Université du Québec à Chicoutimi

THÈSE

Présentée à l'Université du Québec à Chicoutimi
Département des Sciences Appliquées

pour le grade de:
Doctorat en Ingénierie

Discrimination Parole/Musique et étude de nouveaux paramètres et modèles pour un système d'identification du locuteur dans le contexte de conférences téléphoniques

par
Hassan Ezzaidi

Soutenue le 4 Octobre 2002

Devant le jury composé de:

Directeur de thèse:	Jean Rouat	(Université du Québec à Chicoutimi et Université de Sherbrooke, IMSI),
président de jury :	René Chouinard	(Université du Québec à Chicoutimi),
Directeur de programme:	Marcel Paquet	(Université du Québec à Chicoutimi),
Examinateurs:	Douglas O'Shaughnessy	(Institut national de la recherche scientifique, INRS),
	Daniel Audet	(Université du Québec à Chicoutimi, ERMETIS),
	Luc Morin	(Université du Québec à Chicoutimi, ERMETIS).

Bibliothèque
Paul-Émile-Boulet

UQAC
Université du Québec
à Chicoutimi

Mise en garde/Advice

Afin de rendre accessible au plus grand nombre le résultat des travaux de recherche menés par ses étudiants gradués et dans l'esprit des règles qui régissent le dépôt et la diffusion des mémoires et thèses produits dans cette Institution, **l'Université du Québec à Chicoutimi (UQAC)** est fière de rendre accessible une version complète et gratuite de cette œuvre.

Motivated by a desire to make the results of its graduate students' research accessible to all, and in accordance with the rules governing the acceptation and diffusion of dissertations and theses in this Institution, the **Université du Québec à Chicoutimi (UQAC)** is proud to make a complete version of this work available at no cost to the reader.

L'auteur conserve néanmoins la propriété du droit d'auteur qui protège ce mémoire ou cette thèse. Ni le mémoire ou la thèse ni des extraits substantiels de ceux-ci ne peuvent être imprimés ou autrement reproduits sans son autorisation.

The author retains ownership of the copyright of this dissertation or thesis. Neither the dissertation or thesis, nor substantial extracts from it, may be printed or otherwise reproduced without the author's permission.

Proverbes arabes

Si un savant trébuche, trébuchera dans sa chute tout un monde.

Un roi juste est l'ombre de Dieu sur la terre et un roi sans justice est une rivière sans eau.

UNIVERSITÉ DU QUÉBEC À CHICOUTIMI

Date: **le 12 Novembre 2002**

Auteur: **Hassan Ezzaidi**

Titre: **Discrimination Parole/Musique et étude de nouveaux paramètres et modèles pour un système d'identification du locuteur dans le contexte de conférences téléphoniques.**

Département **des sciences appliquées**

_____

Signature de l'auteur

Table des matières

Liste des tableaux

Table des figures

Remerciements

Je tiens d'abord à remercier de tout cœur mon directeur de recherche, M. Jean Rouat, pour m'avoir témoigné une confiance indéfectible au cours des années de cette thèse. Il m'a fait l'honneur de m'accueillir au sein de son groupe ERMETIS et de me faire profiter pleinement des multiples facettes de ses compétences scientifiques. Je le remercie également pour sa disponibilité, sa simplicité, sa grande sagesse et le souci incessant pour le bien et la satisfaction de ses étudiants. Un grand merci à sa femme Collette et à leurs enfants qui m'ont fait souvent oublier l'ennui et le stress du travail pour profiter de la chaleur d'un foyer familial. Enfin, sans lui, ce travail n'aurait pas été possible.

J'aimerai également remercier les membres de jury : MM. Douglas O'Shaughnessy, Daniel Audet et Luc Morin qui ont consacré une partie de leur temps précieux à examiner et corriger cette thèse. Je joins mes remerciements au président du jury M. René Chouinard et au directeur de programme, M. Marcel Paquet, à qui je dois beaucoup de respect. Évidement, je n'oublierai pas dans mes remerciements tout le personnel du département, surtout Mmes Chantale Dumas et Madeleine Potvin, pour les services administratifs ou techniques.

Enfin, je voudrais exprimer mes plus profonds remerciements à ma famille, ma mère Aziza Chinig pour son amour et son affection, mon frère M'hammed Ezzaidi pour ses encouragements et son soutien et ma jeune sœur Madiha, ma source d'inspiration, ainsi qu'à tous les membres de la famille. Je n'oublierai pas évidement ma conjointe, Chantale Lavoie, qui m'a supporté toutes ces longues années tout en me rendant une vie de couple très paisible, facile et joyeuse.

Remerciements au CRSNG, à la Fondation de l'UQAC, à CSE et à M. Karl Boutin pour le financement.

Résumé

La mise en oeuvre de systèmes de compréhension automatique de parole pouvant fonctionner dans des conditions réelles implique de reproduire certaines aptitudes de l'être humain. Outre les aptitudes à comprendre la parole même lorsqu'elle est corrompue par du bruit, nous sommes capables de tenir une conversation impliquant plusieurs interlocuteurs. Ce dernier point est lié au fait que nous identifions implicitement les interlocuteurs. Cette caractérisation du locuteur nous permet par exemple de réaliser des conversations téléphoniques en mode conférence. En plus de la reconnaissance du vocabulaire ou de l'identification du locuteur, on est également capable de distinguer les séquences de la musique (en alternance, en arrière plan, etc.) qui peuvent apparaître lorsqu'un des correspondants se place en mode attente.

En partant de ce contexte, on s'est intéressé à développer un système capable d'une part de discriminer entre les séquences de Parole/Musique et d'autre part d'identifier le locuteur dans des conditions téléphoniques fonctionnant en mode conférence avec une variabilité des combinés. Autrement dit, cette thèse s'intéresse à deux sujets du domaine du traitement de la parole. Le premier sujet porte sur la recherche de nouveaux paramètres pour améliorer les performances des algorithmes qui identifient les locuteurs en mode téléphonique. Le deuxième sujet est consacré à la proposition de nouvelles approches en discrimination de la parole, de la musique et de la musique chantée.

En discrimination du locuteur, on présentera une première étude visant à caractériser le locuteur par des paramètres AM-FM synchrones à la glotte, extraits à la sortie d'un banc de filtres cochléaires. L'objectif visé est de trouver de nouveaux

paramètres plus robustes aux bruits et à la variabilité des combinés téléphoniques. Comme résultats, on a obtenu des scores presque similaires entre le système proposé et le système de référence. Les meilleures performances ont été enregistrées lorsque le système utilise une architecture parallèle composée de deux reconnaisseurs qui se basent respectivement sur les paramètres MFCC et AM-FM. Dans le même cadre, on s'est intéressé à proposer une nouvelle technique de modélisation qui tient compte de la dépendance temporelle entre la source d'excitation et le conduit vocal. Avec les tests de courtes durées, on a obtenu de meilleures performances en comparaison à l'approche classique. Cependant, quand on augmente la durée de test, on obtient presque les mêmes performances pour tous les systèmes proposés.

En discrimination Parole/Musique, on a proposé deux systèmes, le premier utilise trois modèles paramétriques entraînés respectivement pour la parole, la musique et la musique chantée sans effectuer aucune normalisation sur les vecteurs paramètres. Sur une durée test de 100 ms, on a obtenu un taux de reconnaissance en moyenne de 93,77%. Le deuxième système ne requiert aucun entraînement et se base simplement sur un seuil pour effectuer la classification.

Introduction

L'être humain se caractérise par l'aptitude à reconnaître le contexte phonétique prononcé et également à identifier le locuteur (par exemple au cours d'une conversation téléphonique). Cette aptitude a donné naissance à différentes disciplines en recherche dont la reconnaissance automatique de parole, la synthèse de parole, la reconnaissance de la langue et la reconnaissance automatique du locuteur (RAL).

Le domaine de recherche en reconnaissance du locuteur, par une machine, s'intéresse particulièrement à extraire l'information acoustique caractéristique du locuteur, et qui se retrouve grandement noyée dans le signal vocal. Cette information, appelée empreinte vocale, servira comme indice biométrique d'identification d'un locuteur au même titre que l'empreinte digitale ou l'iris (images statiques) [59] [101].

L'intérêt essentiel de cette discipline, consiste à utiliser des systèmes de codage/décodage basés sur l'empreinte vocale via les supports de communication à accès public (téléphone, téléphone/IP), pour garantir la confidentialité, la virtualité (transparence) et l'authentification lors des transactions numérisées. Comme applications, on trouve les transactions bancaires, l'interrogation des bases confidentielles, l'accès à des services privilégiés, l'espionnage et la criminologie. On trouve également, la synthèse qui s'intéresse à incorporer et à doter les machines des caractéristiques humaines telles que le stress, la fatigue, le sexe, etc. Récemment, l'arrivée et l'émergence des applica-

tions multimédia exploitent autrement l'empreinte vocale par exemple pour effectuer le tri et l'archivage de façon automatique de la messagerie vocale [103] ou pour identifier les locuteurs impliqués au cours d'une conversation en mode conférence [58]. Le nouvel enjeu de ces applications est la nécessité de discriminer au préalable entre les portions de la parole, de la musique, du bruit, des silences,etc.

Les recherches en RAL réalisées progressivement depuis une vingtaine d'année, ont fait des progrès considérables et ont obtenu un taux d'identification quasiment parfait (100%) sur une population considérable (650 locuteurs). Cependant, cette performance a été atteinte avec de la parole enregistrée dans des conditions propres, tandis que pour la parole bruitée et/ou la parole téléphonique, les performances obtenues ne sont pas pour autant satisfaisantes. Parmi les facteurs responsables de cette dégradation, on trouve toujours le fameux bruit, l'effet du canal, la non linéarité des combinés téléphoniques et surtout leur variabilité.

Dans ce cadre, cette thèse s'intéresse en grande partie à la caractérisation du locuteur en milieu difficile et hostile en utilisant deux approches d'analyse qui exploitent principalement l'information du paramètre prosodique dit fréquence Fo (fondamental).

La première approche, concernant la paramétrisation, est basée sur des connaissances auditives, inspirées du mécanisme de fonctionnement de l'oreille moyenne. En fait, on applique au signal acoustique vocal, un banc de filtres cochléaires pour estimer la fréquence instantanée (FI) et la modulation d'amplitude (MA) à la sortie de chaque filtre. Avec un détecteur de la hauteur tonale, on localise les segments voisés qu'on suppose mieux caractériser le locuteur. En possession, d'une représentation en trois dimensions de la FI et MA, nous appliquons, de façon synchrone à la glotte,

différentes stratégies afin de trouver des paramètres susceptibles de caractériser le locuteur. On donnera les détails de cette perspective au chapitre 2.

