultat est un algorithme probabiliste qui a obtenu de bonnes, presque optimales, solutions aux problèmes dans lesquels les méthodes classiques ont échoué ou ne sont pas applicables. Un AG particulier est identifié par une forme particulière du codage des solutions en chaines d'un certain alphabet (généralement binaire), une forme particulière des opérateurs génétiques adoptés et une définition particulière de la fonction de fitness ou fonction d'objectif.

5.2.5 Analyse en composantes principales

L'Analyse en Composantes Principales (ACP) est une technique ancienne de l'analyse statistique multi-variable (Jolliffe, 2002). Dans cette méthode, la réduction de la dimension est obtenue par la projection de l'espace d'origine caractérisé par D paramètres vers un espace caractérisé par un sous ensemble caractérisé par d paramètres et qui préserve au mieux l'information contenue dans l'espace d'origine. La première composante principale est la projection des données dans la direction de la

plus grande variance qui contient le plus d'information. Cependant, cette méthode souffre de plusieurs problèmes décrits ci-dessous :

- Les composantes principales sont variables suivant l'échelle des paramètres. Si un paramètre dans le vecteur d'entrée est multiplié par une grande constante, sa variance sera très grande et par conséquent la première composante sera approchée par ce paramètre.
- L'ACP est basée sur la plus grande matrice de covariance des données, c'est-à-dire la matrice est calculée sans avoir d'information concernant les différentes classes existantes.

Les composantes principales définies par les variations maximales n'impliquent pas que ces dernières contiennent plus d'information qui aide à la discrimination des classes. La Figure 5.1 explique ce dernier problème. On voit bien que la direction de la séparation maximale des classes est perpendiculaire à la première composante principale.

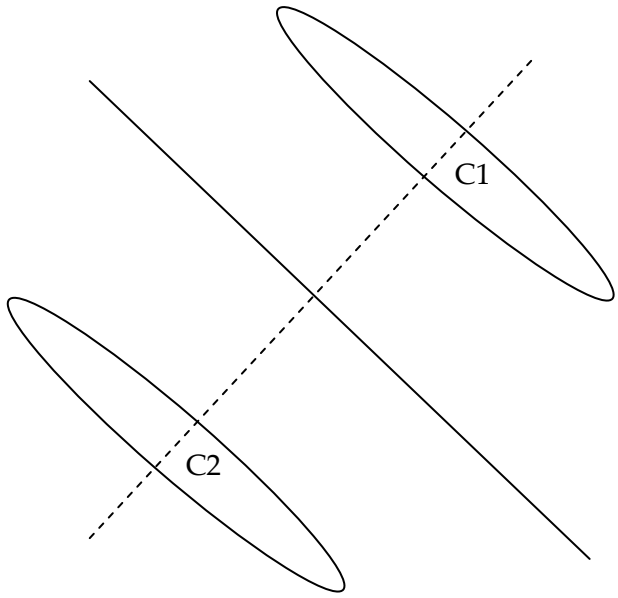


Figure 5.1 Classes Gaussiennes avec des matrices de covariance égales.

5.2.6 L'analyse discriminante linéaire

L'Analyse Discriminante Linéaire (ADL) est une technique statistique utilisée dans la reconnaissance des formes (Duda et al., 2000). Supposons que l'espace acoustique est réparti en un ensemble de classes, chaque classe étant représentée par une gaussienne de matrice de covariance W. Supposons que toutes les classes aient la même matrice de covariance. La technique (ADL) consiste à appliquer une transformation (rotation + dilatation) à tout l'espace acoustique de manière à rendre la variance intra-classe égale

à l'identité. En supposant que les nouveaux centres des classes soient distribués selon une loi gaussienne de matrice de covariance inter-claases B, une deuxième rotation est appliquée à l'espace acoustique de manière à dégager les axes principaux. Ces axes sont les vecteurs propres de la matrice B. Ainsi un sous-ensemble de ces axes correspondant aux directions des grandes variances (les grandes valeurs propres) est utilisé pour former les nouveaux vecteurs de paramétrisation du signal.

Elle est équivalente à l'obtention d'une transformation linéaire qui maximise les mesures de discrimination des classes J_1 et J_2.

$$\begin{cases} \mathbf{J_1 = tr(W^{-1}.B)} \\ \mathbf{J_2 = (W^{-1}.B)} \end{cases} \quad \textbf{(5.2)}$$

Plusieurs groupes de chercheurs (Aubert, 1993 ; Siohan, 1995) ont montré que l'analyse discriminante linéaire permettait d'améliorer les performances des systèmes de reconnaissance ainsi que leur robustesse à certains types de bruits. On peut montrer que l'optimisation des critères J_1 et J_2 conduit aux mêmes paramètres discriminants et que cette optimisation est indépendante du choix des matrices de variance.

L'analyse discriminante linéaire peut être résumée comme une procédure en deux phases : 1) dans la première phase, une fonction de normalisation dépendante des classes recueille les informations statistiques, 2) et dans la phase deux, une fonction de discrimination est dérivée des classes afin que les éléments résultants par (ADL) sont moins corrélés, et classés en fonction d'un critère d'objectif.

5.2.6.1 Principaux points de l'ADL

Nous définissons B la matrice de dispersion entre les classes (ou la matrice de covariance, en anglais between class scatter) et la matrice intra-classes W pour le problème de reconnaissance de formes de M classes. B indique l'écart entre les vecteurs attendus pour chaque paire de classe, tandis que W donne la dispersion des échantillons autour du vecteur attendu de leur propre classe. On essaye de minimiser $\mathrm{Det}(W)$ et de maximiser $\mathrm{Det}(B)$ simultanément, c'est-à-dire :

$$\mathbf{Max}\frac{|B|}{|W|} = \mathbf{Max}\left|\mathbf{W^{-1}.B}\right| \quad \textbf{(5.3)}$$

Une mesure de séparabilité de classe bien connue est la trace de la matrice de discrimination $W^{-1}.B$ définie par :

$$\mathbf{T = tr(W^{-1}.B)} \quad \textbf{(5.4)}$$

L'objectif est de sélectionner un nombre réduit de caractéristiques m (m < n) en appliquant une matrice de transformation A au vecteur original à n dimensions. Le choix de la matrice A est tel T(m) de l'espace transformé de dimension m est maximisé. Il peut être démontré que ceci est réalisé en sélectionnant les m premiers vecteurs propres de la matrice de discrimination, dont les vecteurs propres sont :

$$\emptyset_i, i = 1, 2, \dots, k, \dots, n \tag{5.5}$$

et sont rangés par la domination de leurs valeurs propres :

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n \tag{5.6}$$

et A est définie par :

$$A = [\emptyset_1, \emptyset_2, \dots, \emptyset_m]^T \tag{5.7}$$

Le nouveau vecteur de caractéristiques est donné par :

$$y = A.x \tag{5.8}$$

Les nouvelles matrices intra-classes et entre classes $\tilde{W}^{-1}$ et $\tilde{B}$ pour les nouveaux échantillons sont diagonales, ce qui signifie que les coefficients ne sont pas corrélés et peuvent être calculés comme suit :

$$\begin{cases} \tilde{B} = A.B.A^T \\ \tilde{W} = A.W^{-1}.A^T \end{cases} \tag{5.9}$$

5.3 Analyse et sélection des paramètres

Il est difficile de juger a priori de l'efficacité des paramètres issus d'une analyse. Une première possibilité est de comparer des taux de reconnaissance obtenus avec un système de classification commun aux différents paramètres. Mais, dans ce cas, les résultats peuvent être biaisés par l'adéquation du système de classification à l'un ou l'autre des paramètres. Une autre possibilité est d'estimer le degré de séparation des locuteurs dans l'espace de paramètres.

5.3.1 Sélection par F-ratio

Une mesure qui peut être utilisée pour évaluer un paramètre particulier est le F-ratio (Campbell, 1997). Il est défini comme le rapport entre variance inter-classe et variance intra-classe.

$$F - ratio = \frac{\sigma^2_{inter}}{\sigma^2_{intra}} \tag{5.10}$$

Dans le cadre de sélection des paramètres qui caractérisent au mieux les locuteurs, le F-ratio peut être utilisé pour sélectionner les paramètres qui maximisent la séparation des classes de locuteurs et minimisent la dispersion à l'intérieur de chaque classe. En utilisant le F-ratio, les assomptions suivantes sont faites :

- Le vecteur des paramètres dans chaque classe a une distribution Gaussienne.
- Les paramètres sont non corrélés.
- Les variances dans chaque classe sont égales.

En pratique, les conditions précédentes sont rarement satisfaites, et dans ce cas on ne peut pas évaluer plus d'un seul paramètre à la fois, vu que généralement dans la parole, les paramètres sont souvent corrélés. Pour remédier à ce dernier problème, on peut transformer l'espace des paramètres corrélés en un espace de paramètres non corrélés et ceci en utilisant par exemple une analyse en composantes principales.

5.3.2 Sélection basée sur les performances de reconnaissance

Cette méthode consiste à calculer la contribution de chaque paramètre dans les performances du système de reconnaissance, c'est-à-dire le taux d'erreur, et à utiliser cette dernière pour sélectionner ce paramètre ou non. Paliwal (Paliwal, 92) a utilisé chaque paramètre pour reconnaître toutes les réalisations du corpus d'apprentissage (Tableau 5.2).

Tableau 5.2 Coefficients LPCC rangés en utilisant plusieurs critères.

F-ratio		Taux d'erreur	
1 à 19	20 à 38	1 à 19	20 à 38
C_2	Δ^2C_1	ΔE	C_8
ΔE	ΔC_9	Δ^{2E}	Δ^2C_6
C_1	ΔC_{11}	C_4	Δ^2C_4
C_3	ΔC_8	C_6	Δ^2C_1
C_4	ΔC_{10}	C_1	C_9
C_{10}	Δ^2C_3	ΔC_4	ΔC_8
C_{11}	ΔC_6	C_5	Δ^2C_5
ΔC_2	Δ^2C_2	ΔC_6	Δ^2C_7
C_5	ΔC_7	ΔC_5	Δ^2C_9
C_8	ΔC_{12}	C_2	ΔC_{10}
ΔC_1	Δ^2C_4	ΔC_3	Δ^2C_8
ΔC_3	Δ^2C_{10}	C_3	C_{11}
C_9	Δ^2C_9	C_7	ΔC_{11}
Δ^2E	Δ^2C_5	ΔC_2	C_{10}
C_7	Δ^2C_{11}	ΔC_1	Δ^2C_{10}
C_6	Δ^2C_8	ΔC_7	C_{12}
C_{12}	Δ^2C_7	ΔC_9	ΔC_{17}
ΔC_5	Δ^2C_6	Δ^2C_3	Δ^2C_{12}
ΔC_4	Δ^2C_{12}	Δ^2C_2	Δ^2C_{12}

Cela est équivalent à faire tourner le système de reconnaissance N fois, avec N le nombre de paramètres dans l'espace d'origine, puis ranger les paramètres selon leurs performances et ne maintenir qu'un sous ensemble qui contient les premiers coefficients qui correspondent à ceux qui ont un faible taux d'erreur. Cette méthode a l'avantage de considérer les paramètres individuellement ce qui assure l'indépendance (non corrélation des paramètres). Cependant, les résultats obtenus peuvent dépendre du système de reconnaissance utilisé et non pas des paramètres choisis.

5.4 Conclusion

Dans ce qui suit nous avons utilisé les deux approches pour la sélection des paramètres, à savoir : 1) l'approche en mesurant la pertinence des paramètres par F-ratio et 2) l'approche par mesure de des performances de classification c'est-à-dire une mesure du taux de reconnaissance. Pour les méthodes de sélection des caractéristiques, nous avons utilisé les méthodes de l'Analyse en Composante Principale (ACP), la Sélection Séquentielle Directe (SFS en anglais) et l'Analyse Discriminante Linéaire (ADL).

5.5 Bibliographie

(Atal, 1974) B. S. Atal. Efectiveness of linear prediction characteristics of the speech wave for automatic speaker identication and verication. J. Acoust. Soc. Am., 55(6):1304-1312, 1974.

(Atal, 1972) B. S. Atal. Automatic speaker recognition based on pitch contours. J. Acoust. Soc. Am., 52:1687-1697, 1972.

(Aubert, 1993) X. Aubert & al. Improvement in connected digit recognition using linear discriminant analysis and mixture densities. Proceedings of the ICASSP, vol. 2, pp. 648-651, 1993.

(Campbell, 1997) J. P. Campbell, "Speaker Recognition: A Tutorial", in Proceedings of the IEEE, 85(9)(1997), pp. 1437–1462.

(Davis et Mermelstein, 1980) S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech, Signal Processing, 28(4):357-366, 1980.

(Duda, 2000) Duda R.O., Hart P.E. and Stork D.G.(2000), "Pattern Classification", Wiley Interscience.

(Furui, 1981) S. Furui. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust., Speech, Signal Processing, ASSP-29(2):254-272, 1981.

(Harrag et al., 2005) Harrag A., Mohamadi A., Serignat J.F. (2005), "LDA Combination of Pitch and MFCC Features in Speaker Recognition", INDICON, Chennai, India, 11–13, 237–240.

(Holland, 1975) Holland J. (1975), "Adaptation in Natural and Artificial Systems", University of Michigan Press, Ann Arbor, MI.

(Imprel et al., 1997) B. Imperl, Z. Kacic, and B. Horvat. A study of harmonic features for speaker recognition. Speech Communication, 22(4):385-402, 1997.

(Jin et Waibel, 2000) Q. Jin and A. Waibel. Application of LDA to speaker recognition. In Proc. Int. Conf. on Spoken Language Processing (ICSLP), 2000.

(Jollife, 2002) I. T. Jolliffe. Principal Component Analysis. New York : Springer-Verlag, 2002.

(Mclachlan, 1992) G. J. McLachlan. Discriminant Analysis and Statistical Pattern Recognition. New York : Wiley,, 1992.

(Paliwal, 1992) K. K. Paliwal. Dimensionality reduction of the enhanced feature set for the HMM-based speech recognizer. Digital Signal Processing, pp. 157-173, 1992.

(Siohan, 1995) O. Siohan. On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition. Proceedings of the ICASSP, pp. 125-128, 1995.

(Sonmez et al., 1998) K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling dynamic prosodic variation for speaker verification. In Proc. Int. Conf. on Spoken Language Processing (ICSLP), pages 3189-3192, 1998.

(Soong et Rosenberg, 1988) F. K. Soong & A. E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 36, no. 6, pp. 871-879, 1988.

(Thyes et al., 2000) O. Thyes, R. Kuhn, P. Nguyen, and J.-C. Junqua. Speaker identification and verification using eigenvoices. In Proc. Int. Conf. on Spoken Language Processing (ICSLP), 2000.

Chapitre

6 Expérimentation I: Base BDSONS

Sommaire

Résumé

Ce chapitre donne les premiers résultats obtenus suite aux expérimentations réalisées sur la base BDSONS. Des expérimentations pour évaluer la pertinence des paramètres prosodique (durée et F_0) et des paramètres acoustiques MFCC et leurs dérivées ainsi qu'une première série de tests de fusion des données issues des deux espaces.

6.1 Introduction

Pour évaluer le potentiel et la pertinence des paramètres bas-niveau (acoustique et prosodique), plusieurs séries d'expériences sont proposées dans ce chapitre. Le premier objectif, considéré comme essentiel, est l'apport de ces paramètres dans la caractérisation du locuteur.

La première série d'expériences est conduite sur une base de données BDSONS (Descout et al, 1986), de très bonne qualité du point de vue de la numérisation et de l'environnement d'enregistrement, mais qui présente des lacunes par son caractère générique et par peu de locuteurs et peu de réalisations par locuteur, ce qui limite grandement ces séries de tests (plus de détails sur la base BDSONS sont disponibles dans l'annexe B).

Cette première série comporte cinq expérimentations qui se divisent en quatre parties :

- Etude de la durée.
- Etude de la fréquence fondamentale F_0 (pitch)
- Etude des paramètres acoustiques MFCC, ΔMFCC et ΔΔMFCC.
- Fusion du pitch et des paramètres acoustiques par la méthode (ADL).

6.1.1 Exp 1 : La durée

Dans cette expérimentation, nous avons réalisé une première étude sur les paramètres temporels, particulièrement la durée de voisement/non voisement, la durée des mots et enfin la durée entre les débuts, les centres et les fins de chaque phonème.

Nous avons commencé notre étude par un travail sur la durée sans donner trop d'importance au contexte linguistique ou au types de sons (plosives, fricatives, voyelles, …), puis nous avons essayé d'affiner les résultats par 1) la séparation des zones voisées, zones non-voisées et pauses, et 2) la création des classes de phonèmes.

6.1.1.1 Utilisation d'un logiciel de segmentation

Dans cette partie, nous avons utilisé le logiciel de Segmentation développé par l'équipe de Régine André (Dépambour et Obrecht, 1997) pour obtenir une segmentation grossière des corpus sur lesquels nous avons travaillé. La limitation avec ce logiciel est le réglage des paramètres pour l'obtention du même nombre de segments pour le même corpus et pour les différents locuteurs, ce qui permettrait par la suite une comparaison adéquate. Cette limitation nous a poussé à abandonner l'utilisation de ce logiciel de segmentation pour ne pas avoir de résultats biaisés par le choix des paramètres (du logiciel) qui s'avère une tâche un peu délicate voir presque impossible sur cet outil.

6.1.1.2 Utilisation du logiciel WaveEdit

Cette partie a été réalisée en utilisant le logiciel WaveEdit (Akbar, 1997) avec un corpus de 3 phrases prononcées par cinq locuteurs. Les figures 6.1 à 6.3 donnent les résultats de cette étude sous forme d'histogrammes représentants les zones voisées et non voisées (Figure de gauche), les zones voisées (Figure du centre) et les zones non-voisées (Figure de droite), et cela pour les corpus 28, 29 et 30 prononcés par les locuteurs BP, FC, JO, LC et LT de la base BDSONS, respectivement.

Cette deuxième partie a fait l'objet de nombreux essais. En plus des calculs d'histogrammes ci-dessus, nous avons calculé les matrices des distances et de corrélations pour chaque type de sons (voisé et non-voisé). Les tableaux 6.1 et 6.2 donnent un exemple des matrices de distances et celles des corrélations pour des segments voisés.

Ces essais ont été freinés par 1) le manque de répétition pour les phrases utilisées (une phrase par locuteur) et 2) la non considération des pauses dans le logiciel WavEdit ce qui biaise quelque part le calcul des histogrammes. La seule possibilité restante est de comparer les différents résultats des locuteurs par rapport à leurs différentes réalisations pour voir la corrélation (Tableaux 6.1 et 6.2). Malheureusement, ces résultats ne sont pas concluants vu qu'ils sont eux-mêmes biaisés par le contenu linguistiques qui diffère entre les trois phrases utilisées.

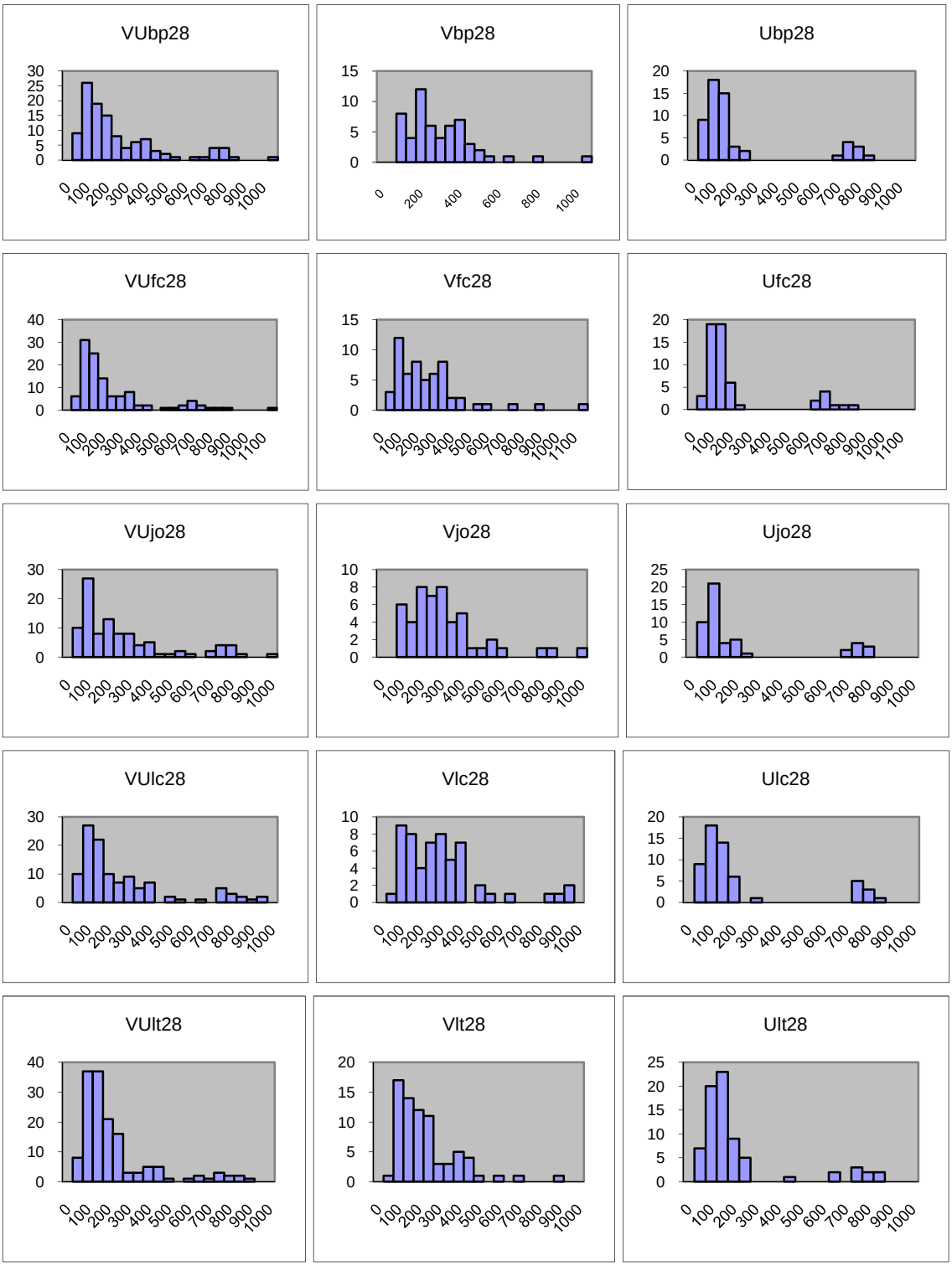


Figure 6.1 Histogrammes des durées du corpus 28 prononcé par les 5 locuteurs.