La deuxième approche est une nouvelle technique qui tient compte de la corrélation source et conduit vocal pendant les phases d'entraînement et de classification d'un système d'identification du locuteur indépendant du texte. On suppose que le lien de dépendance obéit à une loi de probabilité conjointe dont on donnera les motivations et le formalisme mathématique avec l'ensemble des expériences réalisées dans le chapitre 3.

Au chapitre 4, on s'intéressera à un autre type de problématique qui vise essentiellement à discriminer de façon automatique entre les segments de musique et ceux de la parole. On proposera deux systèmes de discrimination Parole/Musique fonctionnant en temps réel. Ils se basent conjointement sur des paramètres standards, les Coefficients Mels Cepstraux (MFCC), utilisés dans la grande majorité des systèmes en traitement automatique de la parole. Le premier système ne requiert aucune session d'apprentissage et donne des décisions sur de courtes unités temporelles. Le deuxième système utilise les modèles paramétriques avec un mélange de Gaussiennes. Trois modèles sont entraînés, respectivement pour la parole, la musique et la musique chantée.

Quant au premier chapitre, il sera consacré à décrire les principaux concepts, méthodes et techniques publiés et utilisés dans ce domaine avec les références appropriées.

Chapitre 1

État de l'art des systèmes actuels en caractérisation du locuteur

Dans le présent chapitre, on décrit le mécanisme mis en jeu lors du processus de phonation. Ensuite, on donne quelques terminologies très utilisées dans le domaine. On exposera enfin un bilan bibliographique sur l'ensemble des travaux réalisés en identification du locuteur. La bibliographie concernant la problématique en discrimination Parole/Musique, ne fera pas l'objet de discussion dans ce chapitre. Par ailleurs, elle sera examinée dans le chapitre 4 qui sera consacré entièrement à ce sujet.

1.1 Introduction

Le signal acoustique de la parole est riche en information et en redondance. La redondance constitue sa robustesse contre le bruit environnant, les distorsions et les dégradations subies par le signal vocal. La richesse exprime les informations simultanées qui sont véhiculées par le contexte linguistique du message, les caractéristiques anatomiques, l'état personnel et les contraintes socio-culturelles du locuteur.

Le phénomène de la phonation implique la coordination et la mise en jeu de

plusieurs acteurs du système phonatoire. Principalement, on y trouve les poumons qui se comportent comme un générateur d'air qui servira ensuite à alimenter le larynx. Le volume d'air est directement lié à l'énergie estimée sur de courts intervalles à partir du signal acoustique. Au niveau du larynx, on retrouve les cordes vocales qui en mode vibratoire génèrent des ondes glottiques de formes triangulaires, asymétriques et périodiques. En mode non vibratoire, le signal à la sortie de la glotte correspond à des explosions de bruit. La cavité supra-glottique (conduit vocal et cavités nasales) est le dernier acteur qui sous l'excitation du signal glottique produit le signal acoustique vocal. Le volume d'air dans les poumons dépend essentiellement des dimensions du thorax, il varie donc avec le sexe et l'age de l'individu. La fonction du larynx dépend directement des propriétés myoélastiques, du couplage entre le larynx et les cavités sub et supra-glottiques de la tension et de l'interaction mécanique des cordes vocales. Pour plus de détails concernant l'anatomie et la morphologie du système, on peut consulter les ouvrages [22] [102].

Même si cette description du système phonatoire, ne tient pas compte de l'aspect perceptif, ni du fonctionnement intégral des organes anatomiques, il nous permet, tout de même, de retenir que le mécanisme mis en jeu au niveau du conduit vocal et cavités nasales semble moins complexe que la partie infra-glottique.

1.2 Terminologie

1.2.1 Identification et vérification du locuteur

On distingue deux approches pour reconnaître un locuteur. La première est la vérification du locuteur, qui consiste à valider l'identité réclamée par une personne à

partir d'un message vocal. La deuxième approche est l'identification du locuteur, qui consiste à déterminer l'identité d'un individu parmi un groupe de personnes.

1.2.2 Dépendance et indépendance du texte

En mode dépendant du texte, le texte prononcé par le locuteur, afin d'être identifié par le système, est le même que celui utilisé lors de la session d'apprentissage. En mode indépendant du texte, le locuteur est libre de dicter un texte de son choix. Les systèmes dépendant du texte se distinguent considérablement sur la nature et l'exploitation du texte en question [15]. En général les systèmes dépendants du texte donnent les meilleurs scores en reconnaissance.

1.2.3 Variabilité interlocuteur et intralocuteur

On définit par variabilité interlocuteur les caractéristiques qui sont propres à chaque locuteur et qu'on ne retrouve pas chez d'autres locuteurs (variabilité induite par le changement du locuteur). La grande variabilité entre les locuteurs est due, d'une part aux variations anatomiques des organes responsables de la production vocale et d'autre part à l'héritage linguistique et au milieu socioculturel de l'individu [18] [22] [29] [50].

La variabilité intralocuteur se présente lorsqu'un même locuteur prononce la même phrase ou le même mot à plusieurs reprises (parfois à différentes sessions) [18] [29] [50]. Plusieurs sources peuvent être attribuées à cette variabilité telle que la fatigue, le stress, le sommeil, l'horaire de la journée (matin, soir), le débit d'élocution, l'état émotionnel, etc.

1.2.4 Caractéristiques non anatomiques

Dans cette catégorie, on peut citer tous les éléments non liés à l'anatomie de l'individu et qui jouent un rôle influant sur la voix. On trouve le contexte, la syntaxe, l'intonation, l'accent régional, la durée syllabique, la façon d'articuler, le lieu géographique, etc.

1.2.5 Caractéristiques anatomiques

Il s'agit des organes ayant un rôle dominant pendant le mécanisme de production et qui sont considérés comme très caractéristiques du locuteur (les poumons, le larynx, le conduit vocal, les lèvres et les cavités nasales [22]).

1.3 Architecture du système en RAL

Un système d'identification du locuteur tel qu'illustré par la figure 1.1 se compose principalement de trois modules : paramétrisation, classification et décision.

Tout d'abord, le module de paramétrisation est un module d'analyse acoustique qui consiste essentiellement à réduire l'information du signal vocal en quantité et en redondance tout en augmentant la discriminibilité. À la sortie, le signal est représenté par un ensemble de vecteurs coefficients.

Ensuite, le module de classification consiste d'une part, lors de la phase d'apprentissage, à estimer les paramètres d'un modèle mathématique approprié pour les vecteurs de coefficients. On parle dans ce cas d'une modélisation paramétrique. D'autre part, lors de la phase de test, la tâche de ce module consiste simplement à évaluer une certaine métrique entre les vecteurs testés et les données stockées comme références

pour chaque locuteur.

Finalement, le module de décision, qui selon les critères établis, permet de désigner le locuteur reconnu.

Cependant, les systèmes en RAL se distinguent entre eux par l'approche adoptée dans l'implémentation et le design au niveau de chaque module. Leurs performances sont directement liées à la base de données (parole propre, bruitée ou téléphonique) utilisée, au nombre de locuteurs considérés, à la quantité de données d'apprentissage et à la durée de test choisie. On peut classer les systèmes de RAL en plusieurs catégories selon l'architecture adoptée, la tâche à effectuer et l'environnement envisagé. Il sera donc difficile de dresser un bilan comparatif de l'ensemble des travaux réalisés dans ce domaine, mais on se limitera à passer en revue les travaux les plus importants dans les prochaines sections.

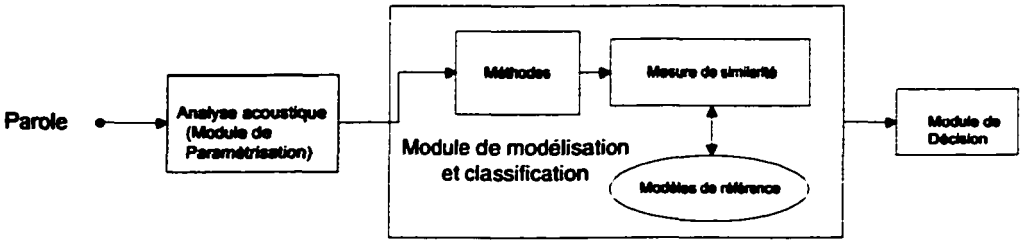

FIG. 1.1 - Schéma d'un système de reconnaissance du locuteur

1.4 Revue sur les méthodes de paramétrisation

Avant de faire une revue sur les travaux à propos de la paramétrisation, on propose de rappeler la définition des paramètres prosodiques, à temps-invariant (indépendants du temps) et à temps-variant.

1.4.1 Paramètres prosodiques

Dans plusieurs domaines d'analyse et de traitement du signal vocal, on définit par paramètres prosodiques :

- la fréquence fondamentale (vibration des cordes vocales) ;
- l'intensité de la voix (ou énergie) ;
- la durée successive des segments syllabiques.

Ces paramètres prosodiques 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 (age, 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 des 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.

1.4.2 Paramètres à temps-invariant et à temps-variant

On appelle paramètres à temps-invariant ceux qui sont estimés sur une fenêtre d'analyse de longue durée alors que les paramètres à temps-variant sont estimés sur une fenêtre d'analyse de courte durée.

Les paramètres à temps-invariant 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 [41] [7]. Pourtant, ils ne tiennent pas compte de la variabilité intralocuteur ce qu'il les rend facile à imiter par les imposteurs. Les paramètres à temps-variant 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.

1.4.3 Paramètres spectro-temporels et prosodiques

D'après Wolf [118], les paramètres idéaux et efficaces pour identifier un locuteur devront être facilement mesurables, difficiles à imiter par les imposteurs, robustes aux bruits et canaux de transmission et être très fréquents dans le signal vocal.

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 [5] [6] [7] a expérimenté une série de paramètres censés à représenter la contribution de différents organes mis en jeu lors du mécanisme de production (vibration des cordes vocales, volume d'air dans les poumons, résonances 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 fondamental 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 des conditions d'utilisation. Les articles suivants discutent d'une éventuelle paramétrisation dans le domaine spectral [29] [56] [76] [88]. On trouve les méthodes de PLP, RASTA, RASTA-PLP et MFCC [53] [54], qui se distinguent par l'incorporation de certaines propriétés perceptives. En effet, ces approches ont déjà été utilisées dans les systèmes en reconnaissance du vocabulaire et sont reconnues par leur robustesse à la variabilité intralocuteur et au contexte d'environnement. Openshaw et al [77] ont comparé la robustesse de ces méthodes dans la situation où les données sont affectées (entraînements 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-RASTA, etc. Ils confirment que l'utilisation du vecteur différence semble intéressant dans la seule situation où le rapport signal bruit est faible.

La méthode du bispectre (statistiques d'ordre supérieur) a été introduite par Wenndt et al [116]. 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 de l'absence du fondamental.

Bojan et al [17] 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 contraintes 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 voir la non-stationnarité du signal vocal.

L'enveloppe spectrale du résiduel est considérée caractéristique du locuteur [29]. 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. À ma connaissance seul l'article de Thévenaz et Hugli [113] a étudié la contribution du résiduel. Les résultats obtenus 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 [66] 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 parole ayant le même contexte phonétique du même locuteur. La deuxième matrice tient compte de la variabilité interlocuteur. 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 [57] 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, suivie par Teager [111] (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% 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 [2] [51] [70] [60] [81] [62] [119] [120] [121] [87].

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.

1.5 Techniques en modélisation/classification

La classification par spectre moyenné de Pruzansky [85] est considérée parmi les premiers travaux proposés en RAL indépendante du texte prononcé. D'autres travaux ont succédé, en moyennant d'autres types de vecteurs paramètres tels les coefficients cepstraux [38] [41], les coefficients LPC [97] [96], les coefficients de "Log Area Ratios", le contour du pitch [5] [6] ainsi que d'autres paramètres d'ordre statistiques [16] [46] [47] [65] [68] [69]. Ce genre d'approche est facile à implémenter mais il est considéré sensible au bruit [29] et ne tient pas compte des propriétés individuelles à court terme et de la dépendance des séquences temporelles.

D'autres travaux ont proposé la classification par réseaux de neurones, connue sous le nom de "méthodes connexionnistes" [10] [11] [13] [52] [95]. Cependant, cette approche nécessite beaucoup de temps pour l'apprentissage du réseau et elle est considérée moins flexible dans la mesure où il faut ré-entraîner le réseaux à chaque fois qu'un nouveau locuteur est ajouté [4](apprentissage discriminant).

Plus tard, les méthodes non-paramétriques ont pris place et sont devenues plus populaires. L'avantage de ces méthodes est qu'elles ne supposent aucune forme à priori sur la distribution des données ni une segmentation explicite (à priori). En effet, les données sont groupées par classe et chaque classe est représentée par son centroïde. Particulièrement, la quantification vectorielle introduite par Soong [106] inspirée des applications en codage de parole a été utilisée ensuite par plusieurs auteurs [19] [21] [52] [93].

Récemment, Campbell [23] confirme que l'utilisation de la matrice de covariance comme méthode de classification, en procédant par une analyse "Line Spectrum Pair" (L.S.P.) permet de réaliser des systèmes en RAL indépendants du texte avec un taux

de reconnaissance élevé pour des conditions non fortement bruitées. Elle est également facile à implémenter en temps réel.

Les méthodes qui supposent une forme à priori sur la distribution des données sont dites méthodes paramètriques. Poritz [83] a utilisé un modèle HMM à 5 états pour caractériser les vecteurs paramètres et prendre en compte la séquence temporelle. La modélisation de chaque état d'un modèle HMM par un mélange de Gaussiennes a été proposée par Tishby [115].

Grâce aux travaux de Matsui et Furui [72], une première comparaison entre la VQ et les HMM discrets et continus a été mise en évidence pour la RAL indépendante du texte. Il ressort que les meilleures performances sont partagées par la VQ et les HMM continus. Cependant, les auteurs mentionnent que les performances pour les HMM sont directement liées au nombre de mixtures par état sans pour autant être affectées par le nombre d'états utilisés. Reynolds [89] propose dans sa thèse, un système à un seul état avec un certain nombre de mélanges de Gaussiennes qui au fil des années devint le modèle le plus utilisé et le plus performant [47] [8] [88] [91] jusqu'à nos jours en RAL comparativement à ce qui existe sur le marché actuel.

1.6 Méthodes d'adaptation et de normalisation

L'adaptation des modèles au locuteur recouvre la problématique liée au vieillissement ou à la mise à jour des modèles et plusieurs travaux ont été publiés [19] [49] [42] [73]. Les techniques de normalisation sont utilisées comme approche pour compenser les effets liés aux conditions d'environnement et d'utilisation (bruit, bande téléphonique, variabilité de combiné, etc.). On distingue deux types de normalisation, la première s'effectue dans le domaine de paramétrisation et la deuxième dans le domaine de

classification [40]. En procédant avec une fenêtre d'analyse assez large, la normalisation dans le domaine paramétrique consiste souvent à soustraire la moyenne de l'ensemble des vecteurs coefficients, à partir de chaque vecteur de coefficients estimés sur chaque trame. Cette technique s'est avérée efficace pour réduire l'effet du canal et les variations à long terme du spectre [6] [39]. Elle est spécialement adaptée pour les applications dépendantes du texte dont la durée des mots prononcés est assez longue. Soong et Rosenberg montrent que la dérivé de premier ordre des coefficients cepstraux semble plus robuste à la variabilité linéaire du canal. Higgins et al [55] proposaient une autre approche censée à réduire l'effet du canal et du bruit. Ils utilisent un banc de filtres dont la fréquence centrale est distribuée selon une échelle non linéaire. L'originalité du traitement consiste à compresser et normaliser les composantes spectrales à la sortie du banc de filtres. Ensuite pour chaque trame, une différence est effectuée entre les canaux adjacents. Gish [46] avait démontré qu'un simple filtrage (avec un filtre fixe) des différents enregistrements téléphoniques permet d'améliorer considérablement les performances du système en RAL.

La normalisation dans le domaine de classification est surtout dédiée aux applications de vérification de locuteurs [40]. Gish et al (1995) ont utilisé une fonction de densité Gaussienne pour modéliser l'effet du canal en disposant d'un volume assez important des données d'apprentissage.

D'après Reynolds [91] la variabilité quand à l'utilisation des combinés téléphoniques représente le facteur principal à la dégradation des performances. Les articles suivants ont proposé une série de techniques pour réduire l'effet du combiné [9] [44] [45] [75] [91].

1.7 Modèles hybrides

Les modèles hybrides correspondent à une autre alternative pour la conception de système robuste aux différences entre sessions d'entraînement et de test. Le principe de ces modèles repose sur le fait de faire coopérer différents classificateurs (systèmes) montés en cascade et/ou en parallèle pour la prise d'une seule décision finale en faisant combiner les scores à la sortie de chaque classificateur. En effet, chaque classificateur utilise son propre espace de représentation pour coder des informations partielles à partir du signal acoustique. Par exemple, on peut faire coopérer deux systèmes, le premier se base sur les coefficients MFCC et le deuxième se base sur le mouvement des lèvres ou la prosodie, etc. L'enjeu majeur de ces approches est la technique employée pour homogénéiser et pondérer les scores à la sortie de chaque classificateur. On peut consulter la thèse de Besacier [14] qui donne une bonne revue sur ces modèles avec un nombre assez important de références bibliographiques.

1.8 Tendances actuelles

Malgré le nombre important de travaux effectués dans ce domaine, les solutions apportées restent relatives car elles ne peuvent pas être appliquées dans n'importe quelle condition d'utilisation.

Toutefois, on retient que l'utilisation des coefficients MFCC en paramétrisation et les modèles GMM en classification demeure la combinaison offrant le plus de performance dans les systèmes actuels en RAL. Pourtant, on avait mentionné que les paramètres MFCC sont particulièrement sensibles aux bruits et à la variabilité due aux conditions d'enregistrement et de transmission. En plus, les modèles GMM sont

des modèles statistiques qui ne tiennent pas compte de l'information de la séquence temporelle. Ceci, reste encore la motivation de plusieurs auteurs à la recherche de nouveaux paramètres ou de retrouver des modèles plus robustes. On peut citer la caractérisation par des paramètres prosodiques (le pitch et les formants) dans [51] ou par des paramètres d'ordre statistique dans [87] [60]. Par ailleurs, Andrews et al [2] ont introduit un nouveau système qui se base sur la séquence phonétique-dialectique. Kajarekar et Hermansky [61] ont proposé un système qui se base sur 4 catégories phonétique en RAL. D'après les auteurs, les voyelles, les "diphtongues" et les fricatifs sont les catégories les plus caractéristiques en vérification du locuteur.

L'effet de codage de la parole utilisé dans les réseaux téléphoniques (GSM, G.729, G.723, MELP "Mixed Excitation Linear Prediction"), sur les performances des locuteurs a été aussi l'objet d'étude pour plusieurs auteurs. Dans des conditions impliquant la variabilité des combinés téléphoniques, Dunn et al [30] observent une dégradation en performance suivant la qualité du codage.

1.9 Conclusion

Dans ce chapitre, on a présenté un état de l'art et les principaux aspects et concepts utilisés en RAL. On a présenté la structure générale d'un système en RAL et ses composantes 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.

Comme récents défis, on trouve le problème de détection de l'activité vocale de chaque locuteur au cours d'une conversation sans avoir à priori une information sur les locuteurs ni sur leur nombre. Je cite le travail de Meignier et al [74] qui ont introduit une approche d'apprentissage itérative et adaptative basée sur un modèle HMM pleinement connecté. Le principe consiste à ajouter un état au modèle HMM à chaque fois qu'un nouveau locuteur est détecté. Chaque état est représenté par un mélange de Gaussiennes. Cependant, cette approche nécessite la possession de la conversation intégrale avant le début de traitement.

Chapitre 2

Caractérisation du locuteur par Modulation d'Amplitude et par Fréquence Instantanée synchrone à la glotte

Ce chapitre est entièrement consacré à l'étude de la forme et de la distribution spectro-temporelle de l'enveloppe et de la fréquence instantanée (ou la phase) dans l'optique de trouver de nouveaux paramètres en caractérisation du locuteur.

On présente d'abord une étude préliminaire en reconnaissance automatique du locuteur qui nous servira à établir un lien entre l'information convoyée par la distribution spectro-temporelle et les propriétés caractéristiques du locuteur. Divers aspects liés à la distribution de l'enveloppe et de la fréquence instantanée ont été examinés pour cibler une technique d'extraction de paramètres robuste et efficace.

À l'issu, différents paramètres potentiels sont proposés et expérimentés sur l'ensemble de locuteurs de la base de données SPIDRE. À des fins de comparaison, les coefficients standards MFCC sont considérés, dans ce chapitre, les paramètres de référence. On discutera enfin l'intérêt et la potentialité d'une telle approche.