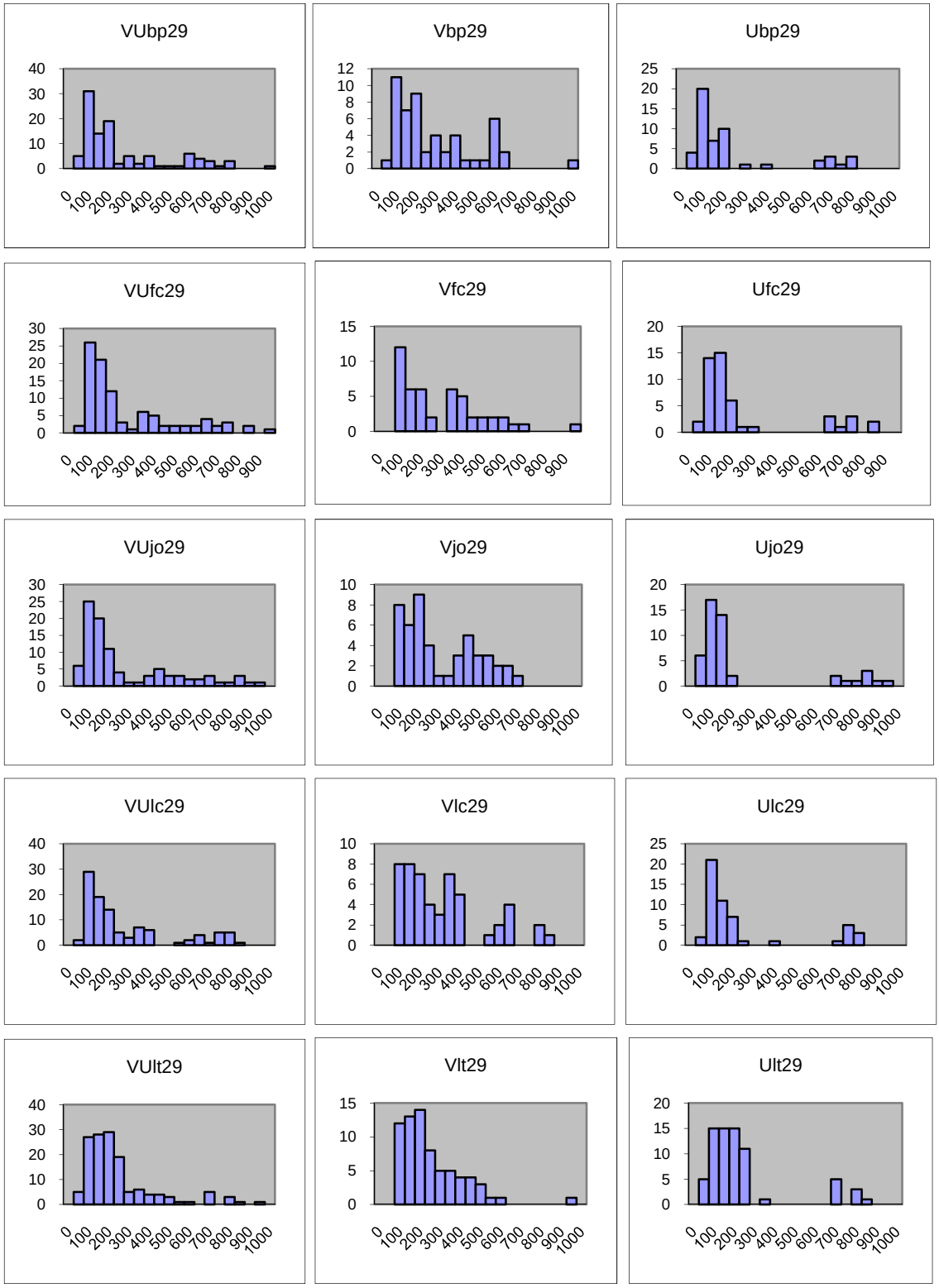


Figure 6.2 Histogrammes des durées du corpus 29 prononcé par les 5 locuteurs.

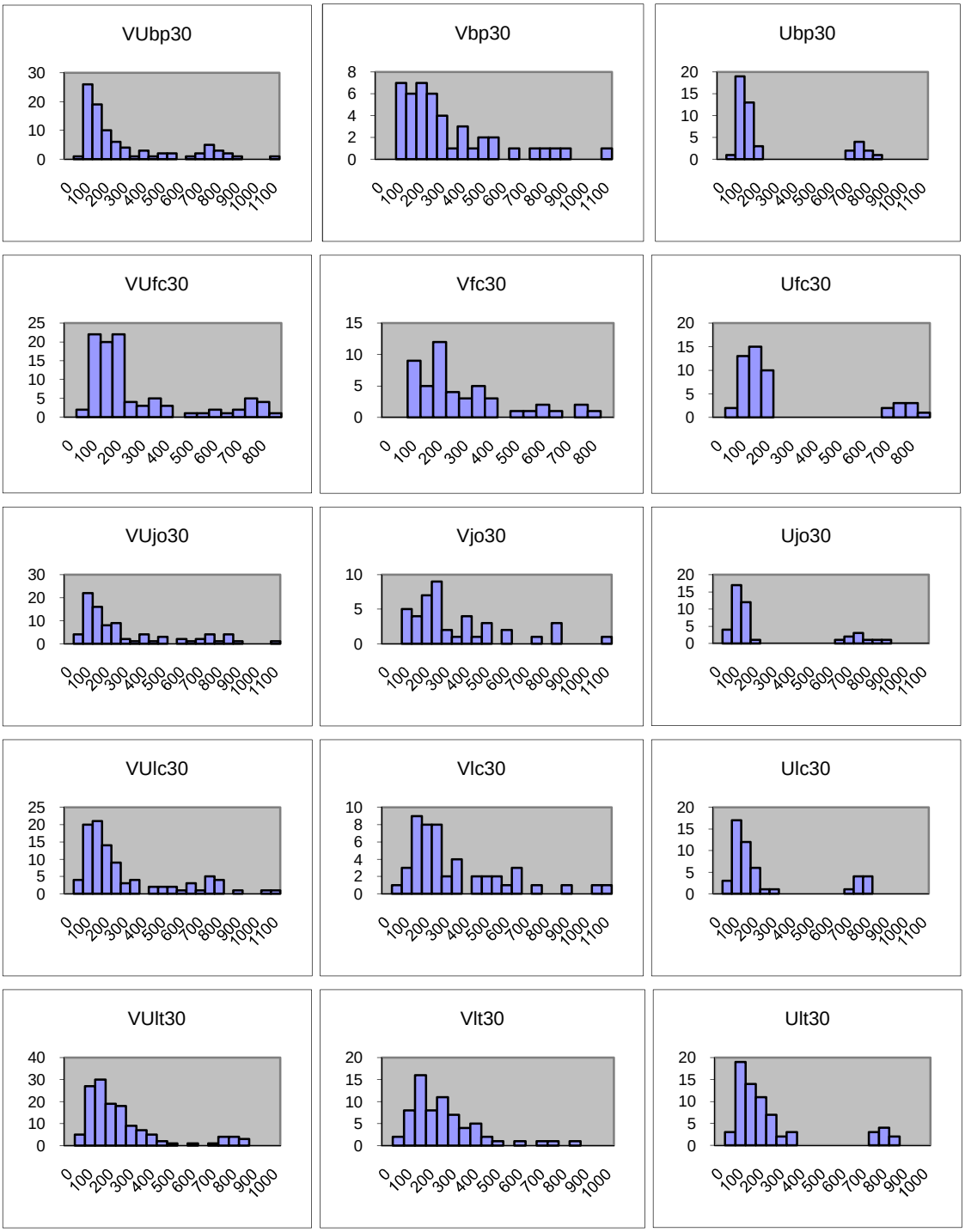


Figure 6.3 Histogrammes des durées du corpus 30 prononcé par les 5 locuteurs.

Tableau 6.1 Matrice des distances (ex. segments voisés).

	bp28	fc28	jo28	lc28	lt28	bp29	fc29	jo29	lc29	lt29	bp30	fc30	jo30	lc30	lt30
bp28	0,00	9,43	7,42	10,77	15,13	10,58	10,00	9,11	8,89	11,05	8,89	6,48	10,15	11,75	14,66
fc28	9,43	0,00	9,49	9,54	14,00	9,64	9,00	11,49	8,83	11,53	6,48	8,06	13,04	12,37	13,56
jo28	7,42	9,49	0,00	7,68	17,26	10,72	12,04	10,49	9,27	13,53	6,48	8,66	8,25	10,63	13,34
lc28	10,77	9,54	7,68	0,00	15,46	11,58	11,31	12,77	8,89	13,27	8,66	11,83	11,18	12,49	10,91
lt28	15,13	14,00	17,26	15,46	0,00	14,80	15,39	15,36	15,81	7,55	15,62	15,13	17,49	17,12	11,22
bp29	10,58	9,64	10,72	11,58	14,80	0,00	8,12	8,19	8,54	12,33	9,11	7,48	11,79	12,65	14,87
fc29	10,00	9,00	12,04	11,31	15,39	8,12	0,00	8,43	7,68	13,64	9,85	8,72	12,29	13,11	16,34
jo29	9,11	11,49	10,49	12,77	15,36	8,19	8,43	0,00	10,00	12,29	6,78	8,43	9,17	9,33	15,23
lc29	8,89	8,83	9,27	8,89	15,81	8,54	7,68	10,00	0,00	12,77	8,49	7,68	11,14	9,85	13,11
lt29	11,05	11,53	13,53	13,27	7,55	12,33	13,64	12,29	12,77	0,00	12,77	11,22	15,07	13,34	9,64
bp30	8,89	6,48	6,48	8,66	15,62	9,11	9,85	6,78	8,49	12,77	0,00	7,81	6,16	7,81	12,81
fc30	6,48	8,06	8,66	11,83	15,13	7,48	8,72	8,43	7,68	11,22	7,81	0,00	10,15	10,49	14,93
jo30	10,15	13,04	8,25	11,18	17,49	11,79	12,29	9,17	11,14	15,07	6,16	10,15	0,00	9,11	14,42
lc30	11,75	12,37	10,63	12,49	17,12	12,65	13,11	9,33	9,85	13,34	7,81	10,49	9,11	0,00	12,29
lt30	14,66	13,56	13,34	10,91	11,22	14,87	16,34	15,23	13,11	9,64	12,81	14,93	14,42	12,29	0,00

Tableau 6.2 Matrice des corrélations (ex. segments voisés).

	bp28	fc28	jo28	lc28	lt28	bp29	fc29	jo29	lc29	lt29	bp30	fc30	jo30	lc30	lt30
bp28	1,000	0,750	0,838	0,812	0,776	0,910	0,878	0,809	0,770	0,738	0,815	0,667	0,800	0,872	0,688
fc28	0,750	1,000	0,787	0,799	0,840	0,866	0,697	0,828	0,623	0,683	0,814	0,574	0,835	0,841	0,684
jo28	0,838	0,787	1,000	0,788	0,712	0,848	0,866	0,837	0,850	0,825	0,777	0,789	0,920	0,928	0,863
lc28	0,812	0,799	0,788	1,000	0,827	0,862	0,796	0,689	0,595	0,799	0,824	0,640	0,840	0,851	0,739
lt28	0,776	0,840	0,712	0,827	1,000	0,811	0,595	0,800	0,550	0,692	0,844	0,514	0,832	0,803	0,612
bp29	0,910	0,866	0,848	0,862	0,811	1,000	0,802	0,811	0,720	0,668	0,849	0,705	0,829	0,893	0,686
fc29	0,878	0,697	0,866	0,796	0,595	0,802	1,000	0,648	0,782	0,852	0,744	0,647	0,743	0,812	0,773
jo29	0,809	0,828	0,837	0,689	0,800	0,811	0,648	1,000	0,710	0,561	0,695	0,718	0,867	0,891	0,673
lc29	0,770	0,623	0,850	0,595	0,550	0,720	0,782	0,710	1,000	0,679	0,615	0,721	0,792	0,783	0,760
lt29	0,738	0,683	0,825	0,799	0,692	0,668	0,852	0,561	0,679	1,000	0,798	0,590	0,791	0,786	0,850
bp30	0,815	0,814	0,777	0,824	0,844	0,849	0,744	0,695	0,615	0,798	1,000	0,711	0,807	0,835	0,774
fc30	0,667	0,574	0,789	0,640	0,514	0,705	0,647	0,718	0,721	0,590	0,711	1,000	0,766	0,836	0,840
jo30	0,800	0,835	0,920	0,840	0,832	0,829	0,743	0,867	0,792	0,791	0,807	0,766	1,000	0,950	0,882
lc30	0,872	0,841	0,928	0,851	0,803	0,893	0,812	0,891	0,783	0,786	0,835	0,836	0,950	1,000	0,887
lt30	0,688	0,684	0,863	0,739	0,612	0,686	0,773	0,673	0,760	0,850	0,774	0,840	0,882	0,887	1,000

Malgré le caractère intéressant avec des profils d'histogrammes différents pour les différentes réalisations et les différents locuteurs, les résultats sont lissés par la valeur de la durée vu qu'on calcule des fréquences d'occurrences des différentes durées, en plus la non-prise en compte des pauses dans ces durées mesurées rend ces résultats peu fiables.

6.1.1.3 *Utilisation des fichiers d'étiquettes de la base BDSONS*

Vu les problèmes rencontrés dans la première partie, à savoir les problèmes de segmentation automatique et le problème de la non-prise en compte des pauses dans la deuxième partie, nous avons décidé d'utiliser les fichiers d'étiquettes (étiquettes manuelles) de la base BDSONS pour étudier la pertinence de la durée. Pour cette partie, nous avons utilisé le corpus texte « la bise et le soleil » prononcé par 14 locuteurs. Pour chaque locuteur, nous avons calculé les durées suivantes: début-centre, centre-fin, et début-fin. Nous avons réalisé deux séries de tests, la première sans prise en compte du contenu linguistique et une deuxième en prenant en compte le contenu linguistique, ce qui nous a conduit à créer cinq classes linguistiques (classes de phonèmes) : plosives, fricatives, nasales, voyelles orales et voyelles nasales.

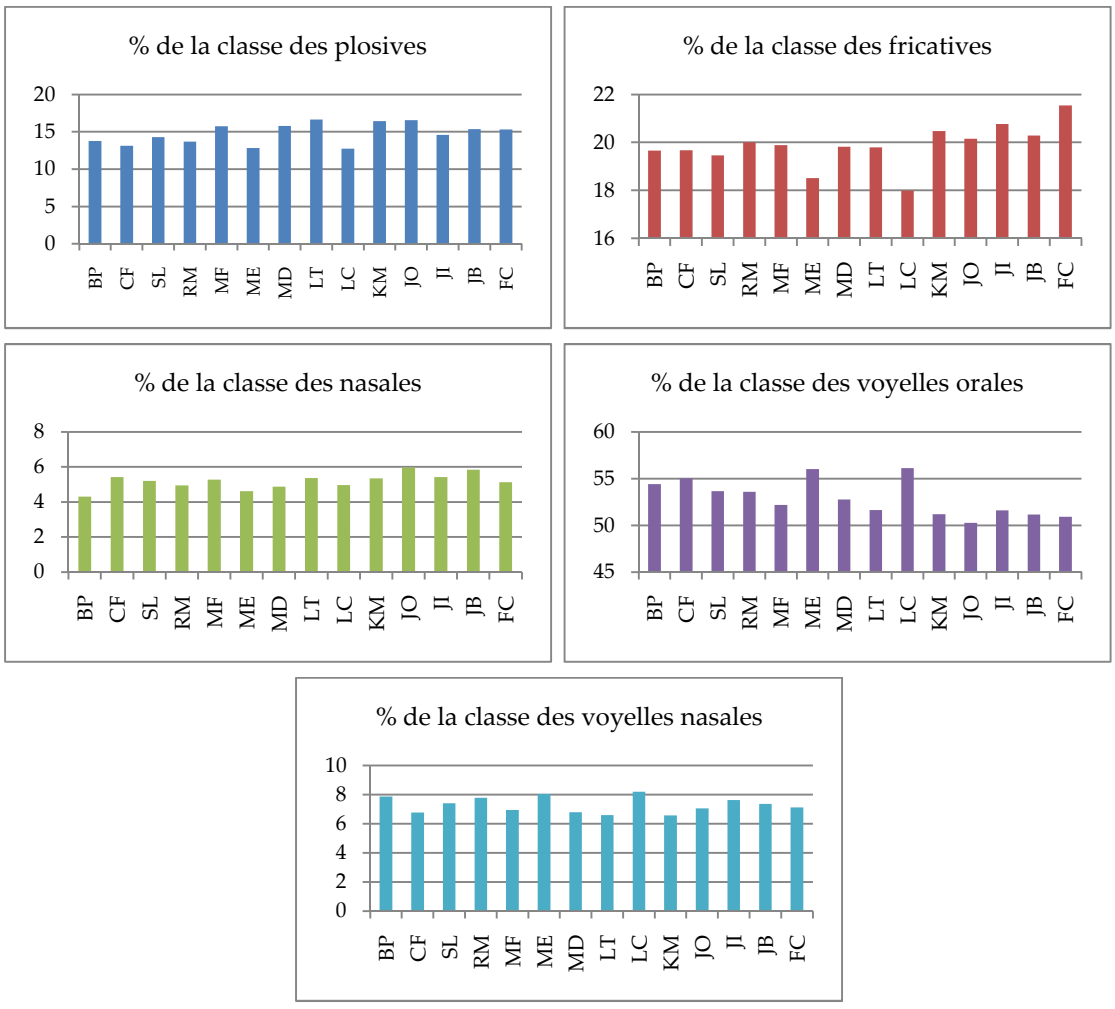


Figure 6.4 Pourcentage des différentes classes par rapport à la durée totale.

Malgré la différence de forme d'histogrammes surtout dans le cas des nasales, malheureusement, il n'y a qu'une réalisation par locuteur ce qui ne nous permet pas de calculer des distances ou des corrélations entre les différentes occurrences pour valider la différence de motifs entre les différents locuteurs.

6.1.2 Exp 2 : Pitch

Pour ce paramètre nous avons calculé la valeur moyenne et la déviation standard sur chaque réalisation de chaque voyelle pour tous les locuteurs. Les résultats obtenus sont présentés dans le tableau 6.3.

Tableau 6.3 Valeur du pitch.

	Moyenne (en Hz)					Déviation standard				
	/a/	/e/	/i/	/o/	/u/	/a/	/e/	/i/	/o/	/u/
Loc 1	100.96	104.85	108.48	104.94	107.54	1.36	3.48	0.95	1.65	2.82
	101.14	101.03	106.62	105.87	111.77	2.71	2.22	2.39	2.09	5.41
	98.52	105.74	116.20	109.89	111.62	1.54	3.81	3.17	3.07	1.90
	99.03	101.00	110.95	111.64	106.29	3.52	4.30	6.99	1.27	1.53
Loc 2	218.99	227.36	233.77	240.63	229.05	6.91	7.51	1.79	2.98	2.26
	220.74	231.62	218.94	223.78	222.69	3.63	4.38	4.82	1.71	6.74
	228.99	234.36	237.24	230.68	227.56	5.02	8.81	1.85	3.93	7.28
	223.80	231.55	226.60	237.41	226.09	9.28	3.54	3.83	7.06	3.05
Loc 3	215.69	216.33	221.56	231.27	226.79	5.50	10.29	9.36	7.61	4.45
	209.80	220.80	231.13	217.67	221.75	8.44	5.54	4.93	8.60	3.56
	211.75	222.23	219.92	224.61	221.78	6.56	10.00	9.42	3.92	9.00
	209.60	219.41	217.51	214.65	216.38	5.09	7.69	6.92	9.60	6.33
Loc 4	102.40	109.62	115.75	114.67	114.03	6.87	5.11	9.88	4.49	6.48
	108.19	113.96	112.89	111.21	113.70	5.63	5.72	5.23	6.34	5.40
	105.04	112.41	114.83	116.39	116.29	5.59	5.34	6.55	8.64	8.73
	104.00	111.71	117.50	117.45	117.73	6.85	8.55	8.00	5.42	5.21
Loc 5	231.16	248.49	253.03	250.47	264.20	7.97	10.84	12.87	15.76	11.72
	226.14	246.12	251.69	252.64	243.78	9.92	14.60	11.74	15.29	12.86
	231.83	235.93	246.80	254.38	245.36	9.02	12.71	12.69	14.93	12.31
	225.73	237.28	239.84	247.10	247.34	9.64	12.16	13.72	15.72	10.76
Loc 6	211.93	215.03	224.13	220.50	242.49	1.53	4.04	5.95	9.14	4.59
	189.97	178.54	251.53	231.39	234.83	5.57	4.68	7.91	6.77	3.11
	211.32	234.93	272.68	241.80	264.80	10.01	7.84	5.79	3.60	5.34
	198.37	222.15	240.64	223.86	255.31	10.87	8.29	3.42	8.02	6.04
Loc 7	131.44	131.16	132.49	124.36	139.91	4.69	5.04	4.29	4.98	5.39
	132.15	138.39	131.09	130.53	133.73	3.51	4.98	5.64	5.05	2.17
	129.93	130.69	132.97	130.81	127.72	2.25	3.50	4.18	4.29	4.77
	128.17	136.78	131.29	132.29	129.40	4.91	5.44	4.76	3.29	1.93
Loc 8	206.90	208.26	213.64	207.72	216.63	4.38	6.78	2.49	5.89	3.61
	206.99	216.07	210.30	208.73	209.35	1.28	5.97	3.60	6.78	3.29
	207.02	208.99	215.73	211.62	212.72	2.84	5.17	3.45	3.65	2.36
	207.86	206.60	215.86	210.35	215.45	4.05	1.40	1.02	4.92	2.64
Loc 9	112.79	116.46	125.91	123.24	122.19	3.76	4.35	2.70	1.97	2.78
	110.72	116.21	135.61	122.00	121.42	2.75	1.39	2.05	3.05	1.72
	112.50	115.86	120.86	118.27	118.32	2.68	2.95	2.28	2.43	1.95
	113.68	115.02	122.62	117.07	119.14	3.15	2.83	1.60	2.05	4.08
Loc10	255.65	260.22	268.34	268.65	266.26	4.37	4.00	3.60	5.87	3.27
	264.92	265.61	285.84	288.94	279.68	2.25	3.84	6.47	6.52	11.50
	260.29	267.20	280.06	272.44	292.04	6.01	6.11	5.88	5.45	4.80
	250.80	257.70	269.83	252.49	288.28	4.77	5.21	3.88	5.93	9.27

La figure 6.5 donne la répartition des 10 locuteurs dans l'espace moyenne-déviation standard du pitch pour les 5 voyelles /a/, /e/, /i/, /o/ et /u/.

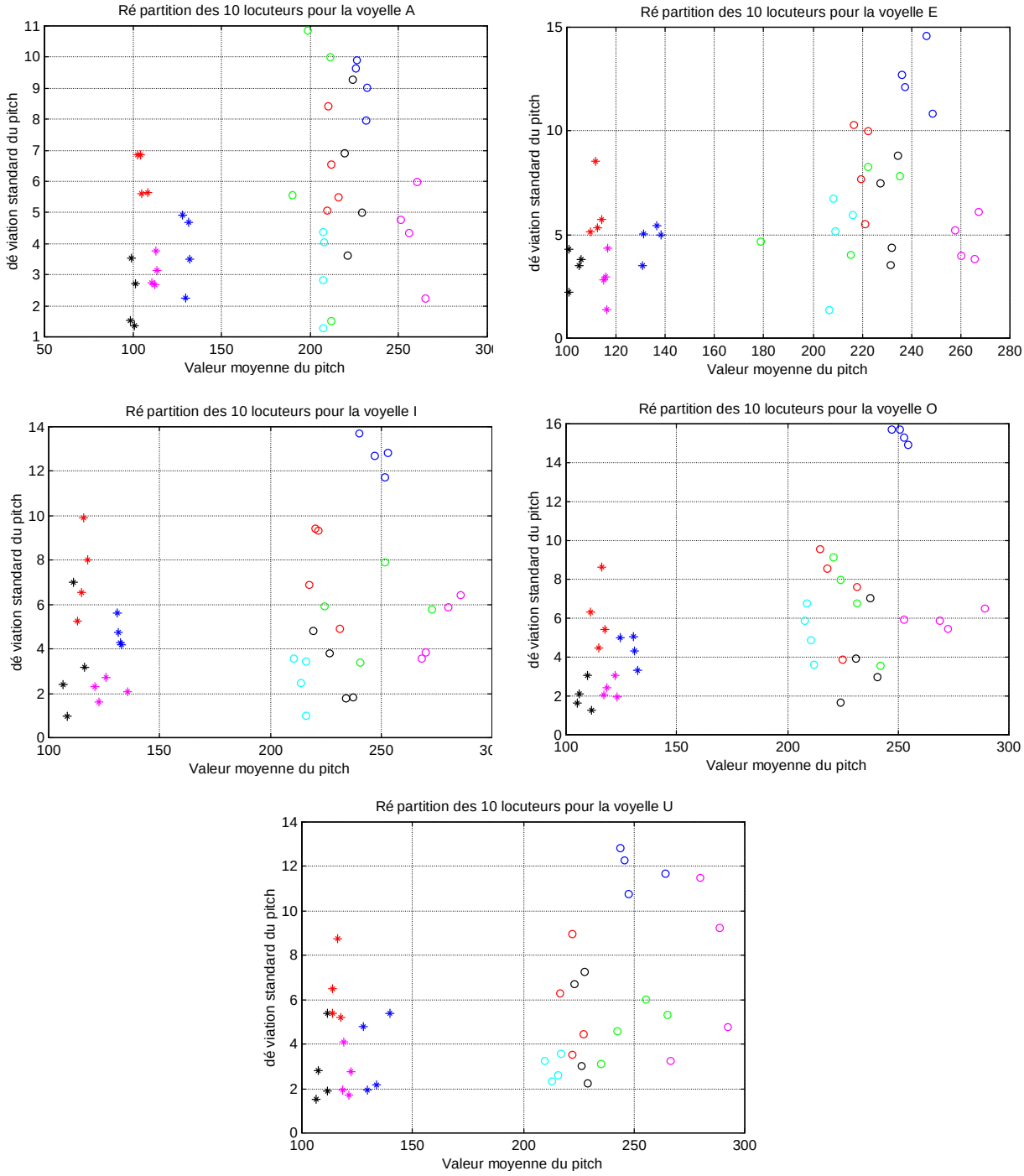


Figure 6.5 Répartition des locuteurs dans l'espace moyenne-déviation standard.