2.1 Introduction

Dans plusieurs travaux, en vue d'application à la parole, le signal acoustique vocal a été considéré comme un signal analytique, modulé en amplitude et en fréquence. En effet, la représentation de la parole par son enveloppe et sa fréquence instantanée est riche en information intégrant conjointement la structure spectrale et l'excitation temporelle. Particulièrement, elle tient compte d'un phénomène non linéaire lié à l'interaction entre la source et le conduit vocal pendant le processus de production [110] [112] [109]. D'après Rosen [92], l'enveloppe du signal analytique a été considérée comme une caractéristique temporelle véhiculant simultanément des informations acoustiques, auditives et linguistiques. Plusieurs auteurs ont exploité l'information MA-MF ou ses dérivées en vue d'application à la parole. En détection de mélodie, Foldvari [37] a utilisé la fréquence instantanée (FI) et l'enveloppe pour l'estimation du fondamental, alors que Demars et al [27] ont utilisé ces mêmes paramètres pour la reconnaissances des chiffres. L'utilisation de la FI seule a été utilisée par Qiu et al [86] pour détecter le voisement et également pour déterminer les instants de fermeture et d'ouverture de la glotte. Potamianos et al [84] ont proposé une modélisation MA-MF multi-bande, en démodulant les composantes par l'opérateur non-linéaire Teager, pour le suivi de la fréquence centrale et la largeur de bande des formants. Un autre type d'opérateur dit "Dyn" a été proposé par Rouat et al [94] comme technique de démodulation des composantes MA-MF. Simultanément, l'opérateur Teager a été utilisé pour la même raison par Bovik et al [20] et Maragos et al [67].

Récemment, Potamianos et Maragos [78] ont proposé des paramètres dérivés de la distribution MA-MF en reconnaissance automatique de la parole. En comparaison aux paramètres standards MFCC, des performances presque similaires entre les deux

approches, ont été observées.

Pour plus de détails, on invite le lecteur à consulter l'ouvrage de Demars [28], disponible gratuitement sur Internet, qui est consacré aux représentations bi-dimensionnelles (temps-fréquence) dont une bonne partie est réservée aux travaux publiés impliquant la modulation MA-MF et ses dérivés avec un nombre important de références.

2.2 Filtres cochléaires

En pratique, les filtres cochléaires sont censés reproduire de façon partielle le fonctionnement du système périphérique auditif, particulièrement au niveau de la membrane basilaire de la cochlée sur un intervalle de fréquence allant de 300 Hz à 5000 Hz. La conception de ces filtres a été inspirée des travaux de Patterson [79] et ceux de Glasberg et Moore [48]. La version numérique utilisée dans cette thèse, a été développée au sein de notre laboratoire par Y.C. Liu pour l'obtention de son mémoire [122]. Il s'agit de filtres à pentes exponentielles arrondis à leurs sommets dans le domaine spectral et caractérisés par la largeur de bande d'un filtre rectangulaire (ERB "Equivalent Rectangular Bandwidth"). Les filtres sont étroits en basse fréquence donc plus sélectifs en harmoniques et ils s'élargissent au fur et à mesure que la fréquence augmente. Donc, en moyennes et en hautes fréquences les filtres sont moins sélectifs et donnent lieu au phénomène dit "phénomène de battement".

2.3 Calcul de l'enveloppe et de la Fréquence Instantanée

On applique au signal vocal un banc de 24 filtres cochléaires dont la fréquence centrale f_{c_i} est répartie de 330 Hz à 4700 Hz. La sortie $s_i(t)$ à chaque canal i est un signal passe-bande à spectre étroit centré autour de f_{c_i}. Le signal $s_i(t)$ est considéré modulé en amplitude et phase avec une fréquence porteuse f_{c_i}, soit :

$$s_i(t) = A_i(t)cos[2\pi f_{c_i} t + \phi_i(t)]. \tag{2.3.1}$$

$A_i(t)$ est l'amplitude instantanée (enveloppe) et $\phi_i(t)$ la phase instantanée. La fréquence instantanée $F_i(t)$ est définie comme la dérivée par rapport au temps de la phase :

$$F_i(t) = f_{c_i} + \frac{1}{2\pi}\frac{d\phi_i(t)}{dt}.$$

Si $\widehat{s_i(t)}$ est la transformée de Hilbert du signal $s_i(t)$ et $s_i(t)'$ est la dérivée temporelle du $s_i(t)$, on peut déduire l'enveloppe $A_i(t)$ et la FI $F_i(t)$ du signal $s_i(t)$ à la sortie de chaque canal i :

$$A_i(t) = \sqrt{s_i(t)^2 + \widehat{s_i(t)}^2} \tag{2.3.2}$$

$$F_i(t) = \frac{s_i(t)\widehat{s_i(t)}' - s_i(t)'\widehat{s_i(t)}}{s_i(t)^2 + \widehat{s_i(t)}^2}. \tag{2.3.3}$$

2.4 Structure de l'enveloppe et de la Fréquence Instantanée

La structure ou représentation spectro-temporelle de l'enveloppe à la sortie du banc de filtres cochléaires est illustrée par la figure 2.1(a). Dans la figure 2.1(b), on illustre celle de la fréquence instantanée.

2.4.1 Interprétation

Comme hypothèse, on peut dire que pour chaque canal, l'enveloppe observée lors de la réponse du système phonatoire (cordes vocales et conduit vocal) peut être considérée dans un premier temps dominée par l'information glottique ("régime entretenu") puis ensuite par l'information du conduit vocal ("régime non entretenu").

En régime entretenu, le spectre instantané de la parole est suffisamment large vis-à-vis de la fonction de transfert du filtre cochléaire et du combiné téléphonique pour que les réponses impulsionnelles du filtre et du combiné puissent affecter fortement l'enveloppe lorsque celle-ci est mesurée lors de l'explosion glottale.

Par conséquent, on peut supposer pour l'instant que pour un banc de filtres donné, la forme de l'excitation glottale et l'information liée à la position des pics secondaires de l'enveloppe peut-être utile (caractéristiques du locuteur). Ces pics secondaires (le principal étant lié à l'impulsion glottale) sont dus aux battements des harmoniques, à l'influence des formants et peuvent donc caractériser à la fois le locuteur et ce qui a été prononcé.

Sur la base de cette interprétation de la structure MA-MF, la suite de ce chapitre sera consacrée à une série d'expériences exploratoires qui consistent principalement à trouver une approche adéquate pour dériver des paramètres caractérisant le locuteur à partir de l'enveloppe et/ou la fréquence instantanée.

2.4.2 Définition : l'unité élémentaire

On définit l'unité élémentaire comme étant un segment de durée To (période instantanée du signal) et qui inclut une impulsion glottale. L'unité élémentaire sera souvent utilisée dans la suite de ce chapitre pour désigner une analyse synchrone à la

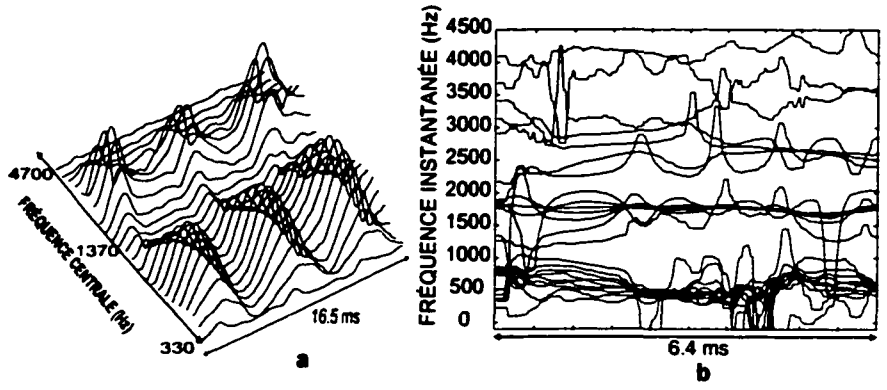

FIG. 2.1 **a** : Enveloppes à la sortie du banc de filtres ; b : FI à la sortie du banc de filtres.

glotte.

2.5 Étude préliminaire des enveloppes et de la FI à la sortie du banc de filtres cochléaires

Cette section portera sur quelques expériences exploratoires réalisées dont l'objectif consistera à mettre aux points des stratégies et des techniques adéquates pour coder de façon pertinente les caractéristiques du locuteur véhiculées par l'enveloppe ou la fréquence instantanée.

2.5.1 Analyse psycho-acoustique

Avant de commencer à proposer et à tester n'importe quel type de paramètres dérivés de l'enveloppe AM ou de la FI, on a trouvé raisonnable de vérifier auparavant si ces composantes en soit maintiennent toujours des propriétés caractéristiques du locuteur.

Comme données, on a utilisé pour cette expérience seulement les voyelles /a/ et /i/ pour 4 locuteurs provenant de la base de données française BDSONS. On s'est organisé de manière à faire en sorte que les mêmes mots des 4 locuteurs aient des durées temporelles presque identiques. Ainsi, on n'est pas obligé de passer par un alignement temporel qui, en quelque sorte, peut introduire des irrégularités au niveau d'amplitude et de phase si on veut resynthétiser le signal.

Le principe de la démarche employée consiste à sélectionner une paire de locuteurs quelconques pour extraire séparément de leurs signaux acoustiques l'amplitude instantanée $A_i(t)$ (i indique le numéro du canal) et la phase instantanée $\Phi_i(t)$ à la sortie du banc de filtres cochléaires. Ensuite, on resynthétise des signaux mixtes, en modulant l'amplitude instantanée du premier locuteur sur une onde porteuse dont la fréquence est la fréquence centrale du canal et dont la phase instantanée est celle du deuxième locuteur. Un traitement identique est effectué pour le deuxième locuteur. Autrement dit, on a effectué une permutation des composantes $A_i(t)$ des deux locuteurs. Finalement, une sommation des signaux modulés de tous les canaux nous permet de récupérer deux signaux acoustiques mixés.

En procédant par des écoutes psycho-acoustiques, on peut ainsi évaluer l'importance de $A_i(t)$ et $\Phi_i(t)$ en caractérisation du locuteur. D'autres tests ont suivi qui consistaient cette fois-ci à évaluer individuellement le rôle des $A_i(t)$ (avec $\Phi_i(t) = 0$) et $\Phi_i(t)$ (avec $A_i(t) = \textit{constante}$) sur chaque canal. Les écoutes psycho-acoustiques ont été effectuées par deux auditeurs travaillant dans le domaine.

Même si les résultats sont préliminaires et les remarques qu'on donne dans cette partie ne sont pas pour autant confirmées complètement. On garde de cette expérience

que l'information contenue par $\mathcal{A}_i(t)$ a tendance à être dominée par l'information linguistique. Par contre on trouve que la phase $\Phi_i(t)$ combine l'information caractéristique du contexte et celle du locuteur. Particulièrement, on est convaincu que $\mathcal{A}_i(t)$ et $\Phi_i(t)$ maintiennent certaines propriétés caractéristiques du locuteur, ce qui nous motive à poursuivre la direction de notre recherche.