D'après les résultats obtenus, on constate que ce paramètre ne permet pas une discrimination entre les différents locuteurs, néanmoins il permet une première subdivision de l'espace en deux classes distinctes : une pour les femmes et une pour les hommes.

6.1.3 Exp 3 : Paramètres MFCC et leurs dérivées

Pour ces paramètres, nous avons utilisé comme mesure de performance le F-ratio (Campbell, 1997) défini dans le chapitre précédent. Les résultats obtenus sont présentés ci-dessous. Les tableaux 6.4 à 6.6 donnent les F-ratio pour chaque type de paramètres. Tandis que les figures 6.6 à 6.8 représentent la répartition des 10 locuteurs avec le paramètre correspondant au F-ratio minimal (Figure de gauche) et celui de F-ratio maximal (Figure de droite) et cela pour chaque voyelle.

Tableau 6.4 F-ratio pour les paramètres MFCC

	/a/	/e/	/i/	/o/	/u/
C_1	0,9824	2,4206	2,2671	1,3818	1,8293
C_2	1,1443	3,1187	5,1243	2,0883	4,6613
C_3	2,2851	2,4698	2,2491	2,6773	2,4866
C_4	1,0280	3,0492	2,3223	1,8988	1,4442
C_5	1,9000	3,2963	3,0096	2,3954	3,2915
C_6	1,1479	1,9420	2,9280	1,1503	3,5302
C_7	0,7389	3,2243	3,0492	1,2787	6,2686
C_8	0,6429	2,8752	3,8096	3,3202	4,2763
C_9	0,8217	3,0588	4,1155	6,9584	4,6085
C_{10}	0,4990	2,4458	2,2347	2,6605	6,0575

Tableau 6.5 F-ratio pour les paramètres ΔMFCC

	/a/	/e/	/i/	/o/	/u/
C_1	0,8241	1,4694	0,4654	1,7741	0,5314
C_2	6,5373	0,4126	0,2795	0,3587	0,4906
C_3	1,5078	0,4306	0,3623	0,2927	0,3994
C_4	0,7521	0,9320	0,3826	0,4750	0,4930
C_5	0,6273	0,3239	0,3754	0,2471	1,1107
C_6	0,2555	0,4150	0,4858	0,5746	0,6645
C_7	0,3934	0,6981	0,2507	0,5062	0,9488
C_8	0,3311	0,9440	0,7269	0,4474	1,2211
C_9	0,6429	0,2471	0,3635	0,4282	0,7677
C_{10}	0,3575	0,6645	0,6741	0,1080	0,6945

Tableau 6.6 F-ratio pour les paramètres ΔΔMFCC.

	/a/	/e/	/i/	/o/	/u/
C_1	0,5710	0,7989	2,6617	1,3111	1,0184
C_2	0,9776	3,2687	7,6205	2,6557	5,9040
C_3	1,3734	2,8129	3,0683	2,9796	2,6905
C_4	1,5906	2,9400	2,8212	1,5594	1,1719
C_5	3,4642	2,4482	4,8088	2,5777	3,3226
C_6	2,0524	1,5894	3,8468	1,3602	3,2819
C_7	1,3878	3,5098	3,6525	1,3099	8,5501
C_8	0,7773	2,5466	3,1763	3,4102	5,9640
C_9	0,8265	2,9772	2,4326	5,3942	3,6249
C_{10}	0,9452	2,8164	3,4258	3,2951	6,5913

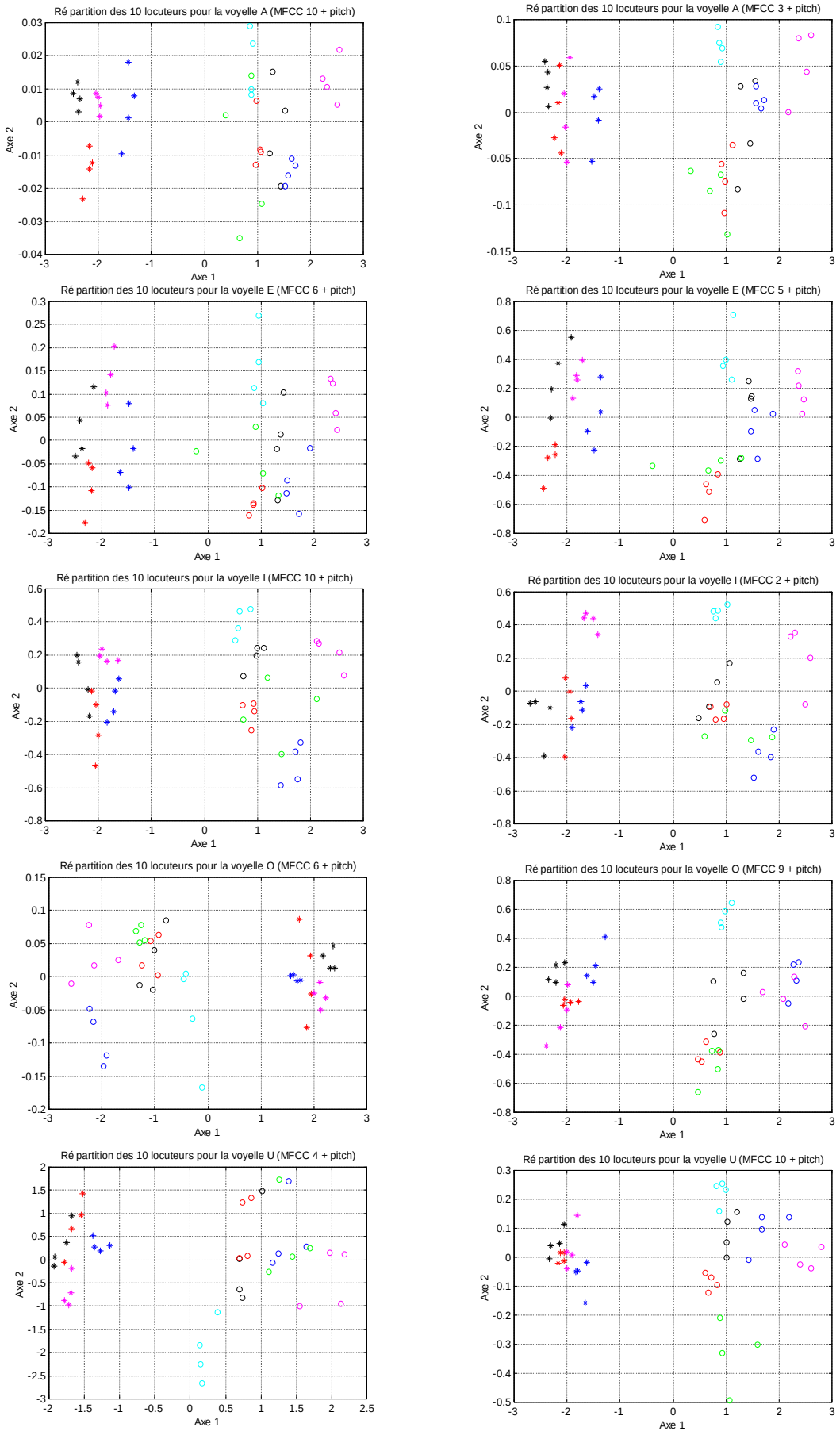


Figure 6.6 Représentation du locuteur dans l'espace pitch + MFCC.

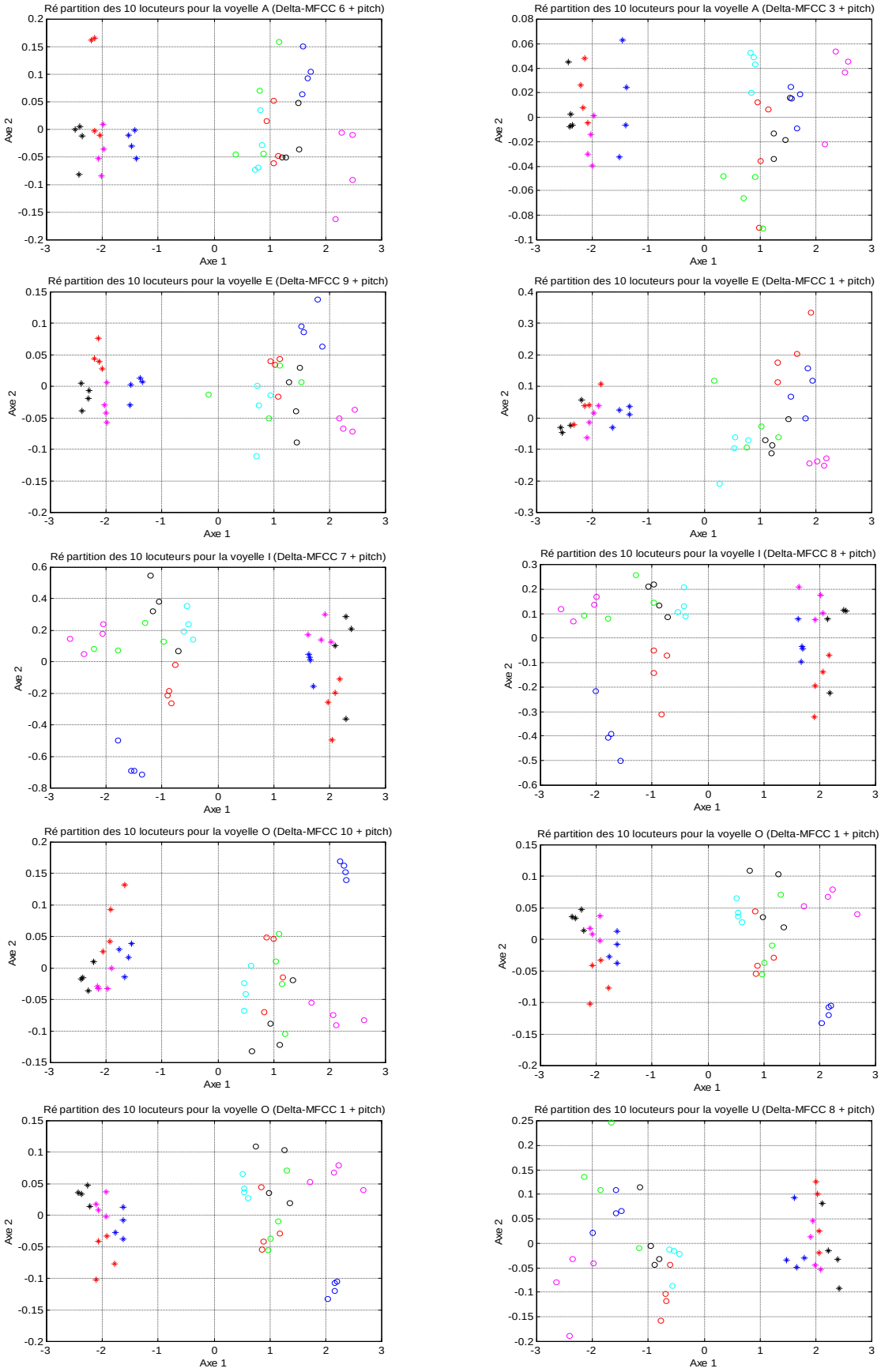


Figure 6.7 Répartition des locuteurs dans l'espace pitch + ΔMFCC.

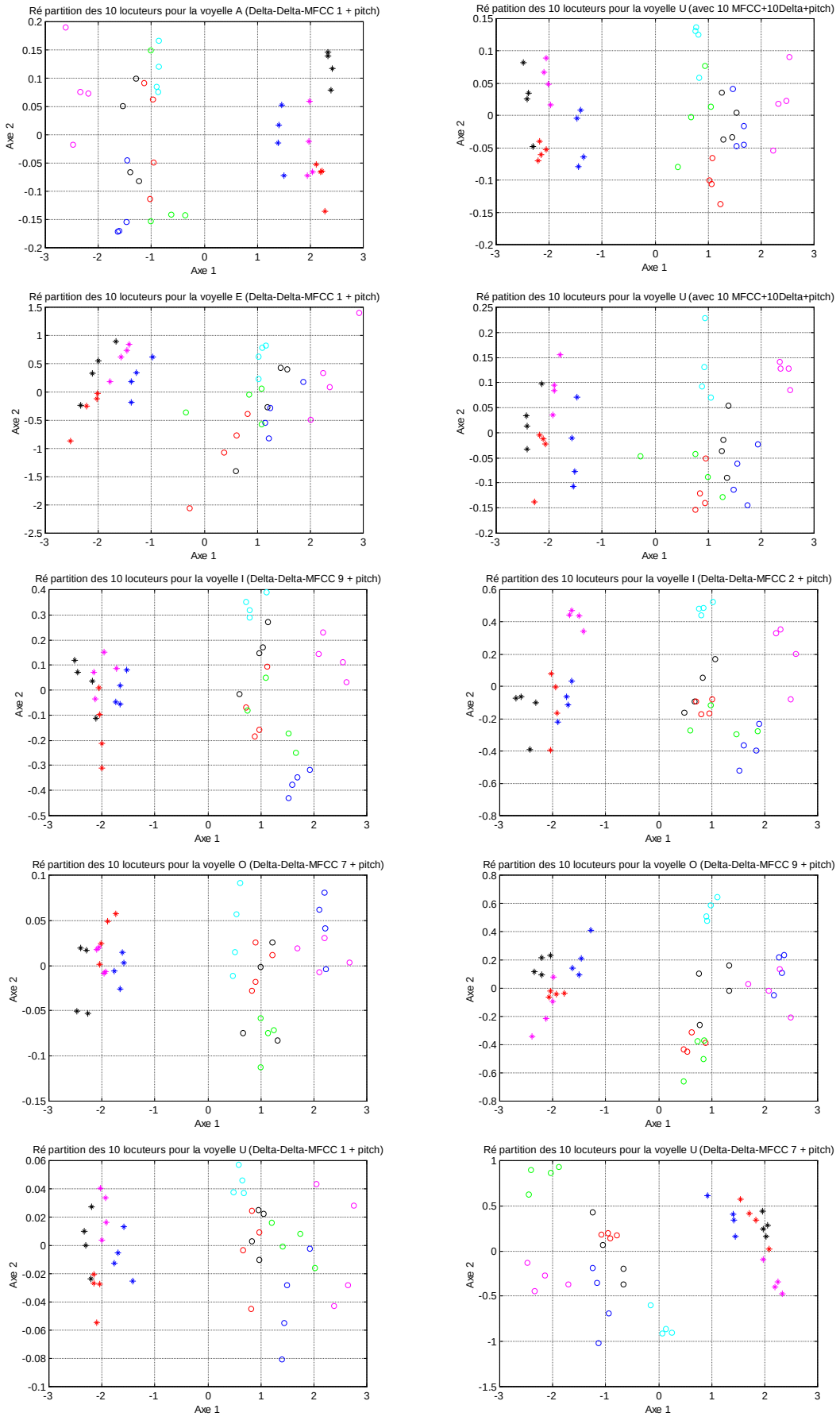


Figure 6.8 Répartition des locuteurs dans l'espace pitch + ΔΔMFCC.

6.1.4 Exp 4 : Les coefficients ADL

Nous avons essayé dans un premier temps d'utiliser une analyse en composante principale qui n'a pas donné de grands résultats. Cela est dû principalement aux problèmes de l'ACP cités dans le chapitre précédent. Pour remédier à ces derniers, nous avons fait appel à l'analyse discriminante linéaire qui permet d'extraire des coefficients qui donnent la séparation maximale des classes de locuteurs et minimisent la dispersion à l'intérieur de chaque classe. Contrairement, à la sélection par F-ratio qui évalue la performance individuelle de chaque paramètre, l'analyse ADL travaille sur des vecteurs de paramètres en faisant l'assomption que chaque nouveau paramètre apporte un plus à la classification (c'est-à-dire sans élimination des coefficients qui ont des performances médiocres).

Les figures ci-dessous donnent les résultats d'une analyse ADL faite sur différents combinaisons de paramètres :

- pitch + 5 MFCC
- pitch + 10 MFCC
- pitch + 10 MFCC + 5 ΔMFCC
- pitch + 10 MFCC + 10 ΔMFCC
- pitch + 10 MFCC + 10 ΔMFCC + 5 ΔΔMFCC
- pitch + 10 MFCC + 10 ΔMFCC + 10 ΔΔMFCC

et cela pour les 10 locuteurs et pour chaque voyelle du corpus utilisé.

D'après les résultats obtenus et les figures présentées ci-dessous, on constate que l'assomption faite ci-dessus est vérifiée, c'est-à-dire qu'à chaque fois qu'on ajoute un nouveau vecteur de paramètres on améliore la séparation entre classes et on minimise la dispersion à l'intérieure de chaque classe (Harrag et al., 2005).

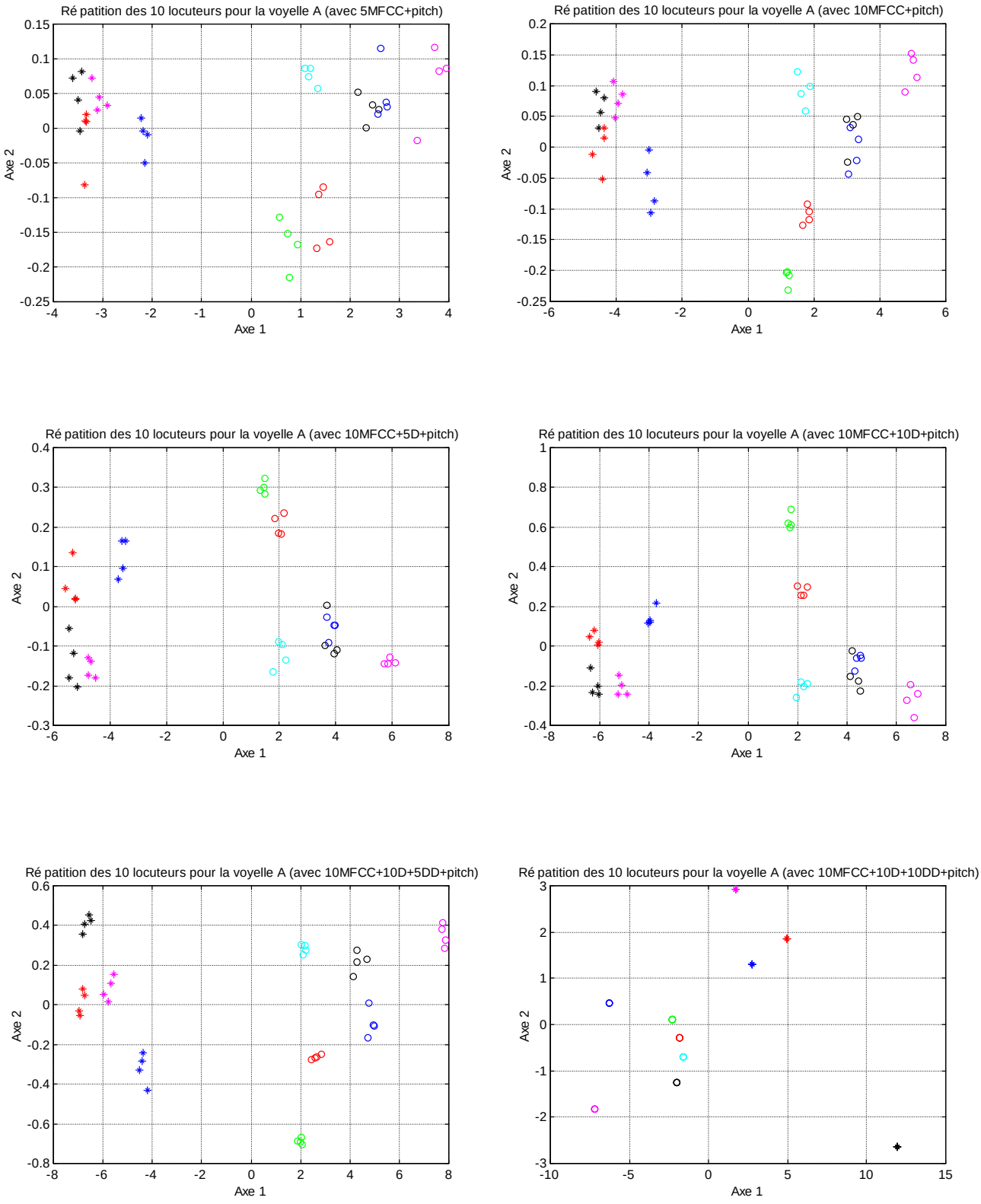


Figure 6.9 Répartition des locuteurs dans l'espace ADL pour la voyelle /a/.

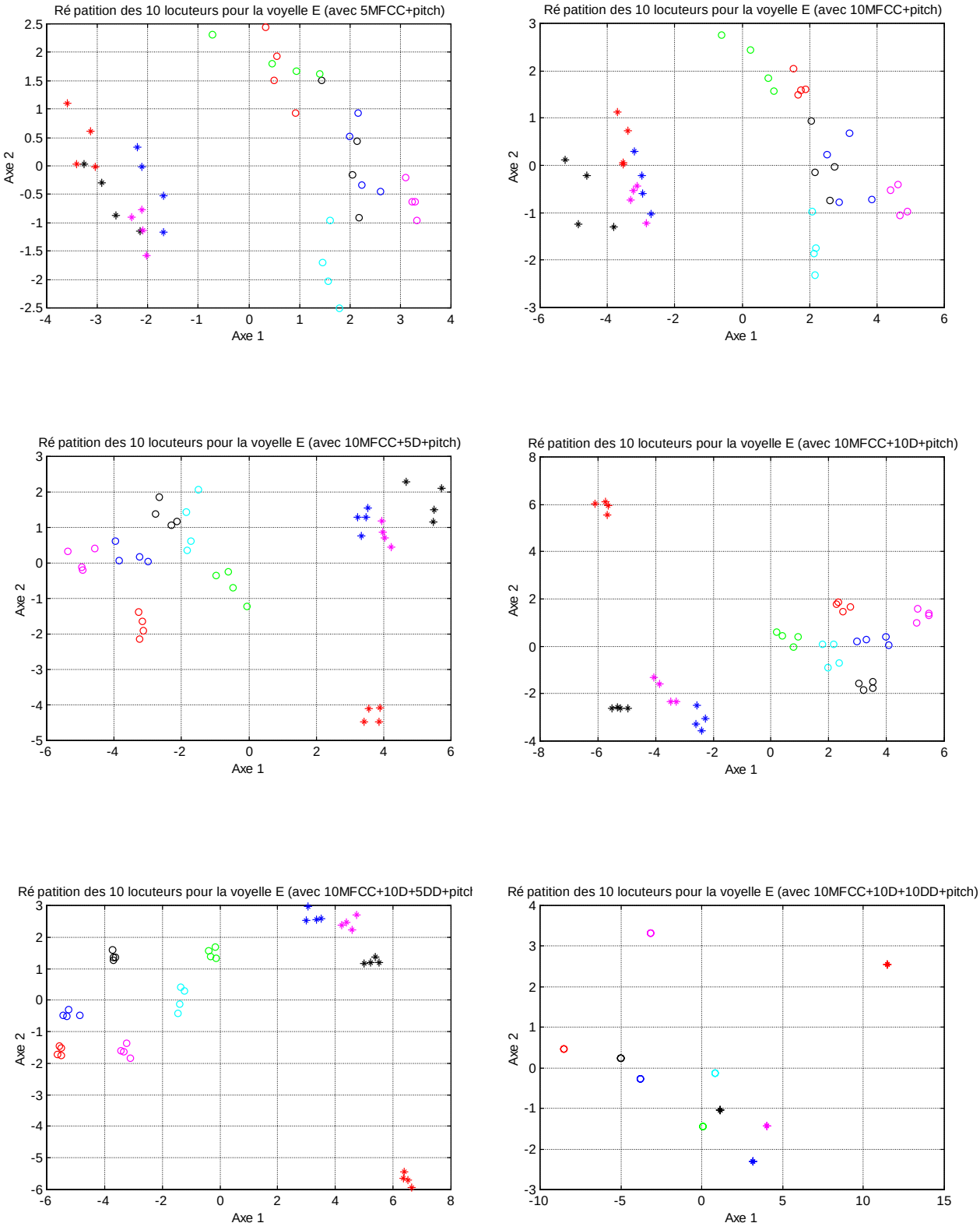


Figure 6.10 Répartition des locuteurs dans l'espace ADL pour la voyelle /e/.