Il sera donc intéressant de considérer la possibilité d'utiliser conjointement l'enveloppe $\mathcal{A}_i(t)$ et la phase $\Phi_i(t)$ (ou la fréquence instantanée) comme paramètres discriminants. On donnera dans la dernière section, les résultats et les discussions en exploitant cette approche.

2.5.2 Données dépendantes du contexte

On tient à noter que les données décrites ici, nous ont servi seulement à effectuer différentes études et analyses dans l'optique de cibler des paramètres à fort potentiels en caractérisation du locuteur. Une fois le choix sur le genre de paramètres fixé, on donnera en fin du chapitre les scores obtenus avec la base de données SPIDRE.

Pour l'instant, on a fixé le contexte phonétique et on a traité uniquement les segments voisés provenant de quelques mots fréquemment utilisés dans les conversations de la langue anglaise tels : "and", "is", "the", "this" et "for". Ceci nous donnera une bonne compréhension sur la variabilité intra- et inter-locuteurs et les effets indésirables introduits par les combinés et la ligne téléphonique. Les mots sont extraits de quatre conversations provenant de 3 combinés différents et pour cinq locuteurs (homme et femmes) : sp1415, sp1497, sp1499, sp1436 et sp1575 de la base SPIDRE. On s'est servi de la transcription phonétique fournie avec la base de données du Corpus SPIDRE pour l'extraction de mots.

2.5.3 Étude des différences de 'phase' entre les canaux cochléaires

On propose ici une première étude préliminaire basée sur des analyses graphiques de la structure spatiale de l'enveloppe. Précisément, elle s'intéresse à vérifier si la non-synchronisation (délai) de la réponse glottale entre les différents canaux du banc cochléaire peut être dépendante du locuteur.

Analyse

Au cours de l'analyse graphique des enveloppes, on a remarqué chez certains locuteurs que le déphasage (positif ou négatif) de l'impulsion glottale sur les 24 filtres du banc cochléaire semble être différente. Donc, on place ici l'accent sur la position relative des pics de l'enveloppe à travers les canaux et dans le temps. La figure 2.2 illustre le principe. En abscisse, on a reporté 25 ms de temps (croissant de gauche à droite). En ordonnée, on a placé les enveloppes des signaux cochléaires pour les canaux 8 (basses fréquences) à 19 (hautes fréquences). Les 24 canaux n'ont pas été représentés afin d'alléger les figures, mais ils entrent en ligne de compte dans la méthode développée. Le locuteur 1575 est tel que les réponses à l'excitation glottale sont à peu près alignées dans le temps. Le locuteur 1497 est tel qu'il y a décalage dans le temps entre les réponses en basses et hautes fréquences.

De façon générale, nous avons constaté que suivant les locuteurs, les maxima correspondant aux impulsions glottales ne se produisaient pas aux mêmes instants pour des canaux cochléaires différents.

L'interprétation possible serait que le spectre de l'excitation glottale est tel que le maximum d'énergie ne se produit pas de façon synchrone entre les bandes de

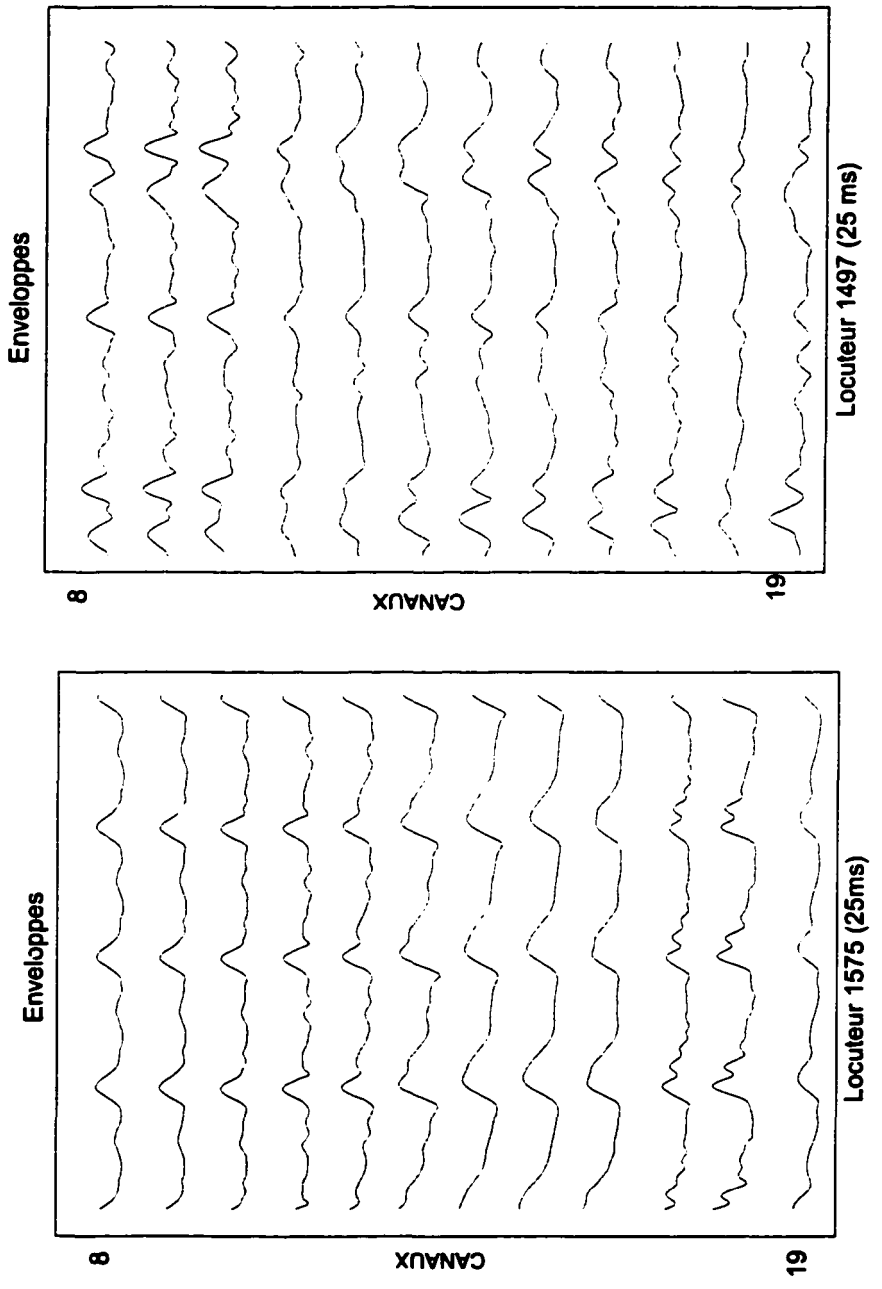

FIG. 2.2 - Enveloppes à la sortie du banc de filtres, filtres 8 à 19 ;

fréquence pour certains locuteurs. Entre autres, le déphasage entre harmoniques peut être différent et dépendant du locuteur et du contexte. Une autre interprétation serait qu'il existe un couplage entre la glotte et le conduit vocal. Cette dernière interprétation pourrait être dérivée des travaux de Teager. Celui-ci a en effet analysé les flux de pression dans le conduit vocal et aurait conclu au couplage non linéaire et à l'existence de mini tourbillons qui pourraient expliquer l'existence de 'bouffées d'énergies' dans le signal de parole [110] [112] [109]. Voyant ses travaux contestés, Teager a d'ailleurs essayé de mettre en valeur ces bouffées d'énergie à la sortie d'un banc de filtres Gaussiens. Malheureusement, son décès n'a pas permis de confirmer ou d'infirmer cette théorie qui demeure toujours très contestée.

Résultats et discussion

Nous avons extrait pour chaque canal cochléaire et pour chaque 'unité élémentaire', l'instant correspondant au maximum de l'enveloppe pendant l'explosion glottale. Ces instants ont été rapportés au temps origine 0 qui est défini comme étant le moment où l'enveloppe du canal 6 (non illustré sur les figures) est à son maximum (toujours pendant une unité élémentaire). Une compilation des résultats a été faite sous forme d'estimation de la moyenne de ces instants pour chaque canal et pour l'ensemble des mots 'and' des 4 locuteurs étudiés. La figure 2.3 présente les résultats sous forme graphique pour différents locuteurs. En abscisse, on a reporté les indices des canaux cochléaires (selon les fréquences centrales croissantes) et en ordonnée les moyennes pour chaque canal.

Il est possible de séparer visuellement certains locuteurs. Les locuteurs sp1415 et sp1436, par exemple sont clairement séparables alors qu'ils le sont difficilement sur la

base de leur fréquence de glotte. Il est important de souligner que les paramètres sont indépendants du combiné (il y a certaines fluctuations entre combinés, mais celles-ci correspondent souvent à des combinés pour lesquels il y a peu de données). En général, les courbes sont similaires pour un même locuteur et pour des combinés différents.

Conclusion

Des améliorations importantes à cette technique devront être apportées. Il y aura lieu d'étudier les distributions de probabilité de ces paramètres et non pas uniquement les moyennes. Une estimation des histogrammes pourrait s'avérer intéressante. Ces paramètres semblent à première vue intéressants et il y aura lieu de les étudier de façon plus approfondie ultérieurement. En effet, on préfère reprendre l'analyse et rechercher d'autres techniques qui pourront être intéressantes, simples et plus prometteuses.

2.5.4 Caractérisation par les lobes secondaires de l'enveloppe

Dans cette sous-section, on passe à une deuxième étude préliminaire déduite également par les analyses graphiques de la structure spatiale de l'enveloppe. Dans ce cas, on cherche à étudier les retombés d'une éventuelle paramétrisation utilisant les lobes secondaires à partir de l'enveloppe pour caractériser le locuteur à la sortie de chaque filtre cochléaire.

Analyse

L'enveloppe des signaux cochléaires est habituellement modulée en amplitude. Les modulations sont en principe plus fortes dans les canaux cochléaires soumis aux battements dus à la résonance de 2 harmoniques provenant de 2 formants différents.

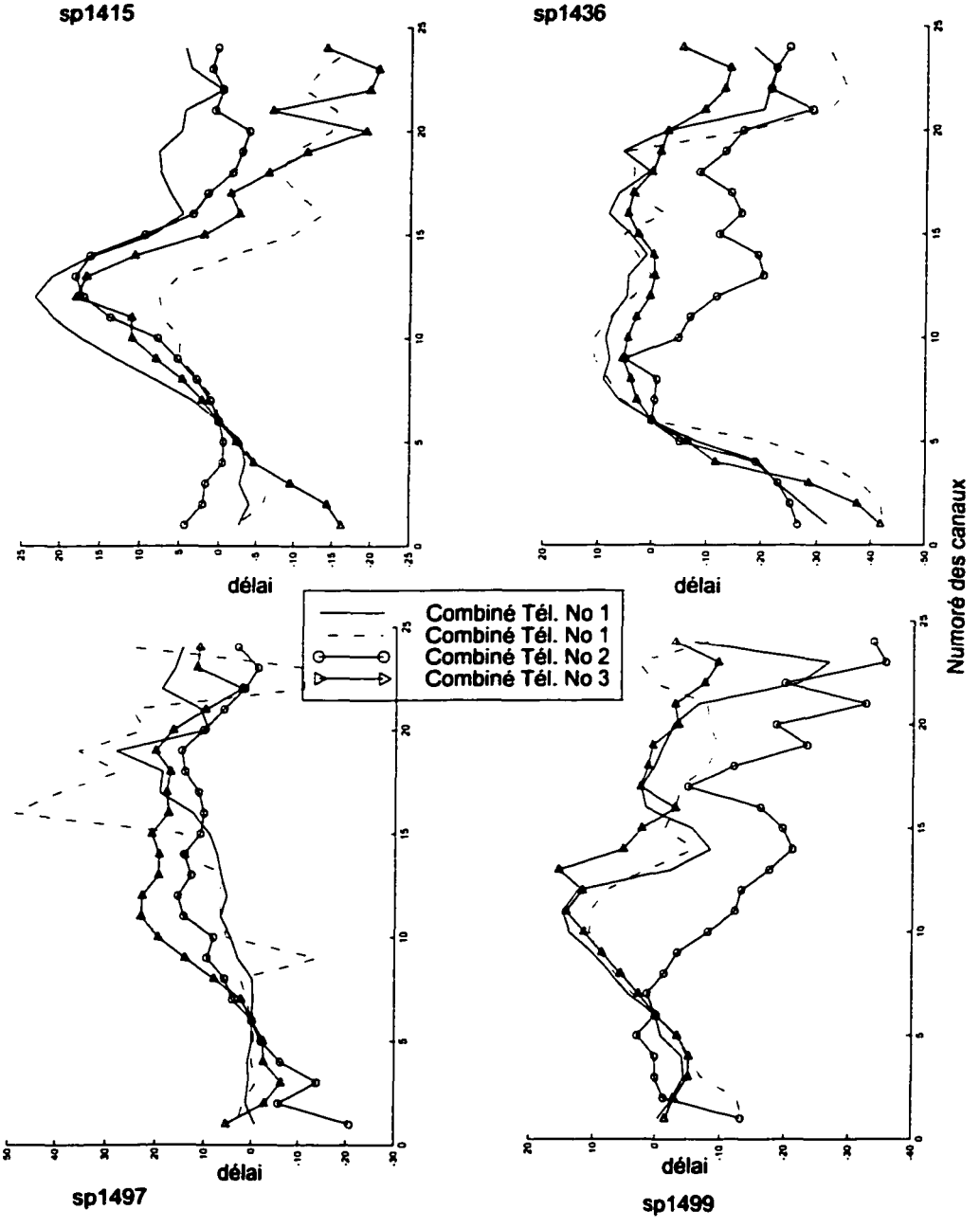

FIG. 2.3 - Différence de phase entre canaux à la sortie de l'enveloppe (pour 4 locuteurs)

Ces battements sont présents dans les canaux pour lesquels il n'y a pas de résolution des harmoniques (canaux cochléaires à partir de 1000 Hz ou 1200Hz). En dessous de 1000 Hz, on estime en général que les filtres cochléaires sont en mesure de résoudre les harmoniques. En conséquence, on pourrait supposer que le codage de l'information se fait principalement en modulation d'amplitude en moyenne et haute fréquence (en MF pour les canaux à harmoniques résolus). Il est important de noter que la fréquence glottique va affecter de façon importante la modulation due aux battements. En effet, pour un filtre cochléaire donné, un homme aura plus d'harmoniques présents dans la bande passante que pour une femme. Cette différence va affecter la position des pics secondaires (pour une même fonction de transfert du conduit vocal).

Par ailleurs, l'apparition de pics secondaires dans les enveloppes n'est probablement pas due uniquement aux battements, mais aussi aux retards respectifs des réponses à l'explosion glottale pour les formants.

Conclusion

Pour l'instant d'après les analyses graphiques, on s'est rendu compte que la forme et la position des pics secondaires ne peuvent pas être utilisées seules pour l'identification du locuteur. Cependant, elles peuvent servir dans un premier temps, pour effectuer une préclassification par groupe de locuteurs. Ensuite, on peut se baser sur d'autres types de paramètres pour discriminer séparément les locuteurs à l'intérieur de chaque groupe (en espérance d'une meilleure performance). Dans ce cas, il y a donc lieu de tenir compte de l'homogénéité des paramètres et de la pondération des scores. Par homogénéité (non homogène) on veut dire définir l'approche pour regrouper les différents paramètres qui appartiennent à des espaces différents. Et par pondération,

on veut dire, trouver la stratégie de pondérer les scores fournis par les différents classificateurs.

2.5.5 Étude des signaux différence des "unités élémentaires"

Dans cette sous-section, on propose une troisième étude préliminaire déduite aussi par les analyses graphiques de la structure spatiale de l'enveloppe. Dans ce cas, on cherche à étudier si une paramétrisation synchrone peut être dépendante du locuteur.

On rappelle que l'unité élémentaire est le segment de durée T_0 (période instantanée du signal) et qui comprend une réponse à l'impulsion glottale.

Analyse

À priori, d'une excitation glottale à l'autre, la contribution du combiné est relativement constante (le combiné est inchangé) tandis que des changements plus importants peuvent se produire en fonction des caractéristiques du locuteur et du contexte phonétique. On pense donc, qu'il serait intéressant d'estimer les caractéristiques du combiné lors de l'explosion glottale et d'en tenir compte pour corriger le signal. En faisant l'hypothèse que l'influence n'est pas la même au cours du temps, il y aurait lieu de mettre en place une technique d'analyse synchrone à la glotte. Il existe, bien sur déjà plusieurs techniques dont l'optique est de faire une estimation statistique des caractéristiques du combiné en étudiant des signaux dont la distribution recouvre l'ensemble des contextes phonétiques et des locuteurs. Malheureusement, ces techniques font appel à des traitements sophistiqués qui sont difficiles à mettre en place pour une application réelle. Par ailleurs, les analyses qui supposent la stationnarité du signal (LPC, analyse de Fourier, etc.) ne peuvent qu'effectuer une estimation moyenne et

peuvent difficilement séparer les caractéristiques observées en 'régime entretenu' de celles qui le sont en 'régime non-entretenu'.

Vers une paramétrisation

On commence par placer une fenêtre chevauchante, dynamique et glissante sur l'enveloppe (ou la fréquence instantanée) de chaque canal. Chaque fenêtre correspond à une unité élémentaire, c-à-d la période instante, du signal T_0. L'unité élémentaire est de durée variable et est mise à jour à chaque impulsion de glotte. En effet, on utilise le système de la détection de la hauteur tonale pour estimer la période T_0 à tous les 10 ms qui servira ensuite pour localiser chaque portion voisée du signal vocal. Donc, on localise les unités élémentaires $seg_k = (x_1, x_2, .., x_N)$ où l'indice k caractérise la position de l'unité dans le temps. Les données x_k sont les composantes d'enveloppe de l'unité k sur une période T de N points. On soustrait ensuite l'unité de position T_{j-1} à l'unité positionnée en T_j. Le traitement permet ainsi d'obtenir le signal différence entre deux unités élémentaires adjacentes. Le signal différence a une longueur variable qui correspond à la période de glotte des 2 unités élémentaires. Les signaux différences calculés sont notés par :

$$d_k = seg_k - seg_{k-1} \tag{2.5.1}$$

À partir de ce signal différence, il est possible d'obtenir divers paramètres dont l'objectif principal est de réduire la dimension. Pour nos premiers tests, on a utilisé quelques variantes aux vecteurs différences qui sont illustrées par les équations 2.5.2, 2.5.3, 2.5.4, 2.5.5. L'équation 2.5.3 mesure la variation absolue aux vecteurs différences. L'équation 2.5.4 découle directement de 2.5.3 et consiste à utiliser l'énergie. L'équation 2.5.5 mesure la valeur absolue de la variation du carré des différences entre 2 unités

élémentaires consécutives. Quant à l'équation 2.5.2, elle mesure l'énergie de l'enveloppe (ou la moyenne de la FI) à chaque explosion glottale, et on reviendra à l'analyse de ce paramètre dans la prochaine section.

$$P1_k = \frac{\sum_{n=1}^{N} seg_k(n)}{N} \tag{2.5.2}$$

$$P2_k = \frac{\sum_{n=1}^{N} abs(d_k(n))}{N} \tag{2.5.3}$$

$$P3_k = \frac{\sum_{n=1}^{N} d_k(n) d_k(n)}{N} \tag{2.5.4}$$

$$P4_k = \frac{\sum_{n=1}^{N} abs(d_k^2(n) - d_{k-1}^2(n))}{N}. \tag{2.5.5}$$

Résultats et discussion

Avec de telle stratégie, on est censé réduire l'effet du canal, puisqu'on traite l'information dans la même largeur de bande (sur chaque canal) et au même instant près. Même si le contexte phonétique et les caractéristiques du locuteur seront également réduits, on supposerait observer la variabilité liée à la source et/ou le conduit vocal au cours du temps.

Le graphique 2.4 illustre le vecteur moyen des paramètres estimés par les équations 2.5.3 à 2.5.5 pour différents combinés et pour un même locuteur.

Le graphique 2.5 illustre le vecteur moyen des paramètres estimés par l'équation 2.5.5 pour différents combinés et pour différents locuteurs.