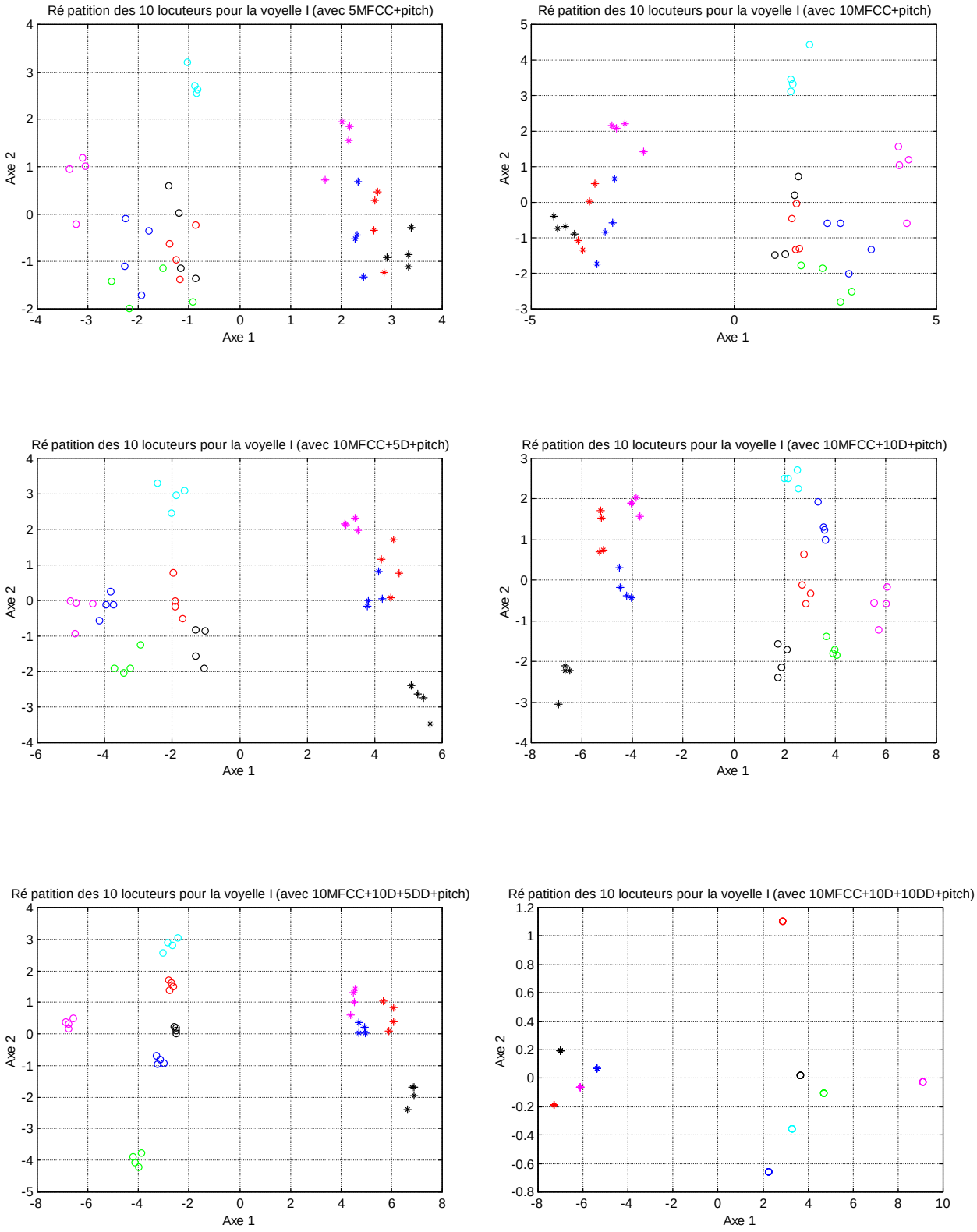


Figure 6.11 Répartition des locuteurs dans l'espace ADL pour la voyelle /i/.

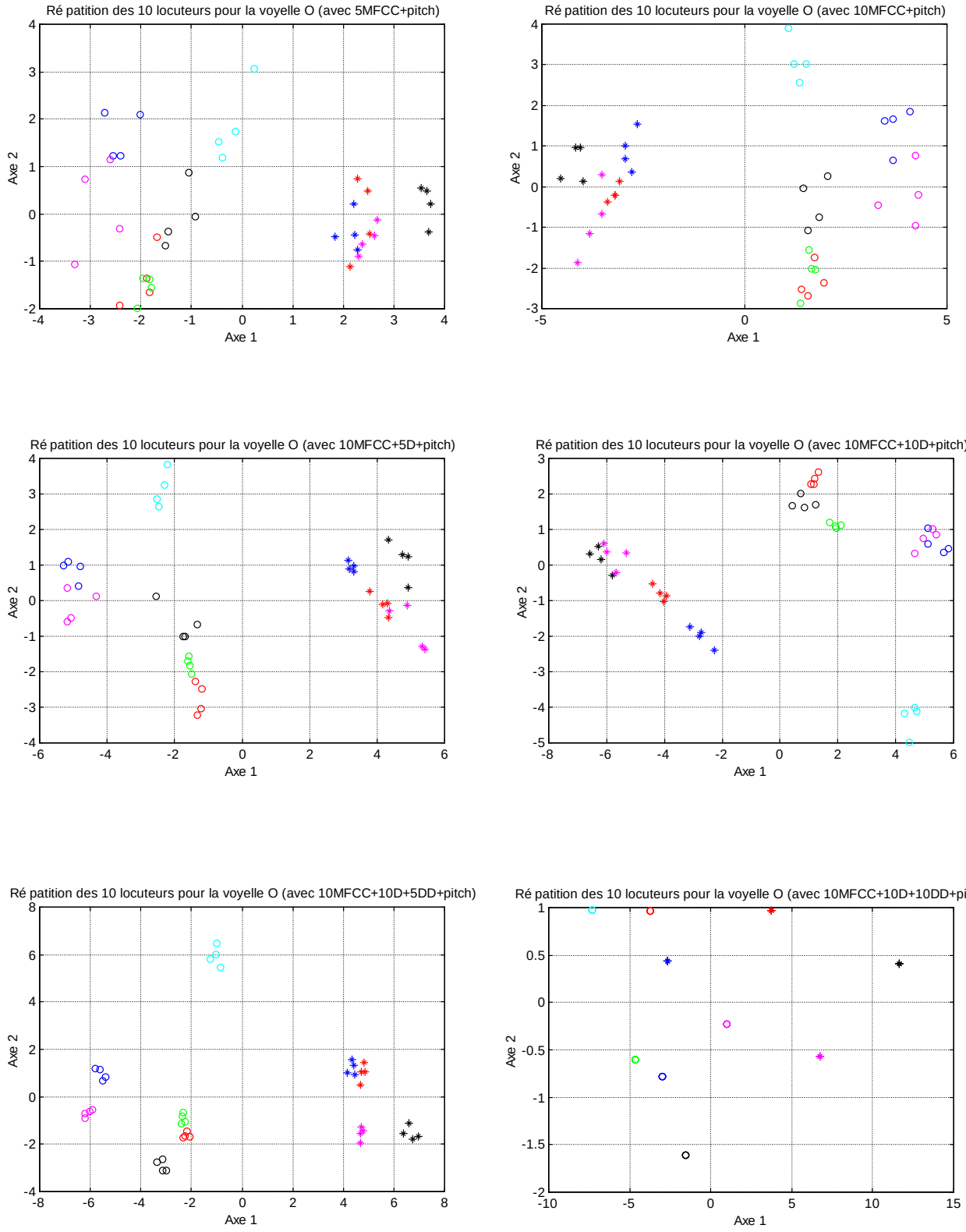


Figure 6.12 Répartition des locuteurs dans l'espace ADL pour la voyelle /o/.

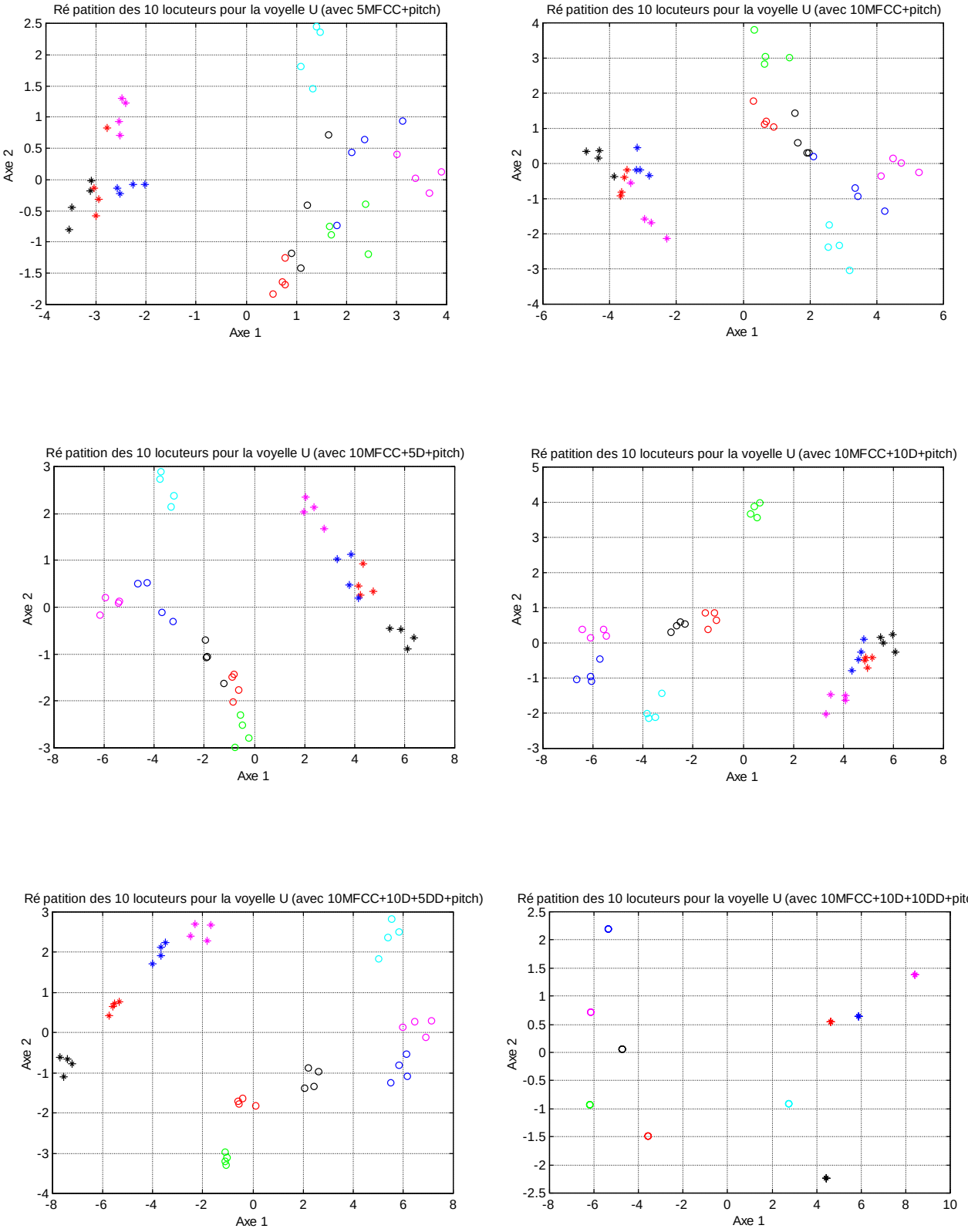


Figure 6.13 Répartition des locuteurs dans l'espace ADL pour la voyelle /u/.

6.2 Conclusion

Dans ce chapitre nous avons essayé de mettre en évidence la pertinence de quelques paramètres utilisés pour la caractérisation du locuteur à savoir la fréquence fondamentale et les coefficients MFCC et leurs dérivées qui sont actuellement les paramètres les plus répandus en reconnaissance du locuteur. Les points cités ci-dessous donnent quelques justifications pour le choix de ces paramètres ainsi que les résultats obtenus et leurs interprétations :

- Les premières séries d'expérimentations concernant nos travaux sur la durée n'ont pas été fructueuses. L'ensemble des problèmes rencontrés peuvent être classés en deux catégories : 1) problèmes d'outils que ca soit pour le logiciel de segmentation de Régine André (problème de réglage des paramètres) ou bien pour le logiciel WavEdit (problème de la non-considération des pauses), et 2) le caractére générique de la base BDSONS ainsi que sa pauvereté pour des travaux de recherches spécifique au domaine de la reconnaissance du locuteur.
- Pour la fréquence fondamentale, elle nous a permis de séparer les locuteurs en deux classes bien distinctes une classe pour les femmes (o sur les figures) et une classe pour les hommes (* sur les figures). Malheureusement, ce paramètre ne permet pas une discrimination nette entre locuteur d'où la nécessité de le coupler avec d'autres paramètres plus discriminant tel que les MFCC et leurs dérivées.
- Concernant les coefficients MFCC et leurs dérivées, nous avons travaillé avec l'approche connue sous le nom du « spectre moyen à long terme », c'est à dire on utilise des caractéristiques vectorielles mesurées uniquement par la valeur moyenne. La similarité entre deux jeux de mesure se réduit donc à un calcul de distance entre leurs valeurs moyennes. Malheureusement, le spectre moyen à long terme est très sensible aux variations du canal de transmission, ce qui le rend inutilisable dès que les locuteurs ne sont pas enregistrés dans un environnement complètement contrôlé (comme celui de la base BDSONS). En plus, l'ensemble des chercheurs s'entendent pour affirmer que ce sont les variations autour du spectre moyen qui sont les plus significatives : le spectre moyen à long terme étant trop sensible à la variabilité intersessions des caractéristiques du locuteur et aux variations du canal de transmission, il ne doit pas être pris en compte mais plutôt utilisé comme élément de normalisation. Néanmoins, vu les propriétés citées ci-dessous, les coefficients MFCC sont actuellement les plus utilisés en reconnaissance du locuteur.
- Le cepstre est une représentation de l'information spectrale pour laquelle une opération de filtrage linéaire se traduit par une modification additive (c'est le principe du traitement de signal homomorphique). Par conséquent, il est théoriquement simple de compenser dans le domaine cepstral l'effet d'un filtrage linéaire (même si le filtre est inconnu). C'est le principe de la normalisation cepstrale qui est, à ce jour, le moyen le plus efficace pour contrer les variations du canal de transmission. Cette propriété est donc en particulier indispensable pour les enregistrements téléphoniques.

- Le cepstre fournit naturellement un moyen très simple de réduire de manière pertinente la quantité d'information. On sait en effet que le lissage de la représentation spectrale obtenu en ne conservant que les premiers coefficients cepstraux conserve l'enveloppe spectrale caractéristique du signal de parole.
- La famille la plus simple de distances (distance euclidienne, éventuellement pondérée) correspond, lorsqu'elle est appliquée aux paramètres cepstraux, à une mesure de distance sur les spectres à court-terme dont on sait qu'elle est pertinente dans le cas de signaux de parole. Cette propriété vient à bout d'un défaut commun à plusieurs systèmes plus anciens pour lesquels les mesures de similarité calculées étaient difficilement interprétables (en particulier le cas pour les systèmes qui utilisaient des distances euclidiennes appliquées directement aux coefficients du modèle AR).
- Les coefficients cepstraux sont statistiquement très faiblement corrélés. Cette propriété rend d'une part inutile les procédures d'orthogonalisation souvent employées avec d'autres jeux de paramètres, d'autre part, elle simplifie la mesure de la distance entre deux jeux de paramètres cepstraux. Notons que la propriété de non-corrélation est une conséquence de la transformée de Fourier. Par conséquent, les coefficients de banc de filtres (même si la puissance dans chaque voie est exprimée sur une échelle logarithmique) ne possèdent pas cette propriété.
- Enfin, de manière plus anecdotique, les paramètres cepstraux fournissent un moyen très simple de s'affranchir de la différence de niveau qui peut exister entre deux spectres identiques : il suffit de ne pas prendre en compte le premier coefficient cepstral pour travailler sur des paramètres normalisés en niveau (par rapport au niveau moyen sur une échelle logarithmique).
- Pour les coefficients ADL, nous avons constaté que l'analyse discriminante linéaire permet de transformer l'espace d'origine vers un autre espace plus discriminant, sauf que cette analyse a été faite sur l'ensemble du vecteur d'entrée c'est-à-dire sans réduction de la dimension et sans élimination des paramètres qui dégradent les performances du système. Pour cela, il faut tenir compte des résultats obtenus par F-ratio sur chaque paramètre et qui donne une idée sur la mesure de performance de chaque paramètre individuellement. Une autre approche peut être appliquée et celle qui a été utilisée par (Paliwal, 1992). Elle consiste à calculer la performance de chaque paramètre, puis classer les paramètres selon la mesure de performance et enfin prendre un sous ensemble caractérisé par les premiers coefficients et contribuent plus à la discrimination entre les classes de locuteurs.

6.3 Bibliographie

(Akbar, 1997) Akbar, Mohammad (1997): "Waveedit, an interactive speech processing environment for microsoft windows platform", In EUROSPEECH-1997, 677-680.

(Campbell, 1997) J. P. Campbell, “Speaker Recognition: A Tutorial”, in Proceedings of the IEEE, 85(9)(1997), pp. 1437–1462.

(Depambour et Obrecht, 1997) Un système d'étiquetage automatique de la parole basé sur la durée des sons. Dans : IVème Congrès Français d'Acoustique, pp. 381-384, Marseille France, 14-18 avril 1997.

(Descout et al, 1986) Descout R., Serignat J.F., Cervantes O. and Carre R. (1986), ”BDSONS : Une base de données des sons du français”, 12th International Congress on Acoustic, Toronto Canada.

(Harrag et al., 2005) Harrag A., Mohamadi A., Serignat J.F. (2005), “LDA Combination of Pitch and MFCC Features in Speaker Recognition”, INDICON, Chennai, India, 11–13, 237–240.

(Paliwal, 1992) K. K. Paliwal. Dimensionality reduction of the enhanced feature set for the HMM-based speech recognizer. Digital Signal Processing, pp. 157-173, 1992.

Chapitre

7 Expérimentation II : Base QSDAS Spécification, mise en œuvre et expérimentations

Sommaire

Résumé

La première partie de ce chapitre est consacré à la nouvelle base de données développée appelée QSDAS (Quranic Speech Database for Arabic Speaker recognition). Cette base qui vient répondre aux différents manques constatés lors des premières expérimentations sur la base française BDSONS. La deuxième partie de ce chapitre est centrée sur les tests effectués sur cette nouvelle base.

7.1 Introduction

Les deux dernières décennies ont connue un intérêt croissant pour les technologies de la reconnaissance du locuteur. Dans le but d'avoir un volume de données parole adéquat pour entrainer et tester les systèmes de reconnaissance de locuteur, la disponibilité de base de données est devenue cruciale pour le développement de ces systèmes que ca soit dans un but de recherche ou d'applications donnant naissance à une diversité de structure et de contenue des bases. Pour ces raisons, plusieurs bases de données dans différents langages ont été collectées sur plusieurs années et dans divers pays : corpus d'Anglais américain TIMIT (Garofolo, et al., 1993) ; plus de 40 langues européennes avec les projets POLYCOST et SpeechDat (Melin, 2000 ; Nataf, 1996) ; La base de données YOHO (Campbell, 1995) ; la base de données CSLU (Cole et al., 1998) ; le corpus chinois (Lee et al., 2002) ; le corpus suédois (Melin, 1996) ; le corpus italien (Falcone, 1996) ; le corpus espaniol (Ortega-Garcia et al., 1998) ; le corpus japonnais (Makino, 2007), …etc, des ressources similaires pour la langue arabe sont très peu ou inexistantes.

Pour combler ce manque et pour répondre à nos besoins de recherche dans le domaine de la parole plus particulièrement le domaine de la reconnaissance du locuteur, nous avons développé une nouvelle base nommée QSDAS (Quranic Speech Database for Arabic Speaker) (Harrag et Mohamadi, 2010) a été développée. Cette base répond non seulement aux différentes limites rencontrées lors de l'utilisation de la base française BDSONS et se positionne comme une ressource qui aide aux développements dans le domaine de la reconnaissance du locuteur ainsi que les études dans le domaine de la parole. La base consiste en 15 .4 Go représentant 6489 fichiers issus de 1617 fichiers audio (Surats) récité par 77 locuteurs. Les 1617 fichiers audio (6.4 Go) sont partionnés en trois sets : SetD, SetS et SetT avec 77, 770 et 770 fichiers audio respectivement. Le setD contient un fichier par locuteur et est destiné au développement tandis que les SetS et SetT contiennent 10 fichiers par locuteur et sont utilisés pour le test. En plus des 1617 fichiers audio, la base QSDAS inclus pour chaque fichier audio un fichier pour la fréquence fondamental F_0 ou pitch, un fichier pour les trois premiers formants F_1-F_2-F_3 et un fichier contenant les paramètres MFCC et l'énergie (paramètres statique et dynamiques 1ère et 2ème dérivées).

7.2 Taxonomie de la base QSDAS

En plus des problèmes soulevés dans la partie expérimentale sur la base BDSONS et qui sont dus principalement à la nature générique de cette base, nous nous sommes confrontés à d'autres questions sur les caractéristiques de la base à développer, les spécificités, le contenu et sur l'existence ou pas d'une base qui répond à nos besoins. Malheureusement, pour ce dernier point, en dehors de quelques études récentes (Awadalla et al., 2005 ; Chouireb et al., 2007 ; Alghamdi et al., 2008), peu d'effort ont été accomplis dans ce domaine et reste généralement insuffisant par rapport à d'autres langues, et plus particulièrement dans le domaine de la reconnaissance du locuteur.

Notre but était de développer une base de données parole :

- riche en messages vocaux pour mesurer les variabilités inter et intra locuteurs ;
- une base dédiée à la reconnaissance du locuteur mais avec beaucoup de données qui lui permette d'être utiliser dans d'autres domaines tels que la reconnaissance et la synthèse de la parole.
- Une base dont des données comparatives peuvent exister.

Dans cette optique, la base QSDAS a été définie et collectée avec comme contenu des sourates coraniques prononcées par 77 locuteurs ce qui a permis d'avoir des échantillons audio approprié pour l'apprentissage des caractéristiques spécifiques au locuteur.

7.2.1 Spécificité RAL de la base QSDAS

Reconnaitre un locuteur est le processus de reconnaissance automatique de qui parle en utilisant les informations spécifiques à ce dernier dans le flux parole. Cette reconnaissance peut être classée en identification ou en vérification (Campbell, 1997). Ces deux taches comprennent une phase d'apprentissage pour générer les modèles de référence et une phase de test afin de déterminer l'identité proclamée. L'état de l'art des systèmes de reconnaissance de la parole et du locuteur sont largement basés sur les modèles statistiques qui nécessitent une quantité importante de données parole pour la phase d'apprentissage.

Pour la reconnaissance du locuteur, les variations intra-locuteur log-terme est un des facteurs clés qui affectent la performe du système (Furui, 1997). Beaucoup de données du même locuteur collectées à différents instants sont nécessaires pour entrainer le modèle du locuteur, et par conséquent implique un important et indésirable effort pour l'apprentissage et restreint les applications pratiques, en particulier lorsque la commodité du client est une exigence fondamentale. Des techniques d'adaptation du modèle permettent d'obtenir de bons modèles du locuteur avec une quantité de données relativement faible (Reynolds et al., 2000). Néanmoins, l'importance des bases de données parole correctement collectées est toujours appréciée par les chercheurs et les développeurs de systèmes.

Avec la croissance rapide de l'économie et les possibilités du marché dans le monde arabe, les technologies de la communication parlée pour la langue arabe sont devenues un domaine actif, ouvert et très attractif ces dernières années ce qui s'explique par les efforts accrus consacrés à l'amélioration de l'infrastructure de recherche dans ce domaine. Beaucoup de corpus pour des systèmes de reconnaissance de locuteur pour la langue anglaise ou pour d'autres langues ont été développés et mis à disposition. Par exemple, les corpus développés par Le Linguistic Data Consortium (LDC, http://www.ldc.upenn.edu/), l'European Language Research Association (ELRA, http://www.icp.grenet.fr/ELRA/) et l'Oregan Graduate Intsitute (OGI, http://cslu.cse.ogi.edu/).