On constate que les moyennes ne sont pas influencées par le type du combiné et le sont encore moins que l'étude reportée sur le délai de la phase. On remarque que les paramètres extraits des différents équations manipulent presque la même information. Les courbes des 4 combinés ont une forme commune qui est en fait due au contexte

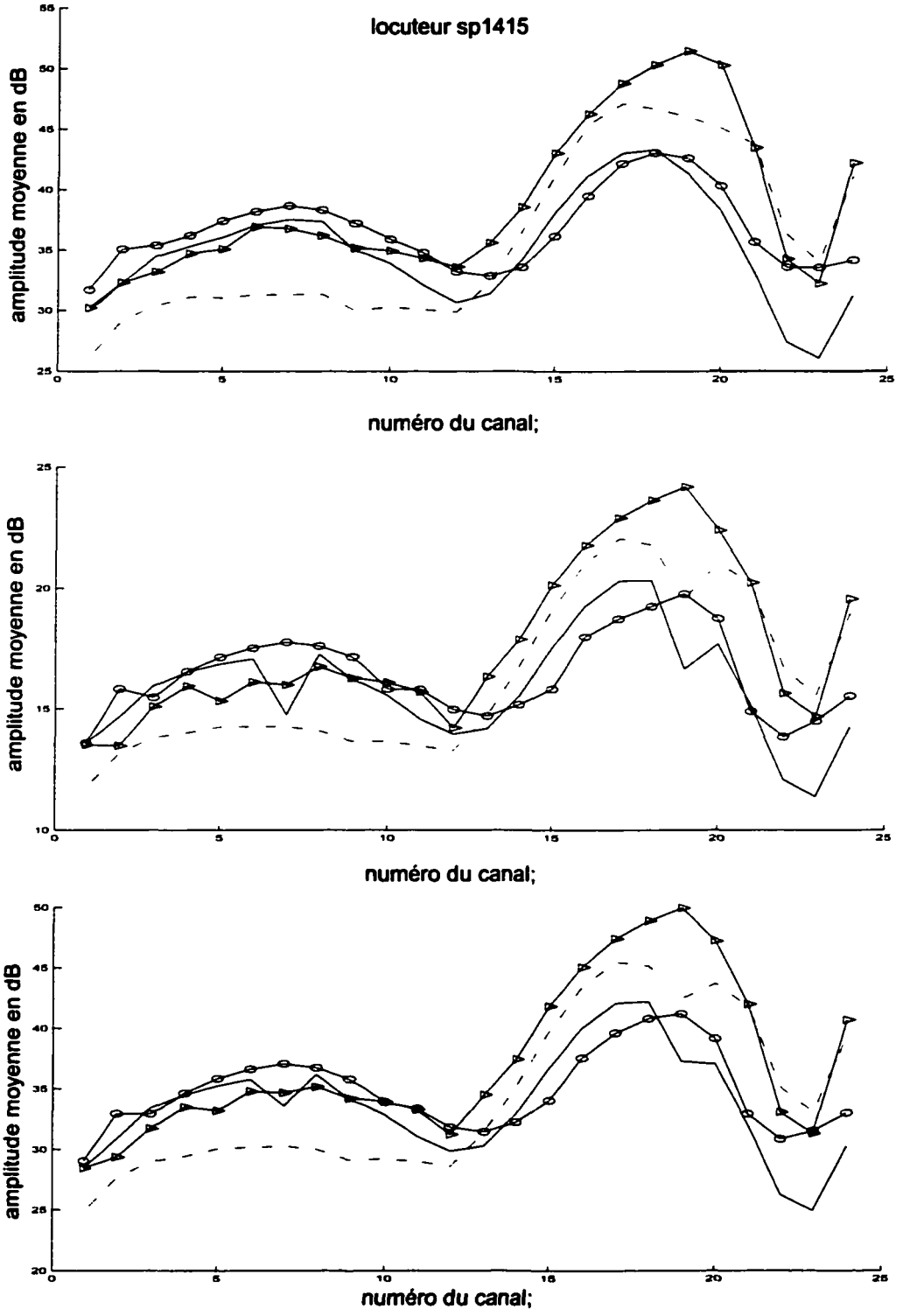

FIG. 2.4 - **a)** : Moyenne de la valeur absolue de la variation du carré des différences entre 2 unités élémentaires consécutives (equ. 2.5.5) ; **b)** : Moyenne de la valeur absolue des différences entre 2 unités élémentaires (equ. 2.5.3) ; **c)** : Moyenne de la valeur absolue du carré des différences entre 2 unités élémentaires (equ. 2.5.4) ; contexte utilisé est le "and".

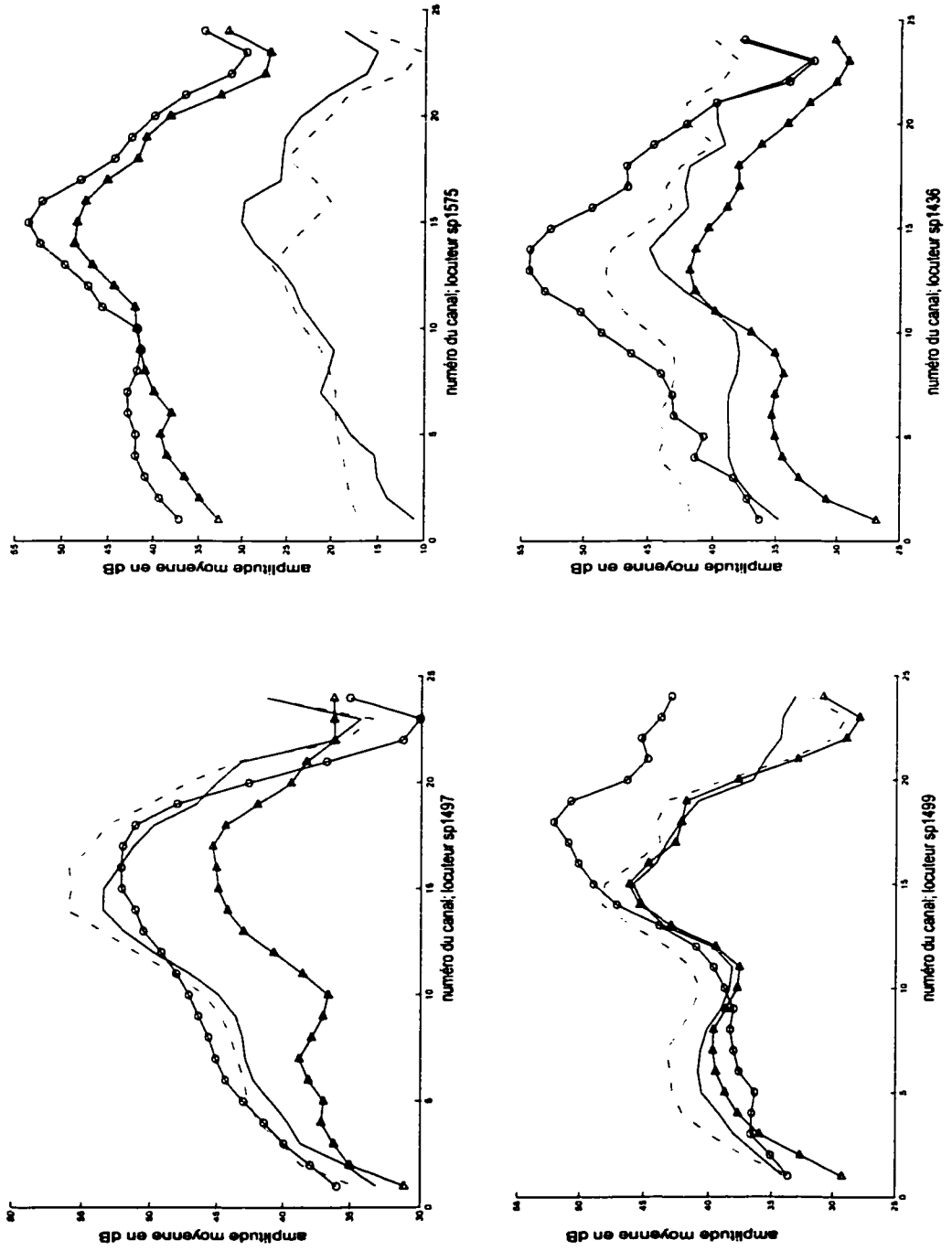

FIG. 2.5 - **Résultats pour 4 locuteurs de la moyenne de la valeur absolue de la variation du carré des différences entre 2 unités élémentaires consécutives**; contexte prononcé est le "and" (equa 2.5.5).

phonétique. Les données provenant du même combiné se caractérisent de plus par une coïncidence au niveau du gain.

Pour des locuteurs différents, on constate que les pics principaux du vecteur moyen ne se positionnent pas à la même fréquence et que parfois pour certains locuteurs la forme spatiale est significativement différente. Évidemment, ceci est dû à la distribution des formants et du contexte, sauf que dans notre cas, le contexte est fixé.

2.5.6 Conclusion

D'après l'étude préliminaire, ce type d'analyse synchrone à la glotte nous apparaît plus intéressant et de potentiel supérieur à ce que nous avons exploré jusqu'à présent (caractérisation par les lobes secondaires et la différence de phases). Il y a bien entendu lieu de pousser les analyses vers une meilleure compréhension des propriétés statistiques des paramètres liés à l'analyse spectro-temporelle des enveloppes. Le fait que les paramètres soient faiblement influencés par le combiné téléphonique tout en préservant le contexte phonétique nous laisse à penser que l'approche doit être explorée. La prochaine section exposera une étude plus détaillée de ces paramètres en utilisant l'ensemble de données de la base SPIDRE.

2.6 Propositions et expérimentations de nouveaux paramètres : MA-MF

Les études exploratoires des sections précédentes visaient à trouver une stratégie pertinente pour extraire de bons paramètres en caractérisation du locuteur. Les paramètres mentionnés sont dérivés de l'enveloppe ou de la FI à la sortie du banc de

filtres cochléaires. Il est maintenant temps d'opter pour un seul type de paramètre (donc une seule stratégie) et pousser son étude de façon plus approfondie. On a donc orienté notre étude vers l'extraction des paramètres synchrones à la glotte à la sortie de chaque canal. On a comparé les performances de ces paramètres aux coefficients standard MFCC. On a aussi augmenté la base de données à 45 locuteurs avec 4 conversations pour chaque locuteur provenant de 3 combinés différents. Les données font partie du Corpus SPIDRE. On invite le lecteur à consulter l'article d'Ezzaidi et al. [36] où plus de détails sont reportés (placé en annexe).

2.6.1 Paramétrisation

L'architecture du système proposé est illustrée à la figure 2.6. La technique d'extraction des paramètres consiste à utiliser au départ un algorithme de détection de la hauteur tonale pour estimer la période fondamentale T_0 des segments voisés. La valeur estimée du fondamental nous permet de reconnaître approximativement la durée de l'explosion glottale à partir de chaque canal. Ensuite, en exploitant conjointement la valeur de la période et la position des pics maximums de l'enveloppe à l'intérieur de trois périodes consécutives, on est capable de déterminer avec plus de précision le début et la fin de l'explosion glottale pour chaque canal. Ainsi, on se trouve à éliminer implicitement les problèmes du retard entre les canaux (délai) liés à la réponse des filtres cochléaires.