Le développement de QSDAS est destiné à fournir des données publiques axées sur les taches de la reconnaissance du locuteur, pour renforcer l'infrastructure de la technologie de reconnaissance du locuteur arabe et faire progresser la recherche et le développement dans ce domaine.

7.2.2 Spécificité arabe de la base QSDAS

La langue arabe n'est pas une seule variété linguistique ; c'est plutôt un ensemble de différents dialectes différents (par exemple, l'arabe du Golfe, L'arabe du moyen orient, et l'arabe de l'Afrique du nord). En général, il existe deux classes de langue arabe : la forme classique utilisé dans le Coran et l'arabe moderne standard utilisés comme langue académique. Dans notre base, nous avons choisi d'utiliser l'arabe classique utilisé dans le Coran, ce choix est motivé par :

- Le Coran est la base et la référence pour la langue arabe ;
- Le Coran est l'élément commun entre les différents pays arabes et musulmans (1,5 Milliard de musulmans et plus de 350 Millions de citoyens arabes) ;
- Le Coran est le seul texte arabe vocalisé;
- Des enregistrements qui permettent des études comparatives sont disponibles, en particulier pour l'identification de l'accent, la reconnaissance des documents vocaux, et la diversité de la parole qui est nécessaire pour d'autres types de recherches ;
- Et comme dernier point, nous voulions apporter notre petite contribution à la science et au développement social de notre communauté en ces temps difficiles que connait le monde musulman, et surtout le monde arabe.

Morphologiquement, la langue arabe comporte 28 lettres, 25 consonnes et trois voyelles longues. Phonologiquement, la langue arabe a 34 phonèmes : 25 consonnes, 3 semi-voyelles, 3 voyelles longues et 3 voyelles courtes non représentées par des lettres mais marquées par des signes diacritiques (de signes de vocalisation placé devant ou derrière une consonne (El-Mosiry, 2000 ; Abd El Hameed, 2003). Syntaxiquement, morphologiquement et sémantiquement, l'arabe diffère des langues Indo-Européennes. C'est une langue sémitique dont la principale caractéristique est que les mots sont construits à partir des racines en suivant certains modèles fixes et en ajoutant des infixes, des préfixes ou des suffixes (Ben Sassi et al., 1999).

L'écriture de l'arabe quant à elle est un élément très important qui sera utile pour la segmentation et l'extraction des phonèmes arabes. La caractéristique de cette écriture est qu'elle ne contient que des lettres pour les consonnes et les voyelles longues. Les voyelles courtes et le doublement de consonnes peuvent être indiqués par des signes diacritiques (Alotaibi et Shahshavari, 1998). Acoustiquement, la différence entre les voyelles courtes et les voyelles longues se situe au niveau du temps, les phonèmes des deux types de voyelles ont les mêmes propriétés.

7.2.3 Spécificité Coran de la base QSDAS

Les dialectes arabes modernes ont dévié de l'arabe classique pur du Coran. En outre, certaines lettres ont pris différents prononciations dans la langue courante. Afin d'acquérir la bonne prononciation des sons de la langue arabe, il est impératif de les écouter à plusieurs reprises puis pratiquer jusqu'à atteindre la précision voulue. Cela est vrai même pour les arabophones quand ils entreprennent l'étude de Tajweed.

Le Tajweed signifie, par définition, améliorer, perfectionner, et de devenir excellent. Fonctionnellement, cela signifie articuler chaque lettre dans son bon timing et de son vrai point d'articulation (Czerepinski, 2004 ; Umm Mohamed, 1997), ce qu'on appelle en arabe « al makhraj ». Tajweed, c'est l'une des sciences islamiques et il sert à préserver la langue de ses erreurs de prononciations c'est-à-dire préserve le texte de la distorsion.

Dans cette science, chaque lettre a son droit et son dû. Les droits d'une lettre sont les caractéristiques qui lui sont toujours connectées, les dûs sont les caractéristiques qui sont présentes, parfois mais pas toujours, comme la répétition « qualquala », la fusion, le ton nasal, …etc. Il ya quatre principe de Tajweed :

- La connaissance du point d'articulation des lettres ;
- La connaissance des caractéristiques des lettres ;
- La connaissance des changements des lettres par rapport aux règles ;
- Et, à l'exercice de la langue par la lecture souvent répétée ou la récitation sonore.

7.2.3.1 Points d'articulation « al makharij »

Dans le corps humain, les parties du corps relatives à la parole se divise en cinq sections principales (Tableau 7.1) :

Tableau 7.1 Différents points d'articulation de la langue arabe.

N° de section	Nom de la section	Nom en arabe	Description
1	Al-jawf	الجوف	Zone de la poitrine
2	Al-halaq	الحلق	La gorge
3	Al-lisan	اللسان	La langue
4	Ash-shafataan	الشفتان	Les lèvres
5	Al-khaychoum	الخيشوم (الأنف)	Le cavité nasale

La figure 7.1 résume les points d'articulation pour chaque son en arabe. Plus de détails concernant cette spécificité coranique ainsi qu'un descriptif des règles de Tajweed sont disponibles dans l'annexe C.

1. Consonnes du pharynx : hamza, ha, ha, ayn, kha, ghayn –
2. Consonnes uvulaire : kaf, qaf –
3. Consonnes palatales : jim, shin, ya –
4. Consonnes molaires : dawd,
5. Consonnes Alvéolaire : ra, lam, noun –
6. Consonnes dantales : ta, dal, Ta –
7. Consonnes alvéo-dentales : tha, dhal, dhaw –
8. Consonnes inter-dantales : zay, sin, sawd –
9. Consonnes labiales : ba, fa, meem, wow -

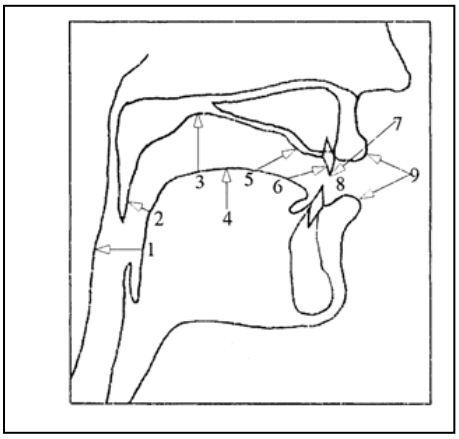


Figure 7.1 Points d'articulation pour les sons de l'arabe.

7.3 Organisation de la base QSDAS : Structure, répertoires et convention

La base de données est structurée comme suit (par niveau du plus haut au plus bas) :

QSDAS - Répertoire racine, contient 3 sous-répertoires (SetD, SetS et SetT) ;

Chacun des sous-répertoires SetD, SetS et SetT contient 4 sous répertoires F_0, FRM, MFP et WAV ;

Pour le SetD, ces sous-répertoires contiennent la fréquence fondamentale F_0, les fréquences des formants, les caractéristiques acoustiques et les fichiers audio ;

Pour les SetS et SetT, chacun d'eux contient un répertoire par locuteur ; le nom de chaque répertoire locuteur est Sdd, où S désigne locuteur (Speaker en anglais) et dd est l'identifiant du locuteur en base 10 (de 00 à 76). Chaque répertoire contient 10 fichiers de données, chaque fichier de données a un nom $sd_2d_1d_0$.ext, où :

S est l'ensemble de référence :

- s='D' pour le SetD,
- s='S' pour le SetS
- s='T' pour le SetT ;
- $d_2d_1d_0$: est la base, d_2d_1 = ID du locuteur (de 00 à 76) et d_0 = ID du fichier (de 0 à 9) ;

ext : est l'extension du fichier :

- ext='F0' pour le fichier pitch (F_0),
- ext='FRM' pour le fichier des formants (F_1-F_2-F_3),
- ext='MFP' pour le fichier des paramètres (12 MFCC + l'énergie + 1ère dérivée + 2ème dérivée),
- ext='WAV' pour le fichier audio au format wav.

La même convention est appliquée aux fichers du SetD. La figure 7.2 donne un aperçu réel de l'organisation de la base de données QSDAS.

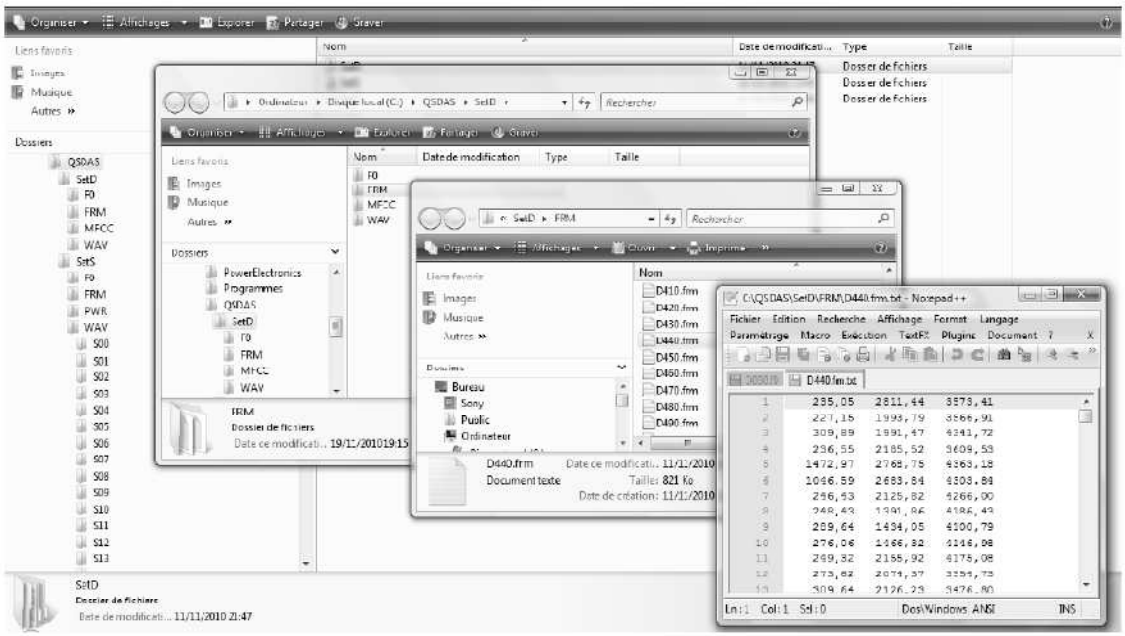


Figure 7.2 Apercu réel de l'organisation de la base de données QSDAS.

7.4 Format de fichiers, taille, répartition et statistiques

- ***Format des fichiers :*** les fichiers sons ont le format suivant :

  Format : WAV

  Fréquence d'échantillonnage : 16 kHz

  Nombre de Bits : 16 Bits

  Nombre de canaux : 1

  Rythme et volume des sons : naturel, aucune spécification particulière

- ***Taille de la base et répartition des fichiers :*** La base QSDAS contient 623 répertoires avec un total de 6489 fichiers. La taille totale de la base est de 15.4Go répartis en 3 sets : SetD, SetS et SetT contenants 308, 3080 et 3080 fichiers, respectivement. La base contient un fichier par locuteur pour le SetD et 10 fichiers par locuteur pour les SetS et SetT, sachant qu'il y a 77 locuteurs masculin. Le tableau 6.2 donne la répartition des fichiers par Set.

- ***Statiques de la base QSDAS :*** la durée totale des fichiers est de 30h répartie sur 1617 fichiers audio. Cela signifie que la durée moyenne de chaque fichier est de 66s tandis que la durée moyenne par locuteur est de 24 min. Les fréquences des différentes durées de fichiers sont données par la figure 7.3.

Tableau 7.2 Répartition des fichiers dans la base QSDAS.

	SetD			SetS			SetT			QSDAS
	Nb de réps	Nb fichs/Loc	Nb de fichs	Nb de réps	Nb fichs/Loc	Nb de fichs	Nb de réps	Nb fichs/Loc	Nb de fichs	Nb de fichs
F0	1	1	77	77	10	770	77	10	770	1617
FRM	1	1	77	77	10	770	77	10	770	1617
MFP	1	1	77	77	10	770	77	10	770	1617
WAV	1	1	77	77	10	770	77	10	770	1617
Nb de réps	4			308			308			620+3*
Nb de fichs			308			3080			3080	6468+21**

* 3 sous-répertoires (SetD, SetS et SetT) ; ** 21 fichiers textes des sourats (1+10+10).

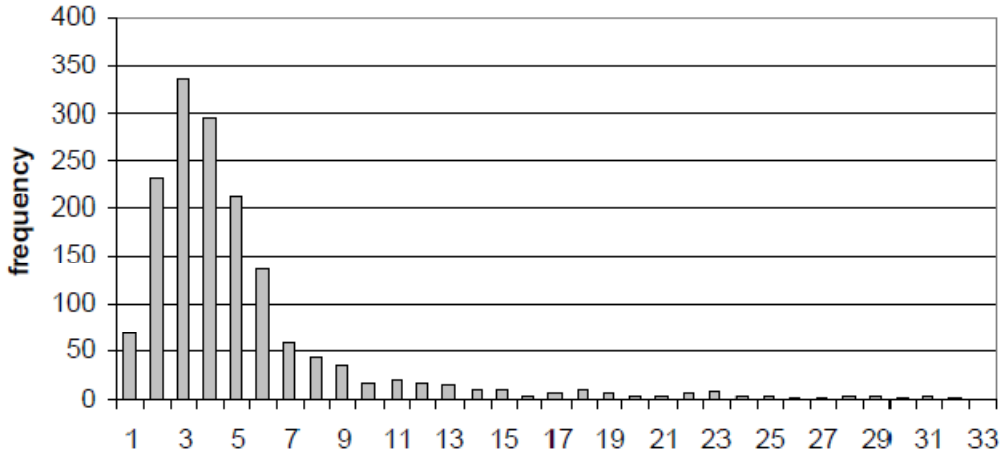


Figure 7.3 Fréquences des durées de fichiers de la base QSDAS.

Ci-dessous quelques statistiques supplémentaires :

- Durée minimale des fichiers audio : 11s
- Durée maximale des fichiers audio : 11 min et 24s
- Durée moyenne des fichiers audio : 1min et 7s
- L'écart-type des fichiers audio : 58s

Le tableau 7.3 donne plus de détails sur le contenu de la base QSDAS.

Tableau 7.3 Répartition des mots et des caractères dans la base QSDAS.

	Nb de mots	% Nb de mots	Nb de mots uniques	% Nb de mots uniques	Nb de caractères	% Nb de caractères
SetD	133	17.71	133	17.71	552	17.80
SetS	417	55 .53	416	55.39	1767	56.98
SetT	201	26.76	200	26.63	782	25.22
Total	751	100.00	749	99.73	3101	100.00

7.5 Paramètres acoustiques et prosodiques de la base QSDAS

En plus de fichiers audio, la base QSDAS intègre les caractéristiques acoustiques et prosodiques suivantes :

- Fréquence fondamentale (F_0 ou pitch),
- 12 MFCC plus les premières et deuxièmes dérivées,
- 1 paramètres pour l'énergie plus première et seconde dérivée,
- Et les fréquences des trois premiers formants F_1-F_2-F_3.

Pour chaque fichier audio (.WAV), nous avons la fréquence fondamentale fichier (.F0), les fréquences des trois premiers formants F_1-F_2-F_3 fichier (.FRM) et les paramètres MFCC ainsi que l'énergie fichier (.MFP). La figure 7.4 donne un exemple de fichier audio QSDAS avec l'ensemble des fichiers associés.

La nouvelle base de données QSDAS (avec plus de fichiers et plus de locuteurs) est actuellement disponible pour des recherches dans le domaine de la reconnaissance du locuteur (Vérification, Identification, suivi de locuteur, …etc). Néanmoins, il reste beaucoup d'effort à fournir pour la rendre adaptée à d'autres travaux dans le domaine de la reconnaissance ou de synthèse de la parole.

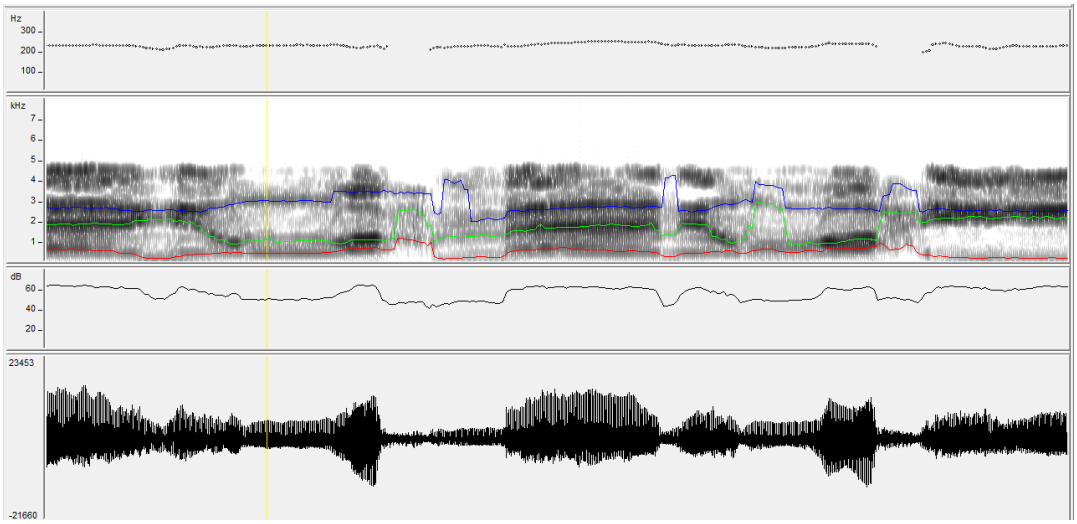


Figure 7.4 Un exemple d'un fichier de la base QSDAS (D300) : pitch, la trajectoire des formants, l'énergie et la forme de l'onde, respectivement.

7.6 Partie Expérimentale II : Base QSDAS

L'extraction des caractéristiques pertinentes est une question clé pour tout système de reconnaissance du locuteur. La suppression du signal de l'information redondante ou qui peut prêter à confusion doit être faite pour n'en retenir que l'information pertinente à la classification. Les meilleures caractéristiques sont celles qui contribuent au mieux à la discrimination entre les locuteurs. Idéalement, les caractéristiques optimales devraient avoir les propriétés suivantes:

- ayant forte variation inter-locuteur ;
- ayant faible variation intra-locuteur ;
- être faciles à mesurer ;
- robustes contre l'imitation ;
- robustes contre le bruit ;
- indépendantes entres elles.

Malheureusement, aucune des caractéristiques actuelles ne répond à ces exigences. Les caractéristiques de haut-niveau telles que la prononciation, les habitudes linguistiques, …etc, sont robustes contre le bruit mais nécessitent une reconnaissance de la parole pour obtenir une séquence de mots, et en plus, elles demandent une quantité importante de données paroles pour estimer les modèles acoustiques et les modèles de langue, ce qui ajoute une complexité intolérable à la tâche de reconnaissance du locuteur. Ces inconvénients des caractéristiques haut-niveau, justifient quelque part, le choix commun d'utiliser les caractéristiques acoustiques bas-niveau qui sont faciles à extraire, n'ont pas besoin d'une reconnaissance de la parole, et en plus, peu de données peuvent suffire à estimer de bons modèles. Leur principal inconvénient est qu'elles peuvent être facilement corrompues par le bruit de fond et par d'autres sources de distorsions.

La plupart des implémentations actuelles utilisent une forme des caractéristiques de l'enveloppe spectrale pour paramétrer la voix (LPCC, MFCC, …etc) avec de bonnes performances (Przybocki et Martin, 2004 ; Reynolds et al., 2000). Cependant des recherches récentes tentent d'inclure des informations complémentaires dans le système afin de réduire les taux d'erreurs. Des informations complémentaires telle que le pitch (Harrag et al., 2005), la phase résiduelle (Sri Rama et Yegnanarayana, 2006), la prosodie (Yegnannarayana et al, 2005), les caractéristiques dialectiques (Chakroborty et Saha, 2009), …etc.

Une approche raisonnable basée sur peu de données d'apprentissage et de test, utilisée pour améliorer les performances d'un système RAL, est d'extraire des données de plusieurs vecteurs de caractéristiques (supposés indépendants), puis extraire des caractéristiques concaténées un vecteur réduit des vecteurs fusionnés renforçant la séparabilité entre les locuteurs. Le vecteur de caractéristiques réduit permet une meilleure classification et nécessite moins de ressources de stockage et de calcul.

7.6.1 Setup expérimental

Le but dans ce qui suit est d'illustrer la valeur de la sélection des caractéristiques lorsque l'on combine des caractéristiques de différents espaces. Nous soutenons que la sélection des informations pertinentes est primordiale pour tout système de reconnaissance du locuteur et démontrant que la suppression des caractéristiques qui ne codent pas d'informations spécifiques au locuteur permet de réduire significativement le taux d'erreur.

Les travaux réalisés peuvent être divisés en trois parties :

- Etude de la meilleure représentation acoustique, y compris l'étude de la pertinence des paramètres de cette dernière ;
- Etude de la pertinence de paramètres prosodiques ;
- Etude de la fusion des paramètres des deux espaces (acoustiques et prosodiques).

7.6.2 Corpus utilisé

Pour les tests de ce chapitre, nous avons utilisé la base de données QSDAS (Harrag et Mohamadi, 2010). Des routines développées sous le logiciel Matlab ont permis d'extraire, à partir de fichiers sonores correspondant à chaque locuteur, les 5 voyelles /a/, /e/, /i/, /o/ et /u/. Ces routines utilisent les trois premiers formants fusionnés avec des coefficients MFCC utilisant la méthode ADL. Nous avons obtenus 92% de taux de reconnaissance avec les trois premiers formants, 98% en les combinant avec des coefficients MFCC et 100% en prenant les 6 premiers coefficients ADL de la combinaison des formants plus MFCC.

7.6.3 Exp 1 : Etude de la meilleure représentation acoustique

Cette première expérimentation a été consacrée à la mise en évidence de la pertinence de chaque coefficient dans un vecteur de paramètres, puis de voir la pertinence de chaque vecteur dans plusieurs jeux de paramètres. L'étude a été faite sur les jeux de paramètres suivants : (LPC, PARCOR, LPCC et les MFCC), chaque vecteur contient 10 coefficients.

Voyelle « /a/ »

Le F-ratio pour chaque coefficient est donné dans le tableau 7.4 et la figure 7.5 :

Tableau 7.4 F-ratio pour chaque coefficient des différents jeux pour la voyelle /a/.

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
LPC	1,482	4,816	1,876	2,948	2,266	3,453	2,360	4,085	3,523	2,119
PARCOR	0,945	2,951	3,185	1,241	1,371	4,386	2,541	4,956	4,307	2,119
LPCC	1,482	2,976	0,622	1,911	1,673	2,870	2,108	3,444	6,610	2,838
MFCC	1,321	3,651	6,181	4,602	3,184	2,589	4,844	5,863	6,093	2,896

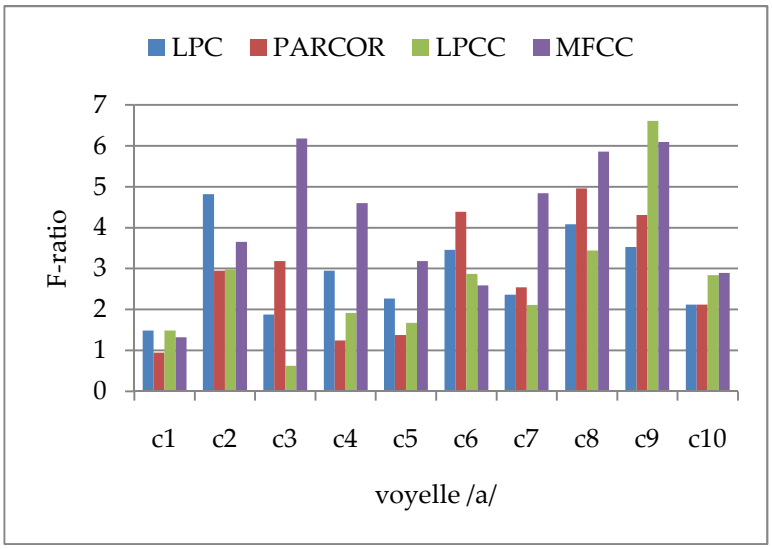


Figure 7.5 F-ratio pour chaque paramètre et pour chaque jeu pour la voyelle /a/.

Voyelle « /e/ »

Le F-ratio pour chaque coefficient est donné dans le tableau 7.5 et la figure 7.6 :

Tableau 7.5 F-ratio pour chaque coefficient des différents jeux pour la voyelle /e/.

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
LPC	1,320	3,958	2,268	1,771	0,929	1,864	0,960	1,203	1,514	0,576
PARCOR	1,086	2,881	2,961	2,363	2,081	2,254	1,201	2,985	1,818	0,576
LPCC	1,320	2,725	1,606	1,200	1,237	1,839	1,707	1,222	4,678	2,041
MFCC	0,582	3,931	2,648	4,160	3,648	1,735	4,175	2,818	6,919	2,244

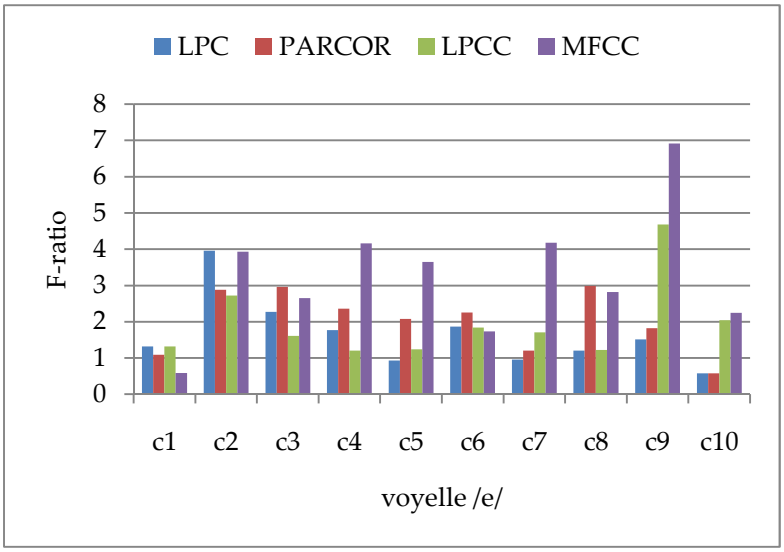


Figure 7.6 F-ratio pour chaque paramètre et pour chaque jeu pour la voyelle /e/.

Voyelle « /i/ »

Le F-ratio pour chaque coefficient est donné dans le tableau 7.6 et la figure 7.7 :

Tableau 7.6 F-ratio pour chaque coefficient des différents jeux pour la voyelle /i/.

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
LPC	1,023	1,632	1,006	1,295	1,222	1,606	1,727	2,338	1,077	0,721
PARCOR	0,821	1,480	1,384	1,781	2,363	1,536	2,468	2,977	0,948	0,721
LPCC	1,023	1,695	1,537	1,358	2,173	1,946	2,530	1,715	1,856	0,885
MFCC	1,356	6,980	4,412	3,640	6,716	4,100	5,789	6,022	5,390	3,443

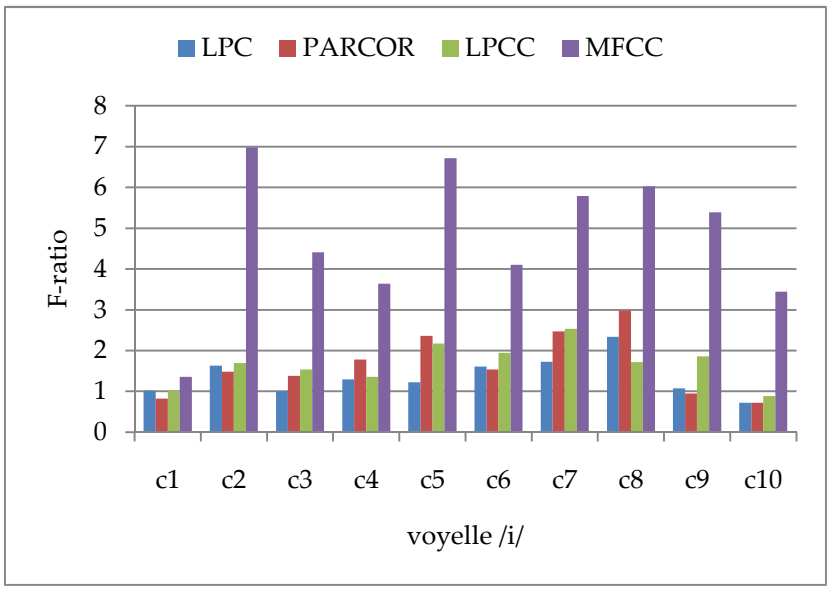


Figure 7.7 F-ratio pour chaque paramètre et pour chaque jeu pour la voyelle /i/.

Voyelle « /o/ »

Le F-ratio pour chaque coefficient est donné dans le tableau 7.7 et la figure 7.8 :

Tableau 7.7 F-ratio pour chaque coefficient des différents jeux pour la voyelle /o/.

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
LPC	1,309	3,232	2,338	1,232	0,783	1,250	1,594	1,586	1,132	0,969
PARCOR	0,557	2,953	2,940	2,022	1,052	1,360	1,218	2,696	1,153	0,969
LPCC	1,309	2,283	1,305	0,903	1,057	3,558	0,788	1,224	2,672	1,888
MFCC	0,896	3,722	2,183	2,047	2,483	1,154	1,754	3,733	8,979	1,838

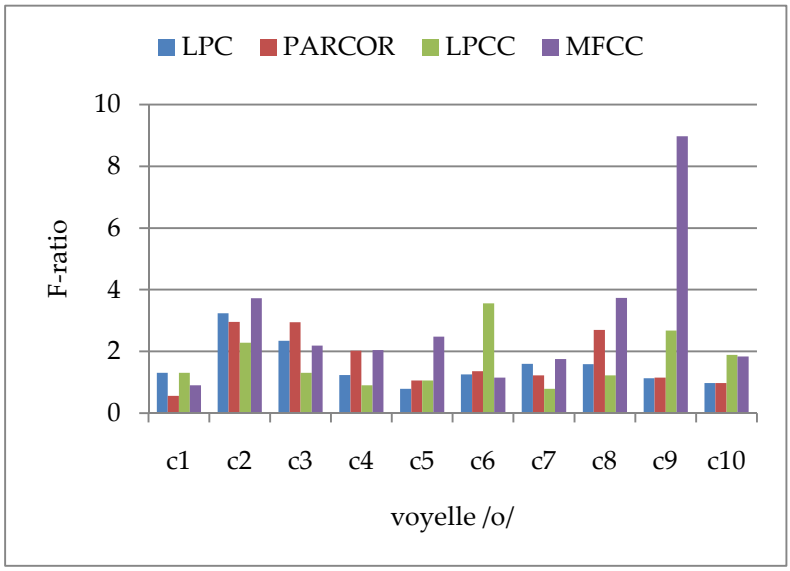


Figure 7.8 F-ratio pour chaque paramètre et pour chaque jeu pour la voyelle /o/.

Voyelle « /u/ »

Le F-ratio pour chaque coefficient est donné dans le tableau 7.8 et la figure 7.9 :

Tableau 7.8 F-ratio pour chaque coefficient des différents jeux pour la voyelle /u/.

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
LPC	0,713	4,110	1,069	2,538	1,532	2,207	1,396	3,601	3,808	1,263
PARCOR	0,882	1,649	1,687	3,078	2,748	2,449	1,701	4,383	3,809	1,263
LPCC	0,713	3,397	2,147	2,032	2,593	3,093	1,430	1,299	2,764	1,189
MFCC	1,554	5,524	3,490	3,081	4,049	3,836	8,836	6,224	4,745	12,138

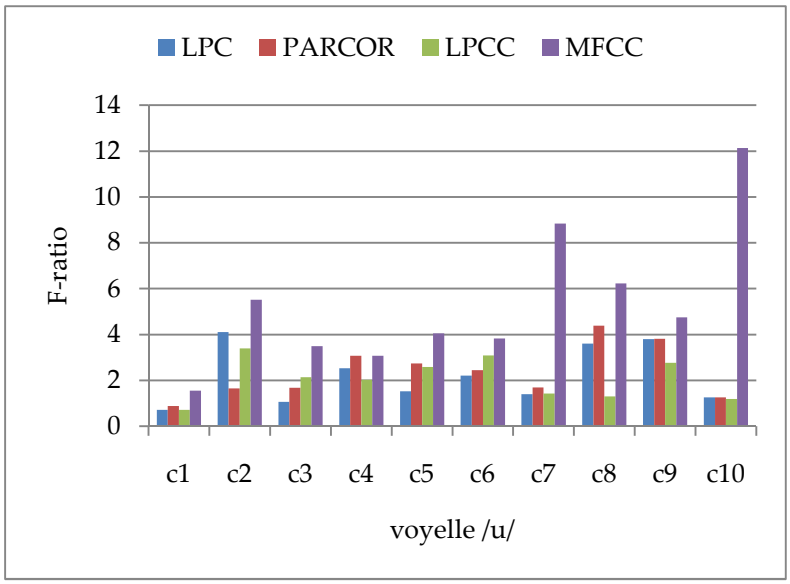


Figure 7.9 F-ratio pour chaque paramètre et pour chaque jeu pour la voyelle /u/.

Le tableau 7.9 et la figure 7.10 donnent le F-ratio par jeu de paramètres pour les 5 voyelles /a/, /e/, /i/, /o/ et /u/.

Tableau 7.9 F-ratio pour chaque jeu de paramètres.

	F-Ratio				
	/a/	/e/	/i/	/o/	/u/
LPC	0,989	0,971	0,969	0,962	0,994
PARCOR	0,984	0,972	0,977	0,954	0,994
LPCC	0,969	0,963	0,962	0,990	0,961
MFCC	0,996	0,991	0,993	0,962	0,996

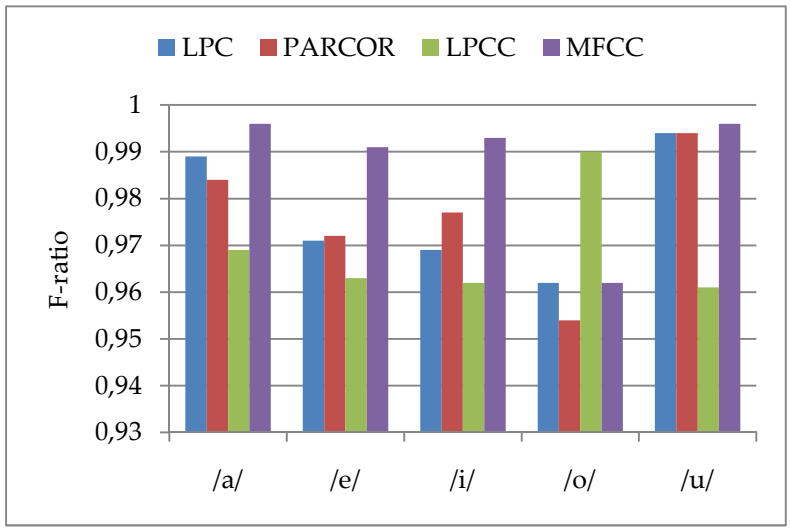


Figure 7.10 illustre les F-ratio pour chaque jeu de paramètres et pour les 5 voyelles.

Les résultats montrent clairement que la représentation MFCC est celle qui donne le meilleur F-ratio. Notons que les coefficients d'ordre élevé codent mieux l'information pertinente spécifique au locuteur que les coefficients d'ordre faible.

Ces résultats sont conforté par une deuxième série de tests basée sur le calcul du taux de reconnaissance (ou plutôt le taux d'erreur de reconnaissance) de chaque jeux de paramètres en utilisant la distance euclidienne, la distance euclidienne pondérée et la corrélation. Les figures 7.11.a, 1.11.b et 7.11.c illustrent les taux d'erreur pour les quatre représentations acoustiques LPC, PARCOR, LPCC et MFCC.

Les meilleurs scores sont détenus conjointement par le coefficient cepstraux LPCC et MFCC, ce qui confirme leur utilisation dans la littérature de la reconnaissance du locuteur. Notons que la distance euclidienne pondérée surpasse la simple distance euclidienne et la corrélation.

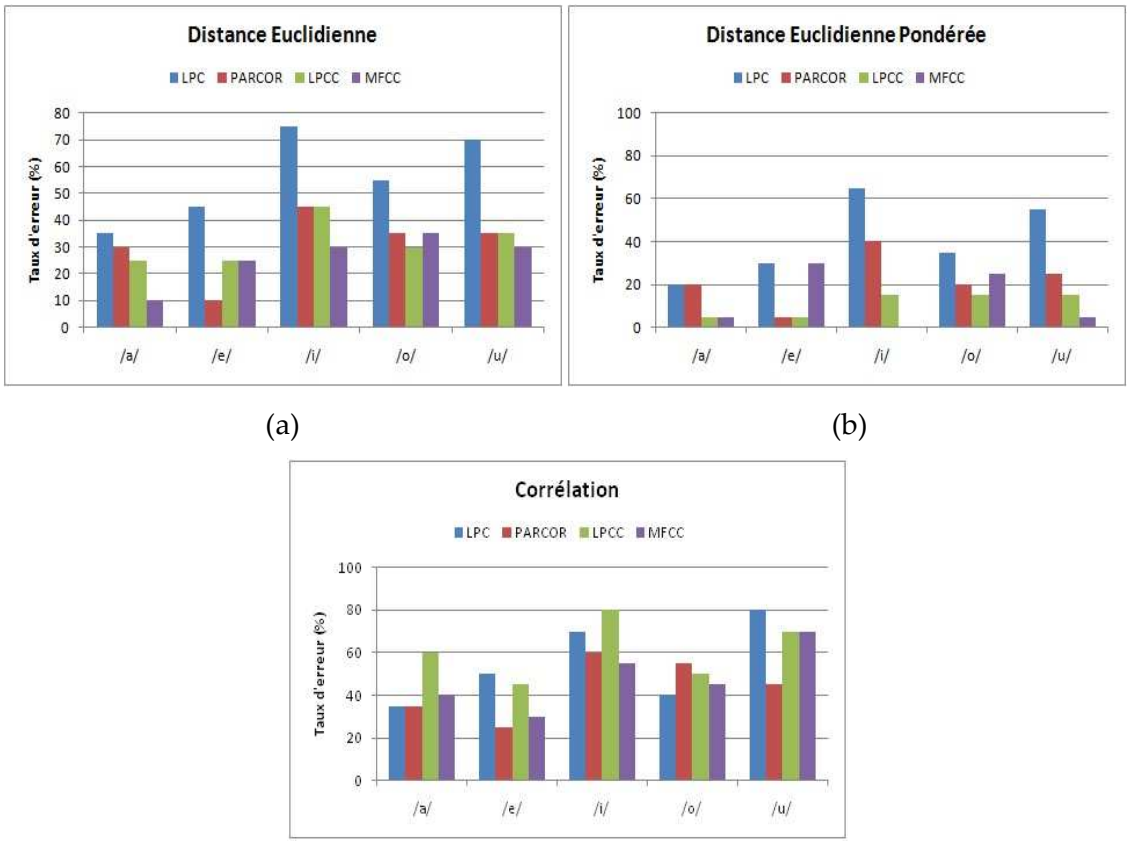
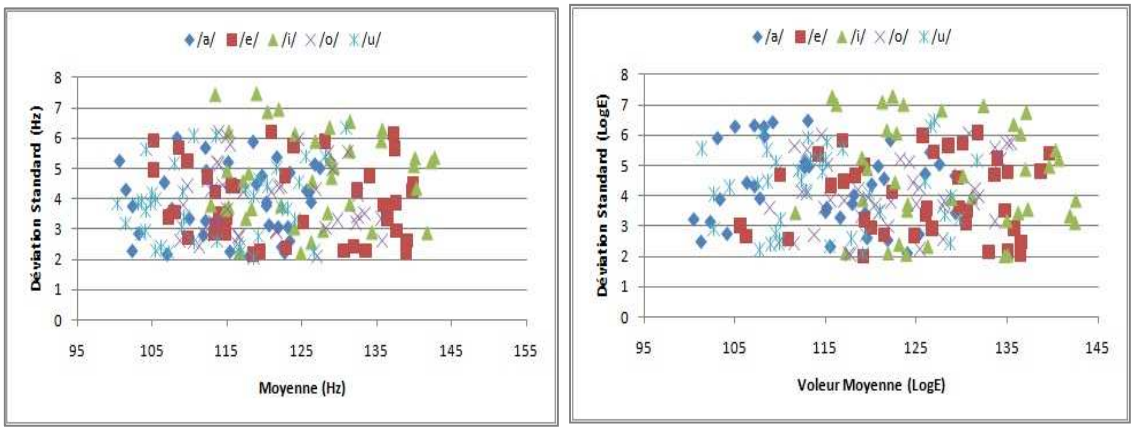


Figure 7.11 Taux d'erreur de reconnaissance pour les quatre jeux de paramètres.

7.6.4 Exp 2 : Etude de la pertinence des paramètres prosodiques

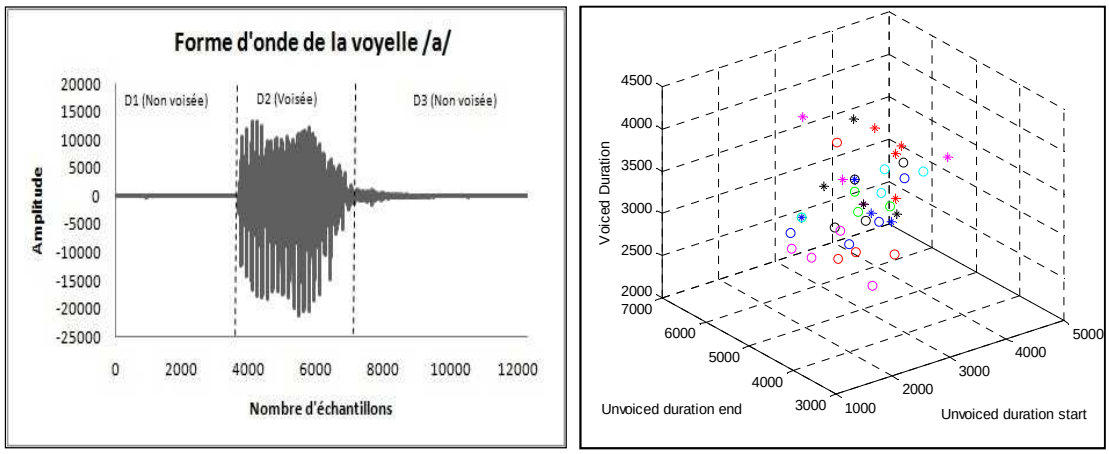
Les tests de cette expérimentation ont été consacrés à l'évaluation du pouvoir discriminant de l'information prosodique (pitch, énergie et durée). Pour le pitch et l'énergie, nous avons pris les valeurs moyennes et les déviations standards (Figure 7.12.)



(a) (b)

Figure 7.12 Répartition des locuteurs : a) espace pitch et b) espace énergie.

Pour la durée, nous avons pris les trois durées non-voisée/voisée/non-voisée (Figure 7.13.a).



(a) (b)

Figure 7.13 a) durées utilisées et b) répartition des locuteurs dans l'espace durées.

D'après les résultats, nous constatons que les paramètres prosodiques, à eux seuls, ne permettent pas une discrimination des locuteurs. Néanmoins, ces informations complémentaires fusionnées à d'autres paramètres tels que les paramètres acoustiques peuvent améliorer les performances du système de reconnaissance du locuteur.

7.6.5 Exp 3 : Fusion des paramètres acoustiques et prosodiques

Cette expérimentation vise à mesurer la pertinence des paramètres issus de la fusion de caractéristiques des deux espaces acoustique (LPC, PARCOR, LPCC et MFCC) et prosodique (Pitch, Energie et Durée). Le tableau 7.10 donne les différentes configurations testées.

Tableau 7.10 Configurations de paramètres testées.

N°	Config. 1	Config. 2	Config. 3	Config. 4	Dim
1	LPC	PAR	LPCC	MFCC	5
2	LPC	PAR	LPCC	MFCC	10
3	LPC	PAR	LPCC	MFCC	15
4	LPC	PAR	LPCC	MFCC	20
5	LPC+E	PAR+E	LPCC+E	MFCC+E	22
6	LPC+E+F_0	PAR+E+F_0	LPCC+E+F_0	MFCC+E+F_0	24
7	LPC+E+F_0+D	PAR+E+F_0+D	LPCC+E+F_0+D	MFCC+E+F_0+D	27

Cette fusion a été réalisée via trois méthodes parmi celles décrites dans le chapitre 5, à savoir : l'analyse en composante principale (ACP), la recherche séquentielle directe (SFS) et l'analyse discriminante linéaire (ADL). Pour chaque méthode et pour chaque jeu de paramètres, ainsi que pour les paramètres concaténés sans aucun traitement spécifique, nous avons calculé le taux d'erreur de reconnaissance en utilisant la technique non paramétrique supervisée connue sous le nom du k-plus proches voisins (KNN)(Darathy, 1991). Les figures 7.14 à 7.18 illustrent les taux d'erreur de reconnaissance pour les paramètres concaténés, les paramètres issus de la fusion par ACP, les paramètres issus de la sélection par SFS et les paramètres issus de la fusion par ADL, et cela pour les quatre jeux de paramètres combinés avec les paramètres prosodiques et pour les cinq voyelles /a/, /e/, /i/, /o/ et /u/, respectivement.

Les résultats montrent un net avantage aux paramètres issus de la fusion par la technique ALD par rapport aux paramètres concaténés directement ou bien fusionnés par les méthodes ACP ou SFS respectivement.

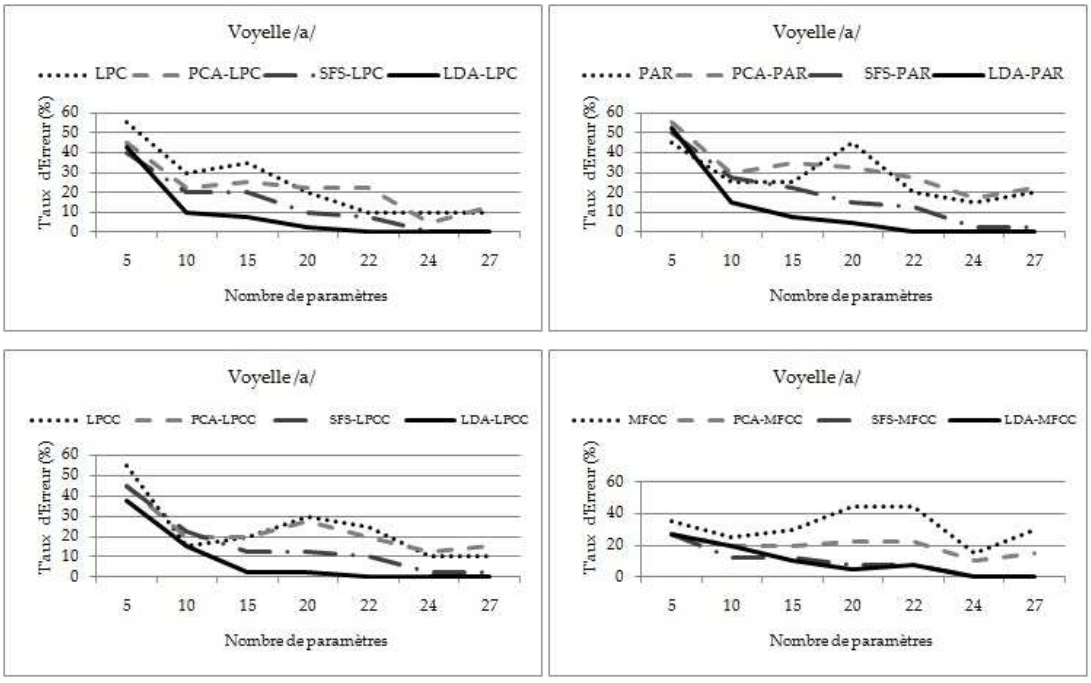


Figure 7.14 Comparaison des différents paramètres (voyelle /a/).

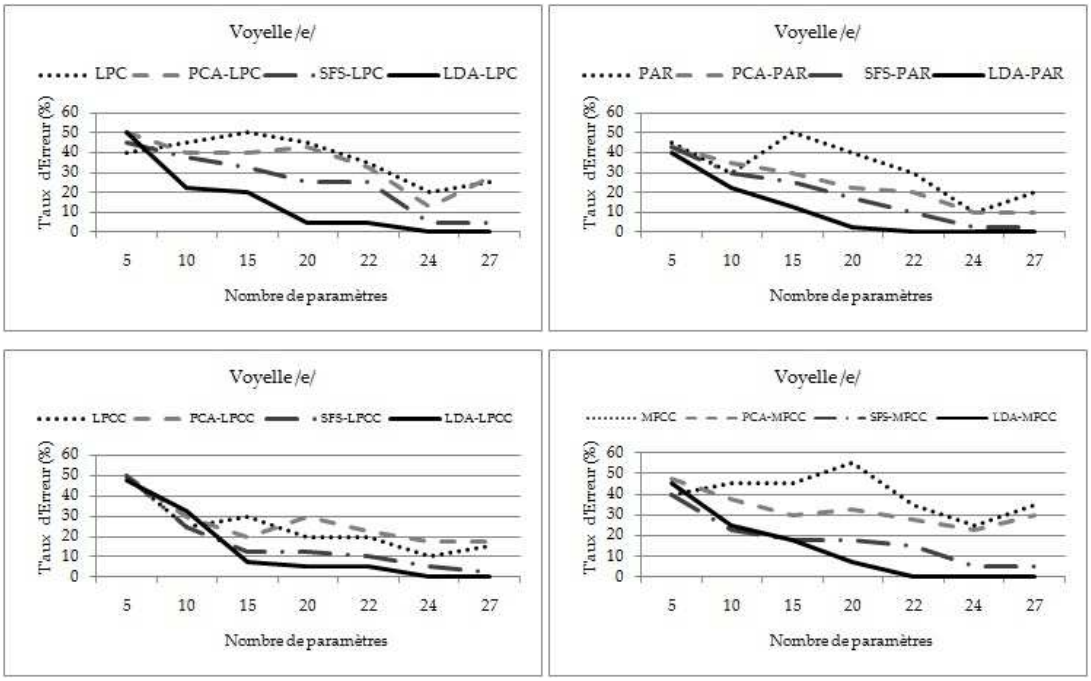


Figure 7.15 Comparaison des différents paramètres (voyelle /e/).

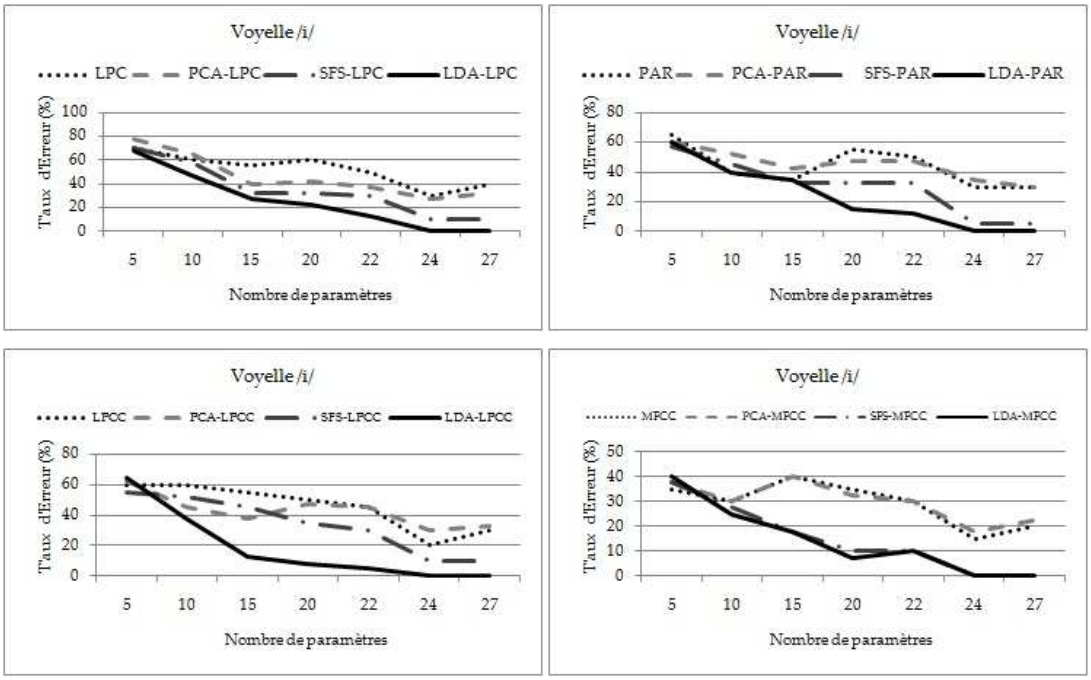


Figure 7.16 Comparaison des différents paramètres (voyelle /i/).

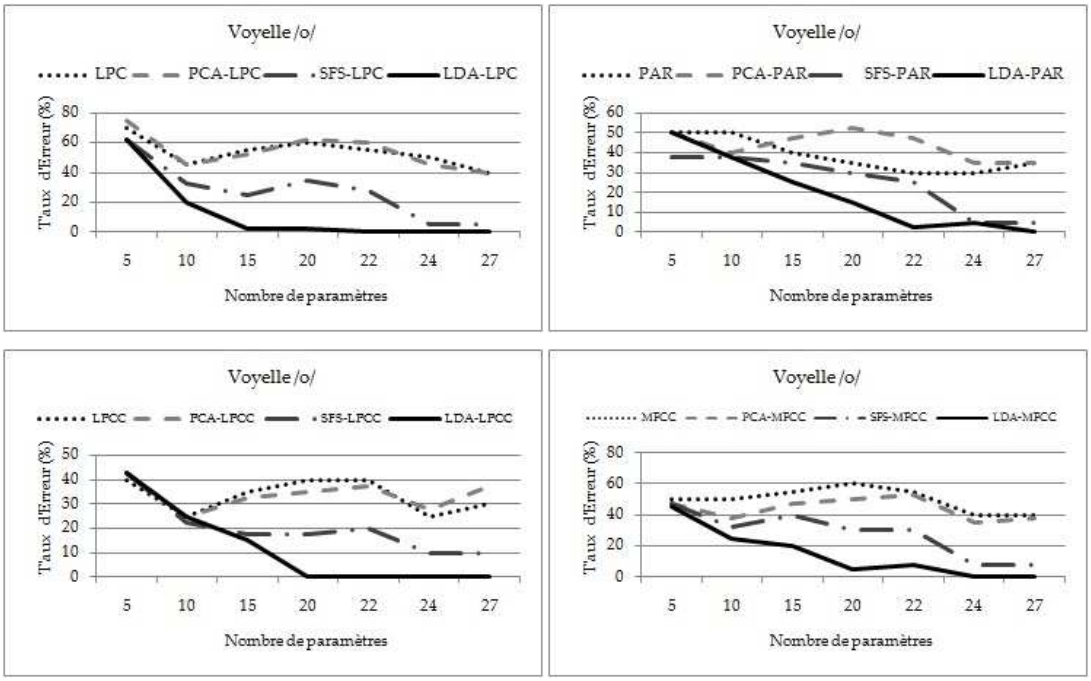


Figure 7.17 Comparaison des différents paramètres (voyelle /o/).

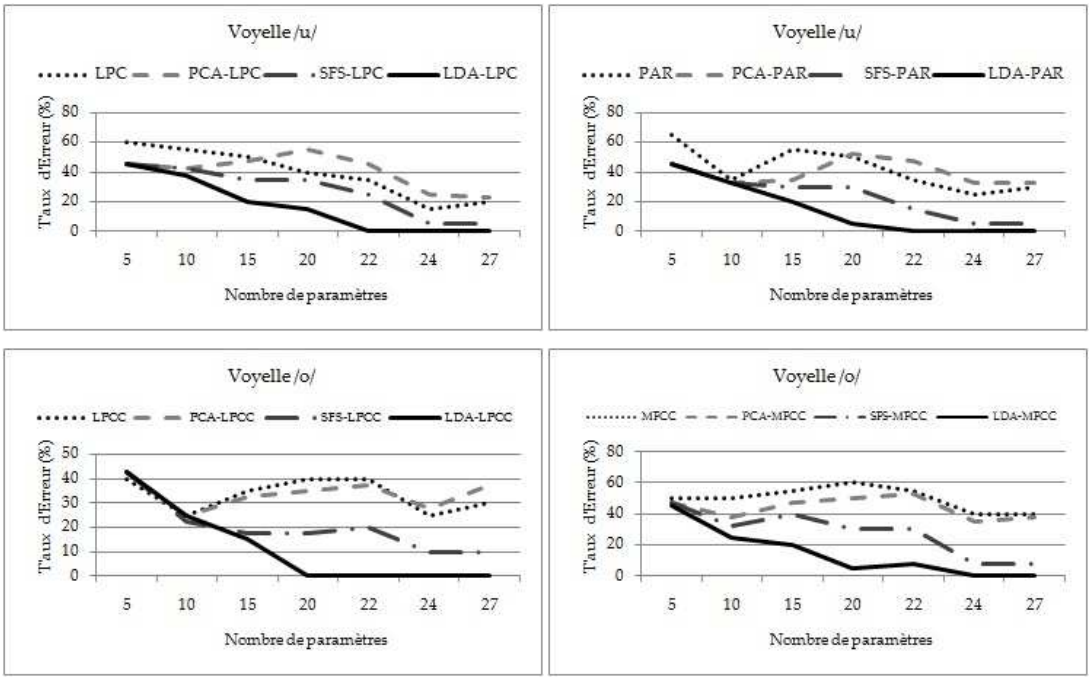


Figure 7.18 Comparaison des différents paramètres (voyelle /u/).

7.7 Conclusion

Ce chapitre illustre nos travaux : 1) pour la collecte d'une nouvelle base de données nommée QSDAS. Cette base qui vient répondre aux différents manques de la base BDSONS signalés dans le chapitre précédent, et 2) sur les tests réalisés sur cette dernière.

Nous avons commencé par évaluer les paramètres issus de différents espaces séparément en analysant dans un premier temps les paramètres acoustiques LPC, PARCOR, LPCC et MFCC utilisant une analyse F-ratio au niveau des paramètres de chaque jeu ainsi qu'une autre au niveau de chaque jeu de paramètres. Ces premiers tests montrent globalement les paramètres MFCC comme étant ceux qui ont les meilleurs F-ratio. Une deuxième série de tests sur les paramètres acoustiques a été faite en calculant les taux d'erreur de classification, en utilisation la corrélation, la distance euclidienne et la distance euclidienne pondérée. Les meilleurs scores de cette série de tests sont détenus par le coefficient cepstraux LPCC et les coefficients MFCC, ce qui confirme leur utilisation dans la littérature de la reconnaissance du locuteur. Notons que la distance euclidienne pondérée surpasse la simple distance euclidienne et la corrélation.

La deuxième batterie de tests couvre une étude comparative entre trois méthodes de fusion ou d'extraction des paramètres codant l'information pertinence du locuteur, à savoir : l'analyse en en composante principale (ACP), la recherche séquentielle directe (SFS) et l'analyse discriminante linéaire (ADL).

Les résultats montrent que pour les paramètres concaténés, sans aucune fusion utilisant une de trois méthodes ACP, SFS ou ADL, l'inclusion de nouveaux paramètres n'améliore pas forcément les performances en réduisant le taux d'erreur. Pire encore, l'inclusion de paramètres supplémentaires peut même dégrader les performances du système. Les mêmes vecteurs de paramètres fusionnés par ACP ne donnent pas nécessairement de meilleurs résultats. Ceci est du principalement aux inconvénients de la méthode ACP qui en découlent du fait que l'erreur quadratique moyenne entres les vecteurs de données dans la projection des K-composantes de l'espace ACP n'est pas nécessairement pareil que minimiser l'erreur de classification. Par contre, les mêmes paramètres fusionnés par SFS ou ADL donnent de meilleur résultat, avec une nette avance pour la technique ADL pour toutes les configurations. Cet avantage de la méthode ADL par rapport à la méthode SFS vient du fait que la méthode SFS ne permet pas de corriger les précédents ajouts de paramètres sélectionnés successivement. En plus, les paramètres ADL sont moins corrélés, ce qui n'est pas le cas des paramètres SFS.

7.8 Bibliographie

(Abd El Hameed, 2003) S. Abd El Hameed, "Tayseer Al Rahman Fe Tagweed Wa Qeraat Al Qoran", Dat Naher El Nile, 2003 (in Arabic).

(Alghamdi et al., 2008) M. Alghamdi, et al., "Saudi Accented Arabic Voice Bank", In Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics Athens, Greece, August, 2008.

(Alotaibi et Shahshavari, 1998) Y. A. Alotaibi and M. M. Shahshavari, "Speech Recognition -What It Takes for a Computer to Understand Your Commands", IEEE Potentials, Feb/March 1998.

(Awadalla et al., 2005) M. Awadalla, F. E. Z. Abou Chadi, and H. H. Soliman, "Development of an Arabic Speech Database", ITI 3rd International Conference on Information and Communications Technology, Cairo, Egypt, 2005, pp. 89–100.

(Ben Sassi et al., 1999) S. Ben Sassi, et al., "A Text-to-Speech System for Arabic Using Neural Networks", in Proceedings of IJCNN' 1999, Washington, USA, pp. 3030–3033.

(Campbell, 1997) J. P. Campbell, "Speaker Recognition: A Tutorial", in Proceedings of the IEEE, 85(9)(1997), pp. 1437–1462.

(Campbell, 1995) J. Campbell, "Testing With the YOHO CD-ROM Voice Verification Corpus ", ICASSP. Detroit, USA, May 1995, pp. 341–344. http://www.biometrics.org/REPORTS/ICASSP95.html

(Chakroborty et Saha, 2009) Chakroborty S. and Saha G. (2009), "Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter", International Journal of Signal Processing, 5.

(Chouireb et al., 2007) F. Chouireb, M. Guerti, M. Naïl, and Y. Dimeh, "Development of a Prosodic Database for Standard Arabic", The Arabian Journal for Science and Engineering, 32(2B)(2007).

(Cole et al., 1998) R. Cole, M. Noel, and V. Noel, "The CSLU Speaker Recognition Corpus", Proc. ICSLP, Sydney, Australia, 1998.

(Czerepinski, 2004) K. C. Czerepinski, "Tajweed Rules of the Quran", Parts 1, 2 & 3, 2004.

(Darathy, 1991) B. V. Dasarathy, "Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques", ISBN 0-8186-8930-7, 1991.

(El Mosiry, 2000) K. El Mosiry, "Al Gamea Fe Tagweed Wa Keraat Al Qoraan", Dar El-Fath, 2000 (in Arabic).

European Lang Resources Assoc. http://www.icp.grenet.fr/ELRA/

(Falcone, 1996) M. Falcone, "The SIVA Speech Database for Speaker Verification: Description and Evaluation", ICSLP'96, Philadelphia, USA, October 1996, pp. 1902–1905.

(Furui, 1997) S. Furui, "Resent Advances in Speaker Recognition", Pattern Recognition Letters, 18(1997), pp. 859–872.

(Garofolo et al., 1993) J. S. Garofolo, et al., "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM", NIST, 1993.

(Harrag et Mohamadi, 2010) Harrag A. and Mohamadi T. 2010. QSDAS: New Quranic Speech Database for Arabic Speaker Recognition, AJSE Journal, Theme Issue on Arabic Computing, Vol. 35(2C), pp. 7-19, December 2010.

(Harrag et al., 2005) Harrag A., Mohamadi A., Serignat J.F. (2005), "LDA Combinatio of Pitch and MFCC Features in Speaker Recognition", INDICON, Chennai, India, 11–13, 237–240.

Linguistic Data Consortium. http://www.ldc.upenn.edu/

(Lee et al., 2002) T. Lee, W. K. Lo, P. C. Ching, and H. Meng, "Spoken Language Resources for Cantonese Speech Processing", Speech Communication, 36(2002), pp. 327–342.

(Makino, 2007) T. Makino, "A Corpus of Japanese Speaker's Pronunciation of American English: Preliminary Research", Phonetics Teaching & Learning Conference UCL, August, 2007.

(Melin, 2000) H. Melin, "Databases for Speaker Recognition: Activities in COST250 Working Group 2", in COST250 –Speaker Recognition in Telephony, Final Report 1999 (CD-ROM), 2000.

(Melin, 1996) H. Melin, "Gandalf—A Swedish Telephone Speaker Verification Database", ICSLP'96, Philadelphia, USA, October 1996, pp. 1954–1957.

(Nataf, 1996) A. Nataf, "Definition of Environmental and Speaker Specific Coverage for SDB", SpeechDat, Technical Report LE2-4001-SD1.2.3, October 1996. http://www.speechdat.com

Oregon Graduate Institute. http://cslu.cse.ogi.edu/

(Ortega-Garcíaet al., 1998) J. Ortega-García, et al., "AHUMADA: A Large Speech Corpus in Spanish for Speaker Identification and Verification", ICASSP'98, Seattle, USA, 1998, pp. 773–776.

(Przybocki et Martin, 2004) Przybocki, M. A. and Martin, A. F. (2004), "NIST Speaker Recognition Evaluation Chronicles", Odyssey, Toledo Espana.

(Reynolds et al., 2000) Reynolds, D. A., Quatieri T. F., and Dunn, R. B. (2000), "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, 10, 19-41.

(Umm Mohamed, 1997) Umm Mohamed, "A Brief Introduction to Tajweed", Abul Kasem Publication House, Jeddah, 1997.

(Yegnannarayana et al, 2005) Yegnanarayana B., Prasanna S.R.M., Zachariah J.M., and Gupta C.S. (2005), "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system'", IEEE Trans Speech and Audio Processing, 13, 575-582.

Conclusions & Perspectives

Cette thèse s'inscrit dans une volonté d'extraire les données pertinentes d'une base de données sonores et, plus particulièrement, les traits du locuteur. L'approche proposée s'appuie, dans un premier temps, sur la mesure de performance des traits issus de différents espaces (acoustique, prosodique, …etc.), puis dans un deuxième temps à fusionner les traits : 1) pour améliorer les performances du système de reconnaissance et 2) de réduire la dimension.

Nous ne pouvons s'attaquer aux travaux de cette thèse sans donner un tour d'horizon qui sert d'introduction au domaine de la biométrie, et plus particulièrement au domaine de la biométrie vocale (vs Reconnaissance Automatique du Locuteur). La première partie de cette thèse a été consacrée à l'introduction du lecteur petit à petit au fond de nos travaux de recherche, en suivant toujours le fil conducteur de la biométrie, à la biométrie vocale, à la production et l'analyse du signal parole qui donnent les bases fondamentales pour attaquer la deuxièmes partie de cette thèse qui est axée principalement sur le choix, l'extraction, la sélection et la fusion des paramètres pertinents.

Quels paramètres utiliser ?

L'étude de l'efficacité de différents types d'analyse pour la RAL a fait l'objet de nombreuses publications. Dans la plupart de ces études, le système d'identification ou de vérification est fixé et les résultats obtenus pour des paramètres variés (LPC, LPCC, MFCC, Pitch, …) sont comparés. A partir des résultats comparatifs publiés, on peut tirer un certain nombre de conclusions :

- Les paramètres dérivés d'une analyse temporelle, les LPCs, LARs, PARCORs, contiennent sensiblement le même type d'information. Les résultats obtenus (p.exp taux de reconnaissance) pour ces différents paramètres sont non seulement comparables en moyenne, mais les confusions interviennent sur les mêmes segments de parole.
- Les coefficients cepstraux LPCC conduisent à de meilleurs résultats que les coefficients précités (LPC, LAR, …).
- Le pitch et sa dynamique ne permettent pas de reconnaître efficacement l'identité d'un locuteur.
- Différents paramètres gagnent à être combinés car ils peuvent contenir des informations complémentaires (p.exp les LPCC et le pitch).

- La dynamique des paramètres (paramètres différentielles contiennent également une information différente des vecteurs instantanés. Ceci a été mis en évidence par l'étude de la corrélation entre ces différentes informations. Les paramètres instantanés gagnent donc à être combinés.
- Les résultats généralement reportés montrent une légère supériorité des paramètres instantanés sur les paramètres différentiels (pour le pitch, les LPCC ou les MFCC), mais il ne s'agit pas d'un comportement systématique (des études ont montré la supériorité des paramètres différentiels sur des données bruitées).
- Cependant, malgré la recherche intense pour trouver de nouveaux paramètres en caractérisation du locuteur, il reste que les coefficients MFCC demeurent le meilleur choix. En effet, ils sont supposés être très bien représentatifs de la forme du conduit vocal. Leurs distributions statistiques sont particulièrement bien modélisées par le modèle à mélanges de Gaussiennes et les composantes du vecteur des coefficients sont convenablement décorrélées.
- A ce jour, l'information de haut niveau n'a pas été fréquemment mise en œuvre dans la reconnaissance du locuteur, cela est dû principalement à la difficulté de mesure automatique et quantitative de ce genre d'information. Néanmoins, récemment, les efforts conjoints de 10 instituts (MIT, IBM, OGI, et d'autres) ont été mis en place afin d'exploiter l'efficacité de l'information haut niveau pour un système de reconnaissance de locuteur précis. Dans ce projet commun, une large panoplie d'approches utilisant des modèles de prononciation, de la dynamique prosodique, les caractéristiques du pitch et de la durée, les flux de phones, et les interactions conversationnelles ont été explorés et développées. Il a été démontré que ces nouvelles caractéristiques et classificateurs, en effet, fournissent des informations complémentaires et peuvent êtres fusionner pour réduire l'erreur de reconnaissance.

Comment fusionner ?

Dans ce qui suit nous avons utilisé les deux approches pour la sélection des paramètres, à savoir : 1) l'approche en mesurant la pertinence des paramètres par F-ratio et 2) l'approche par mesure de des performances de classification c'est-à-dire une mesure du taux de reconnaissance. Pour les méthodes de sélection des caractéristiques, nous avons utilisé les méthodes de l'Analyse en Composante Principale (ACP), la Sélection Séquentielle Directe (SFS en anglais) et l'Analyse Discriminante linéaire (ADL).

Synthèse et interprétation des expérimentations BDSONS

Lors de ces expérimentations, nous avons essayé de mettre en évidence la pertinence de quelques paramètres utilisés pour la caractérisation du locuteur à savoir la fréquence fondamentale et les coefficients MFCC et leurs dérivées qui sont actuellement les paramètres les plus répandus en reconnaissance du locuteur. Les points cités ci-dessous donnent quelques justifications pour le choix de ces paramètres ainsi que les résultats obtenus et leurs interprétations :

- Les premières séries d'expérimentations concernant nos travaux sur la durée n'ont pas été fructueuses. L'ensemble des problèmes rencontrés peuvent être classés en deux catégories : 1) problèmes d'outils que ca soit pour le logiciel de segmentation de Régine André (problème de réglage des paramètres) ou bien pour le logiciel WavEdit (problème de la non-considération des pauses), et 2) le caractère générique de la base BDSONS ainsi que sa pauvreté pour des travaux de recherches spécifique au domaine de la reconnaissance du locuteur.
- Pour la fréquence fondamentale, cette dernière nous a permis de séparer les locuteurs en deux classes bien distinctes une classe pour les femmes (o sur les figures) et une classe pour les hommes (* sur les figures). Malheureusement, ce paramètre ne permet pas une discrimination nette entre locuteur d'où la nécessité de le coupler avec d'autres paramètres plus discriminant tel que les MFCC et leurs dérivées.
- Concernant les coefficients MFCC et leurs dérivées, nous avons travaillé avec l'approche connue sous le nom du « spectre moyen à long terme », c'est à dire on utilise des caractéristiques vectorielles mesurées uniquement par la valeur moyenne. La similarité entre deux jeux de mesure se réduit donc à un calcul de distance entre leurs valeurs moyennes. Malheureusement, le spectre moyen à long terme est très sensible aux variations du canal de transmission, ce qui le rend inutilisable dès que les locuteurs ne sont pas enregistrés dans un environnement complètement contrôlé (comme celui de la base BDSONS). En plus, l'ensemble des chercheurs s'entendent pour affirmer que ce sont les variations autour du spectre moyen qui sont les plus significatives : le spectre moyen à long terme étant trop sensible à la variabilité intersessions des caractéristiques du locuteur et aux variations du canal de transmission, il ne doit pas être pris en compte mais plutôt utilisé comme élément de normalisation.
- Pour les coefficients d'ADL, nous avons constaté que l'analyse discriminante linéaire permet de transformer l'espace d'origine vers un autre espace plus discriminant, sauf que cette analyse a été faite sur l'ensemble du vecteur d'entrée c'est-à-dire sans réduction de la dimension et sans élimination des paramètres qui dégradent les performances du système. Pour cela, il faut tenir compte des résultats obtenus par F-ratio sur chaque paramètre et qui donne une idée sur la mesure de performance de chaque paramètre individuellement. Une autre approche peut être appliquée et celle qui a été utilisée par (Paliwal, 1992). Elle consiste à calculer la performance de chaque paramètre, puis classer les paramètres selon la mesure de performance et enfin prendre un sous ensemble caractérisé par les premiers coefficients et contribuent plus à la discrimination entre les classes de locuteurs.

Synthèse et interprétation des expérimentations QSDAS

Cette partie illustre les deux volets: 1) la collecte d'une nouvelle base nommée QSDAS. Cette base qui vient répondre aux différents manques de la base BDSONS signalés dans le chapitre précédent, et 2) sur les tests réalisés sur cette dernière.

Nous avons commencé par évaluer les paramètres issus de différents espaces séparément en analysons dans un premier temps les paramètres acoustiques LPC, PARCOR, LPCC et MFCC utilisant une analyse F-ratio au niveau des paramètres de chaque jeu ainsi qu'une autre au niveau de chaque jeu de paramètres. C'est premiers tests montrent globalement les paramètres MFCC comme étant ceux qui ont les meilleurs F-ratio. Une deuxième série de tests sur les paramètres acoustiques a été faite en calculons les taux d'erreur de classification en utilisation la corrélation, la distance euclidienne et la distance euclidienne pondérée. Les meilleurs scores de cette série de tests sont détenus par le coefficient cepstraux LPCC et les coefficients MFCC, ce qui confirme leur utilisation dans la littérature de la reconnaissance du locuteur. Notons que la distance euclidienne pondérée surpasse la simple distance euclidienne et la corrélation.

Ces travaux ont été complétés par une étude comparative entre trois méthodes de fusion ou d'extraction des paramètres codant l'information pertinence du locuteur, à savoir : l'analyse en en composante principale (ACP), la recherche séquentielle directe (SFS) et l'analyse discriminante linéaire (ADL). Les résultats montrent que pour les paramètres concaténés, sans aucune fusion utilisant une de trois méthodes ACP, SFS ou ADL, l'inclusion de nouveaux paramètres n'améliore pas forcément les performances en réduisant le taux d'erreur. Pire encore, l'inclusion de paramètres supplémentaires peut même dégrader les performances du système. Les mêmes vecteurs de paramètres fusionnés par ACP ne donnent pas nécessairement de meilleurs résultats. Ceci est du principalement aux inconvénients de la méthode ACP qui en découlent du fait que l'erreur quadratique moyenne entres les vecteurs de données dans la projection des K-composantes de l'espace ACP n'est pas nécessairement pareil que minimiser l'erreur de classification. Par contre, les mêmes paramètres fusionnés par SFS ou ADL donnent de meilleur résultat, avec une nette avance pour la méthode ADL pour toutes les configurations. Cet avantage de la méthode ADL par rapport à la méthode SFS vient du fait que la méthode SFS ne permet pas de corriger les précédents ajouts de paramètres sélectionnés successivement. En plus, les paramètres ADL sont moins corrélés, ce qui n'est pas le cas des paramètres SFS.

Perspectives

Dans cette thèse, nous avons adressé quelques axes importants du domaine de la reconnaissance du locuteur, à savoir : 1) l'extraction des traits pertinents, et 2) la sélection et fusion des traits issus de deux espaces (acoustique et prosodique). Néanmoins, il reste d'autres axes qui peuvent faire l'objet de nos futurs travaux, parmi lesquels :

- L'extraction de nouveaux traits issus d'autres espaces (dialogal, haut niveau, …etc.) et d'autres types de caractérisation (ondelettes, SVM, …etc.) ;
- Utilisation d'autres méthodes de sélection heuristiques et évolutionnaires pour la fusion des données et la fusion des décisions des classificateurs.

Annexes

Annexe A : Listes des contributions scientifiques

1. Harrag A. et Mohamadi T., "Best fusing of acoustic and prosodic features: Application to speaker recognition", IEEE ICMCS'11 Internationale Conference, Ouarzazate Morocco, 7-9 April, 2011.
2. Harrag A. et Mohamadi T., "PCA, SFS or LDA: what is the best choice for extracting speaker features?", IJCA journal, Vol 15(3), pp. 1-3, February 2011.
3. Harrag A. et Mohamadi T., "QSDAS: New Quranic Speech Database for Arabic Speaker Recognition", AJSE journal, Vol 35(2C), pp. 1-12, December 2010.
4. Harrag A.,, Mohamadi T. et Serignat J.F., "LDA Combination and MFCC features in speaker recognition", IEEE INDCON'05, Vol 11-13, pp. 237-240, Chinnai India, December 11-13, 2005.
5. Mohamadi T., Hacine-Gharbi A., Mezaache S. et Harrag A., "Reconnaissance automatique de la parole par la méthode MSDTW", IEEE CCECE'01, Vol 5, pp. 527-530, Toronto Canada, May 13-16, 2001.
6. Mohamadi T., Hacine-Gharbi A., Mezaache S. et Harrag A., "Reconnaissance automatique de la parole par la méthode MSDTW", Séminaire National du Contrôle et Signaux SNCS, pp. 290-295, Djelfa Algérie, Octobre 30-31, 2001.
7. Mohamadi T., Mezaache S., Harrag A. et Hacine-Gharbi A., "Une base de données de la parole arabe BDSONS-Arabe et son système de gestion", IEEE ICMASI'00, Casablanca Maroc, Octobre 23-25, 2000.
8. Harrag A., Mohamadi T. et Moussaoui A., " B.D.P.A : Base de Données de la Parole Arabe ". Revue des Sciences et Technologies, Université de Constantine, Algérie, N° 10, pp. 73-83, Décembre, 1998.
9. Harrag A., Mohamadi T. et Moussaoui A., "Arabic Speech Database and Its Management System ". Proceedings de IEEE-SMC, CESA'98, IMACS Multiconference, Computational Engineering in Systems Applications, Vol. 4, pp. 534-537, Nabeul-Hammametn, Tunisia, April 1-4, 1998.

Annexe B : Base de donnée BDSONS

La "Base de Données des SONS" du Français (BDSONS) est une action menée sous l'égide du GRECO-PRC COMMUNICATION HOMME MACHINE depuis 1983. Elle répond à la demande des chercheurs de disposer d'une large base de sons, utiles tant pour l'étude de la langue française que pour les recherches en traitement automatique de la parole. Ces dernières présentent aujourd'hui deux pôles d'activité principaux : l'évaluation des algorithmes de reconnaissance et l'étude de nouveaux modèles en reconnaissance et en synthèse de la parole.

Les enregistrements qui composent BDSONS, ont été réalisés au CNET Lannion durant le premier semestre de l'année 1985. Les données représentent un volume global d'environ 3,5 Giga octets, stockés sur sept disques CDROM, BDSONS étant la première base de données du GRECO stockée sur CDROM. Les sept disques qui composent GRECO1 ont donc pour nom de volume ID : CD1_GRECO1 … CD7_GRECO1.

Tous les fichiers de sons enregistrés initialement au CNET Lannion, correspondaient au format standard GRECO qui se traduit par un fichier composé d'une entête binaire, d'une partie de signal et d'un listing sam. Ces fichiers ont été mis à la norme européenne d'enregistrement des fichiers de signal (PROJET ESPRIT Nø2589-SAM).

L'ensemble des corpus de BDSONS-GRECO1 forment deux groupes : un premier groupe de type "Evaluation" et un deuxième groupe de type "Acoustique". La fréquence d'échantillonnage du signal est de 16000 Hz, et le codage est fait sur 16 bits. Les cinq premiers disques compacts CD1_GRECO1 … CD5_GRECO1,contiennent les corpus de type "Evaluation" et les deux disques CD6_GRECO1 et CD7_GRECO1, regroupent les corpus de type "Acoustique" qui sont destinés à l'étude de la langue française.

Annexe C : Règles du Tajweed

The nosal tone (الغنة): it is a soft voice which comes from the nose appears when pronouncing certain letters. The duration of the nosal tone is two motions (حركتين).

The motion (الحركة): It is the time spent when you hold OR you open your finger.

Sukun or Jazm (السكون أوالجزم): A letter having motion is called mutaharik letter. If letter is without motion, it is called sakin letter and the sign (ْ) appears over it.

Shadda (ّ): If a sign shadda (ّ) appears over a letter, the letter is pronounced again. The first time having sukun and the second time with motion.

Tanween (التنوين): It is a convention that applies to the double damma, double fatha, or double kasra that appears on the last letter of some words (i.e. (حكيمٌ), (حكيماً), (حكيمٍ)). In fact tanween is considered as nun sakina (نْ) because it is pronounced as nun with sukun (ْ) (i.e. (حكيماً) is pronounced as (حكيمَنْ)).

The madd letters (حروف المد) : The madd letters are:

- The (اَ) which is preceded by a fatha.
- The (وُ) which is preceded by damma.
- The (ىِ) which is preceded by kasra.

Tajweed rules

Our aim here is to present a summary of the rules of tajweed to make it as simple as possible for beginners; we have presented only the significant and important rules. Those who have experienced should consult more rigorous books on this subject [25].

Rule 1: Meem (م) and noun (ن) with shadda(ّ)

Definition: the existence of shadda on meem (مّ) or noun (نّ).

Rule: the nosal tone (الغنة) should be heard in both cases

Examples: النّساء لجّنة، النّار، المزمّل، ثمّ،

Rule 2 : Meem with sukun (مْ)

We have to notice the letter which follows the meem with sukun (مْ). There are three cases:

Case 1: Merging (الإدغام)

Definition: The existence of another letter meem (م) after the meem with sukun (مْ)

Rule: Both meem with sukun (مْ) and the following meem (م) should be pronounced as one meem with shadda (مّ). The nosal tone (الغنة) should be heard in this case.

Examples: (أمْ مَن) pronounced as (أمَّن) with nosal tone, (كمْ مِن), (منكمْ مَريضا)

Case 2: Hiding (الإخفاء)

Definition: The existence of the letter ba (ب) after the meem with sukun (مْ)

Rule: Both the meem (م) and the ba (ب) should be pronounced but without closing the lips completely when pronouncing the meem (م). They are closed only when reaching the ba (ب). The nosal tone (الغنة) should be heard in this case also.

Examples: (آمنتم بِالله), (فاحكم بَينهم), (وفي ذلكم بَلاء).

Case 3: Appearance (الإظهار)

Definition: The existence of the rest of the letters after the meem with sukun (مْ)

Rule: Both the meem (م) and the following letter should appear and be pronounced normally. The nosal tone (الغنة) should NOT be heard in this case.

Examples: (ذلكم الله ربكمْ له الملك), (لعلكمْ تتقون).

Rule 3: Nun with sukun (نْ) and tanween (التنوين)

We have to notice the letter which follows the nun with sukun (نْ) and tanween. There are four cases:

Case 1: Appearance (الإظهار)

Definition: The existence of any of the following six letters after the nun with sukun (نْ) or tanween: (ء), (هـ), (ح), (خ), (ع), (غ).

Rule: The nun with sukun (نْ) or tanween (التنوين) should be pronounced normally. The nosal tone (الغنة) should NOT be heard in this case.

Examples: (حليماً غفورا), (منْ عمل), (عليمٌ خبير), (تنْحتون), (ينْهون), (آتيةٌ أكاد), (إنْ أنا).

Case 2: Conversion (الإقلاب)

Definition: The existence of the letter ba (ب) after the nun with sukun (نْ) or tanween.

Rule: The nun with sukun (نْ) or tanween should be converted into a meem with sukun (مْ). In this case the rule of meem with sukun (مْ) (Rule 2, Case 2) should be applied here.

Examples: (مِنْ بعد) first convert to (مِمْ بعد) and then apply Rule 2-Case 2, (عليمٌ بذات الصدور), (سميعٌ بصير), (ينْبت لكم).

Case 3: Merging (الإدغام)

Definition: The existence of any of the following six letters after the nun with sukun (نْ) or tanween: (ى), (و), (ن), (م), (ل), (ر). These letters are combined in word yarmelun (يرملون).

Rule: Both the nun with sukun (نْ) or tanween and the following letter should be pronounced as one letter with shadda. The nosal tone (الغنة) should be heard for only four letters out of the six. These four letters are combined in the word yanmu (ينمو). The nosal tone should NOT appear for the remaining two letters ra (ر) and lam (ل).

Examples: (منْ يقول) pronounced as mayqul (ميْقول) with nosal tone, (قولٌ معروف), (منْ ربهم) pronounced as (مِرَّبهم) without nosal tone, (هدىً لِلمتقين).

Case 4: Hiding (الإخفاء)

Definition: The existence of any of the rest of the letters after the nun with sukun (نْ) or tanween.

Rule: The nun with sukun (نْ) or tanween should be pronounced midway between the appearance (الإظهار) and merging (الإدغام). The nosal tone (الغنة) should be heard in this case. You are to extend the letter before the nun (ن) during the nosal tone (الغنة) period. Your mouth should be changed to take the shape of the next letter during this period. The nun (ن) is hidden in this time period.

Examples: (عذابٍ شديد), (قولاً كريما), (أيامٍ ثُم), (مُنْقلبَا), (مَنْ ذا الذى), (أنْ صدوكم).

Rule 4: Strong and soft letters (حروف التفخيم والترقيق)

A strong letter is a letter that should be magnified and make strong or thick when pronounced. While soft letter is a letter that should be softened and made fine or thin when pronounced. The Arabic letters are classified into four categories:

Case 1: Strong letters (حروف التفخيم أو الإستعلاء)

Definition: The existence of any of the following seven letters: (ظ), (ق), (ط), (غ), (ض), (ص), (خ). These letters are combined I three words (قظ) (ضغط) (خص).

Rule: These letters are always strong. They should be magnified when pronounced. If you circle your lips when pronouncing them, it will help.

Examples: (بَطِرت), (الإصْباح), (فتصْبحوا), (يضْربون), (يصُدون), (الصَلاة), (خَالدين).

Case 2: Soft letters (حروف الترقيق أو الإستقال)

Definition: The existence of any of the rest letters except for ra (ر).

Rule: These letters are always soft. They should be made fine or thin when pronounced. If you flatten your lips when pronouncing them, it will help.

Examples: (يغفِر), (إصبِر).

Case 3: The majestic lam (لام الجلالة)

Definition: The existence of the majestic lam (ل), i.e., the word allah (الله) or allahum (اللهم).

Rule: It should be magnified when it comes after fatha or damma.

Examples: (يطيعُ الله), (صدقَ الله).

It should be softened when it comes after kasra.

Examples: (قُلِ اللهم), (بِسمِ الله).

Case 4: The letter ra (ر)

Definition: The existence of the letter ra (ر).

Rule: It is sometimes pronounced as a strong letter and sometimes as pronounced as a soft one.

1. It is magnified (strong letter):

- If it has a fatha or a damma on it: (رَحيم)

- If it has a sukun after a fatha or damma: (يرْزقون)

- If it has a sukun after hamzat wasl: (من ارْتضى)

- If it has a sukun after a kasra and before a high letter: (strong letter) (*)

2. It is softened (soft letter):

- If it has a kasra on it: (الغارِمين).

- If it has a sukun after a kasra: (فِرْعون) (**)[12]

Rule 5: The shaking letters (حروف القلقلة)

Definition: The existence of any of the following five letters: (ذْ), (جْ), (بْ), (طْ), (قْ). These letters are combined in two words (جد) (قطب).

Rule: These letters should be shaken or rebounded when pronounced.

Examples: (الحريقْ), (محيطْ), (يُبْدئ), (البروجْ), (الخلودْ).

---

[12] (**) is an exception of rule (*) and there are only five cases in whole Quran that correspond to this rule: (لبالمِرْصاد), (مِرْصادا), (إرْصادا), (فِرْقة), (قِرْطاس).

Rule 6: The Whispering letters (حروف الهمس)

Definition: The existence of any of the following ten letters: (ت), (ك), (س), (ف), (ص), (خ), (ش), (هـ), (ث), (ح). These letters are combined in three words (فسكت) (شخص) (حثه).

Rule: The breath should run along with these letters when pronouncing them.

Examples: (أَهْل), (كَذَّبَتْ), (مسْكين).

Rule 7: Merging similar letters (إدغام الحروف المتماثلة)

Definition: The existence of two subsequent and similar letters. The first one should be have a sukun.

Rule: Both letters should be pronounced as one letter with shadda.

Examples:

- (إِذْ ذَهبَ) pronounced as (إِذْهبَ).- (ربحتْ تِجارتهم) pronounced as (ربحتِّجارتهم)

- (يلهثْ ذَلك) pronounced as (يلهذَّلك)- (لقدْ تَاب) pronounced as (لقتَّاب)

- (ودتْ طَائفة) pronounced as (ودَّطائفة)- (إركبْ مَعنا) pronounced as (إركمَّعنا)

- (قلْ لهم) pronounced as (قلَّهم) - (قلْ رَبِّ) pronounced as (قرَّبِّ)

- (يدركْكُم) pronounced as (يدركُّم)- (نخلقْكم) pronounced as (نخلكُّم)

Rule 8: Extension (المد)

Definition: The appearance of a hamza (ء), shadda (ّ), or sukun (ْ) , after any of the three madd letters (آ), (و), (ى).

Rule: Extend or lengthen the sound of the madd letters by a duration of two, four or six motions (حركتين، أربع أو ست حركات), depending on the different cases. Although there are many cases for madd, we give here five cases only. The rest are in fact consistent with the natural extension (two motions), i.e., normal reading.

Case 1: Connected necessary extension (المد المتصل الواجب)

Definition: The appearance of a hamza after any of the three madd letters in the same word.

Rule: Extend or lengthen the sound of the madd letters by a duration of four motions (أربع حركات)

Examples: (سيئت), (أشداء), (السماء).

Case 2: Separated optional extension (المد المنفصل الجائز)

Definition: The appearance of any of the three madd letters at the end of a word followed by a hamza at the beginning of next word.

Rule: Extend or lengthen the sound of the madd letters by a duration of two (as in natural extension) or four motions (حركتين أو أربع حركات).

Examples: (يا أيهَا النَاس), (إنما أنَا), (بنى إسرَائيل).

Case 3: Compulsory extension (المد اللازم)

Definition: The appearance of a shadda or sukun after any of the three madd letters in the same word.

Rule: Extend or lengthen the sound of the madd letters by a duration of six motions (ست حركات).

Examples: (الطامّة), (الصاخّة), (الحاقّة), (حادّوا), (شاقّوا).

Case 4: Stopping sukun extension (المد العارض للسكون)

Definition: The appearance of any of the three madd letters before the last letter of a stopping word (a word that you stop on it so that its last letter has sukun).

Rule: Extend or lengthen the sound of the madd letters by a duration of two, four or six motions (حركتين، أربع أو ست حركات)

Examples: (عذاب أليمْ), (رب العالمينْ), (خالدونْ), (ذو انتقامْ)

Case 5: Letters extension (مد الحروف)

We have to notice the existence of the letters at the beginning of some Surahs. There are two situations.

First situation:

Definition: The existence of any of the following eight letters at the beginning of some Surahs: (ص), (ق), (ن), (م), (ك), (ل), (س), (ع). These letters are combined in two words (نقص) (عسلكم).

Rule: Extend or lengthen the sound of the madd letters, which appears when pronouncing any of the above letters, by a duration of six motions (ست حركات).

Examples: (حم), (ن), (ق), (كهيعص), (طسم), (الم).

Second situation:

Definition: The existence of any of the following six letters at the beginning of some Surahs: (ر), (هـ), (ا), (ط), (ى), (ح). These letters are combined in two words (طاهر) (حى).

Rule: Extend or lengthen the sound of the madd letters, which appears when pronouncing any of the above letters, by a duration of two motions (natural extension).

Examples: (حم), (كهيعص), (طسم), (الم).

الخلاصة

ووفقا لنظرية انتاج النطق والكلام ،يتم الكلام بحركية الحبال الصوتية تليها صياغة المسالك الصوتية تردد الصوت على الشفاه .وقد شكلت الخصائص الصوتية التي تمثل القناة الصوتية تطبيقها على نطاق واسع في التعرف على الأشخاص .وعلى الرغم من كشف النقاب عن أن معالجة تردد الحبال الصوتية تلعب دورا هاما في تحديد خصائص الناطق، وفائدة خصائص مصدر صوت في التعرف على الأشخاص، ولها أسلوب فعال لاستخراج الخصائص، لم تستغل بالكامل .الهدف من هذه الرسالة هو اقتراح أسلوب أو أكثر لاستخراج سمات الناطق من قاعدة بيانات صوتية، وهذا الاستخلاص تركز على إيجاد تمثيل أفضل للمعلومات ذات الصلة المحددة االناطق. لتحقيق هذا الهدف، نحاول الإجابة على الأسئلة التالية: كيف تمثل المعلومات بشكل فعال الناطقباستعمال الإشارات الصوتية؟ هل حقا يستحق أخذ في الاعتبار المعلومات من مصدر الصوت؟ كيف يمكن الاستفادة الكاملة من انصهار معلومات المصدر والمسالك الصوتية؟ لهذا قمنا بتحليل المعلومات المتعلقة بالتنغيم. كما ناقشنا جدوى معلومات من مصدر الصوتية في التعرف على الناطق، مع التركيز على وجه الخصوص، على تكامل معلومات الناطق المستمدة من مصدر الصوت والناتجة عن القناة الصوتية.

الكلمات الشائعة: تحديد هوية الناطق، قاعدة بيانات صوتية، خصائص التنغيم، خصائص المسالك الصوتية.

Résumé

Selon la théorie de la production de la parole, la parole est produite par la phonation des cordes vocales suivie par l'articulation du conduit vocal et du rayonnement aux lèvres. Les caractéristiques acoustiques représentant le conduit vocal ont été largement appliquées pour la reconnaissance du locuteur. Bien qu'il ait été révélé que la phonation glottique joue un rôle important dans la caractérisation du locuteur, l'utilité des caractéristiques de la source vocale pour la reconnaissance automatique du locuteur, ainsi que sa technique efficace d'extraction des caractéristiques, n'a pas été pleinement exploitée. L'objectif de la thèse est de proposer une ou des méthodes d'extractions des traits du locuteur à partir d'une base de données parole, cette extraction centrée sur la recherche d'une meilleure représentation de l'information pertinente spécifique au locuteur. Pour atteindre cet objectif, nous essayons de répondre aux questions suivantes : Comment représenter efficacement les informations spécifiques au locuteur à partir du signal vocal ? Est-il vraiment utile de prendre en compte l'information de la source vocale? Comment faire pour tirer pleinement parti de la fusion de l'information de la source vocale et de celle du conduit vocal? L'information spécifique au locuteur liée à la phonation glottique est analysée. L'utilité de l'information de la source vocale pour la reconnaissance du locuteur est discutée. En particulier, la complémentarité des informations locuteur, celle issue de la source vocale et celle issue du conduit vocal, est dressée.

Mots clés : caractéristiques du locuteur, paramètres acoustiques et prosodiques, fusion des données.

Abstract

According to the theory of speech production, speech is produced by phonation of vocal cords followed by the articulation of the vocal tract and radiation at the lips. The acoustic features representing the vocal tract have been widely applied to speaker recognition. Although it was revealed that glottal phonation plays an important role in speaker characterization, the usefulness of voice source characteristics for speaker recognition, and its efficient technique for feature extraction 'has not been fully exploited. The objective of this thesis is to propose one or more extraction methods of speaker features from a speech database, this extraction focused on finding a better representation of relevant information specific to the speaker. To achieve this goal, we try to answer the following questions: How effectively represent the speaker-specific information from the voice signal? Is it really worth taking into account information from the voice source? How to take full advantage of the information fusion of the vocal source and the vocal tract? The speaker-specific information related to the glottal phonation is analyzed. The usefulness of the information of the vocal source for speaker recognition is discussed. In particular, the complementarity of information the speaker, the voice coming from the source and the end of the vocal tract, is derived.

Keywords: speaker features, acoustic and prosodic features, feature fusing.