Pour chaque intervalle $[0,T_0]$ qui correspond à la période fondamentale, les vecteurs paramètres FI, DFI, MA, DMAP, MAP1, MAP2 sont estimés tels qu'illustrés à la figure 2.7. Le vecteur paramètre FI , AM correspond respectivement à la valeur moyenne de la fréquence instantanée et la moyenne de l'enveloppe à la sortie de

chaque canal et pour chaque impulsion glottale. Le vecteur DMA est la différence des vecteurs MA adjacents, et il mesure en quelque sorte la dynamique des vecteurs MA. Le vecteur DFI est la différence des vecteurs FI, cette fois ci, à travers les canaux adjacents. Ces vecteurs différences sont supposés réduire les effets du canal et la variabilité du combiné. Les vecteurs MAP1 et MAP2 sont censés caractériser respectivement les lobes secondaires (régime transitoire) et la montée de l'enveloppe (régime entretenu). En effet, il y aura lieu de mieux caractériser le vecteur MAP1 par son énergie propre que de l'estimer par l'énergie ayant lieu par défaut sur le quart de la période l'intervalle $[T_0/2,3T_0/4]$. Une telle modélisation se trouve très difficile à réaliser et peut donc biaiser les résultats. Pour plus les détails de mise en oeuvre du système, j'invite le lecteur à consulter l'article d'Ezzaidi et al [36].

2.6.2 Résultats

Les expériences dans cette section sont divisées en deux parties. Dans la première partie, on trouve les résultats cherchant à évaluer l'impact sur l'utilisation des combinés similaires ou différents entre les sessions d'apprentissage et de test selon deux aspects : un premier aspect où on suggère l'évaluation en se basant sur les paramètres proposés ; le deuxième aspect, propose une étude au niveau psycho-acoustique en comparant la performance de la machine et celle des auditeurs utilisant les combinés dans différentes conditions entre les sessions d'apprentissage et de test. La deuxième partie des expériences portera sur la comparaison des paramètres proposés et les coefficients MFCC pris comme paramètres de référence.

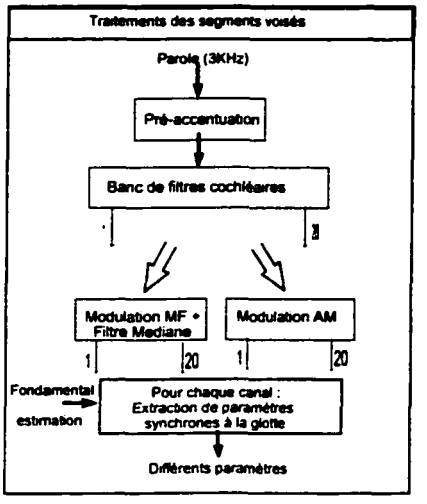

FIG. 2.6 – Architecture du système d'extraction des paramètres MA-FI.

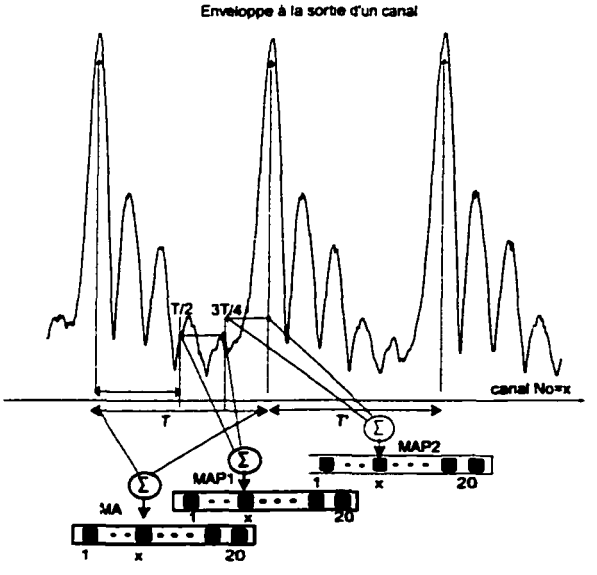

FIG. 2.7 – Illustration d'extraction de paramètres à la sortie d'un canal.

Résultats sur l'impact des combinés téléphoniques

Dans la base de données SPIDRE, on dispose seulement de deux conversations par locuteur enregistrées sur le même combiné téléphonique. On a donc utilisé une conversation pour la session d'entraînement et l'autre conversation pour la session du test afin d'évaluer les performances en reconnaissance du locuteur dans des conditions impliquant des combinés similaires. Dans la table 2.1, on donne les résultats de cette expérience.

Dans les conditions de la variabilité des combinés, on a utilisé trois conversations enregistrées à partir de 2 combinés différents lors de la session d'entraînement et la quatrième conversation enregistrée à partir d'un autre combiné a été présentée pour les tests. Dans la table 2.2, on donne les résultats qui tiennent compte de la variabilité des combinés téléphoniques.

TAB. 2.1 - Résultats avec les mêmes combinés (LVQ), 35 locuteurs (21 hommes et 14 femmes). Taille du dictionnaire est de 256.

Paramètres	IF	DIF	AM	DAM	$AMP1$	$AMP2$
Hommes	81%	90%	75%	62%	57%	76%
Femmes	65%	65%	72%	58%	79%	72%
Hommes+Femmes	74%	74%	63%	58%	58%	69%

TAB. 2.2 - Tests avec des combinés différents (LVQ), 35 locuteurs (21 hommes et 14 femmes). Taille du dictionnaire est de 256.

Paramètres	IF	DIF	AM	DAM	$AMP1$	$AMP2$
Hommes	71%	67 %	57%	33%	52%	71%
Femmes	79%	64%	92%	50%	79%	65%
Hommes+Femmes	69%	58%	52%	32%	32%	38%

Discussion

En comparant les tables 2.1 et 2.2, on peut déduire plusieurs points. Le point le plus important est que la variabilité des combinés a fait baisser les taux de reconnaissance sur l'ensemble des paramètres proposés à quelques exceptions pour le cas des femmes. Il est donc évident que la fonction de transfert du combiné, modifie considérablement les caractéristiques du signal de la parole pour tromper le système de reconnaissance. Sur l'ensemble des locuteurs, on observe avec les paramètres IF, DIF et AM une chute des scores qui dans la mesure reste acceptable contrairement au cas des paramètres DAM, $AMP1$ et $AMP2$. Si on veut utiliser ces derniers, il y aura lieu d'effectuer une pré-classification sur la base des sexes puisque d'après les tables les scores pour les hommes et les femmes restent en quelques sortes proches.

Pour le test impliquant les femmes, on remarque que les paramètres IF, DIF, MA, $MAP1$ donnent des scores supérieurs ou égaux comparativement aux conditions utilisant des combinés similaires alors que pour les autres paramètres la dégradation reste toutefois acceptable. Comme explication, on peut dire que la durée de la réponse en régime transitoire (liée au conduit vocal), est plus grande pour l'homme que la femme. Par conséquent, on peut dire que la longue durée de la période glottale chez les hommes comparativement aux femmes donne naissance à un genre de bruit pendant le régime transitoire (réponse lié au conduit vocal va se stabiliser avant l'arrivée de la prochaine excitation glottale). En se fiant toujours à la réponse de la glotte, on peut attribuer ces résultats aux faits qu'on dispose de plus de données pour la femme que l'homme puisque la durée de la période de l'impulsion glottale chez les femmes est plus courte que chez les hommes.

En résumé, les paramètres IF, DIF et AM semblent présenter plus de robustesse

à la variabilité des combinés que le reste des paramètres proposés.

Comparaison avec un système de référence

Puisque notre objectif est de trouver de nouveaux paramètres robustes spécialement aux conditions téléphoniques, nous étions donc contraints à effectuer une comparaison en performance avec un système de référence. On a privilégié l'utilisation les coefficients standards $MFCC$ comme paramètres de référence puisqu'ils représentent un standard dans le domaine. La tache de classification est restée commune pour les deux systèmes. Pour plus de détails, on invite le lecteur à consulter l'article [36]. On rappelle que seulement les conditions impliquant l'utilisation des combinés différents entre les sessions d'entraînement et de test ont été utilisées. De plus la base de données est composée cette fois-ci de l'ensemble des locuteurs disponibles dans la base de données SPIDRE.

Les résultats sont illustrés par les figures 2.9 et 2.8. Sur un aspect global, les coefficients MFCC enregistrent les meilleurs scores en reconnaissance du locuteur avec 70% pour les hommes, 78% pour les femmes et 73% pour l'ensemble des locuteurs. Alors que les scores respectivement pour la FI (et MA) sont de 63% (68%) pour les hommes, 72% (78%) pour les femmes et 58% (71%) pour l'ensemble des locuteurs. Cependant, les scores sur l'ensemble des paramètres restent proches et comparables bien que les paramètres de référence sont amplement adaptés et mieux optimisés pour ce type de traitement. Quand aux paramètres qu'on présente, leurs composantes ne sont pas complètement décorrélées et peut être seront mieux modélisées avec d'autres méthodes de classification. On trouve également que les femmes semblent mieux discriminées que les hommes. En réalité, les sons voisés des femmes comportent moins de composantes harmoniques que les sons voisés des hommes. Par ailleurs, les filtres

cochléaires sont supposés être capables de mieux démoduler des signaux de femmes à la sortie de chaque canal.

L'essentiel est que les paramètres AM-FM et les MFCC semblent ne pas confondre les mêmes locuteurs, un point qui nous mène à supposer que ces paramètres peuvent être complémentaires entre eux pour concevoir un système d'identification du locuteur plus robuste.

Ceci nous a motivé à simuler un système global d'identification du locuteur composé sous une architecture parallèle par des sous systèmes élémentaires entraînés individuellement sur un seul de type de paramètres MFCC, IF ou AM. Le système global donne en sortie une décision provenant de chaque système élémentaire. La fusion des décisions se fait directement à travers les matrices de confusion. Ceci nous permet d'éviter de faire une homogénéisation des scores, c'est à dire d'essayer de normaliser les scores des différents systèmes élémentaires qui ne sont pas forcément comparables entre eux. Ainsi, la pondération des scores peut aussi constituer un autre problème à résoudre. Dans la figure 2.8, on reporte les résultats sur l'ensemble des locuteurs pour les trois systèmes globaux MFFC+IF, MFCC+AM et MFCC+AM+FM. Particulièrement, on trouve que le système global basé sur la combinaison des paramètres MFCC et AM augmente le taux de reconnaissance à 82%, ce qui est considérablement supérieur aux performances des paramètres de référence avec 73% (coefficients MFCC).

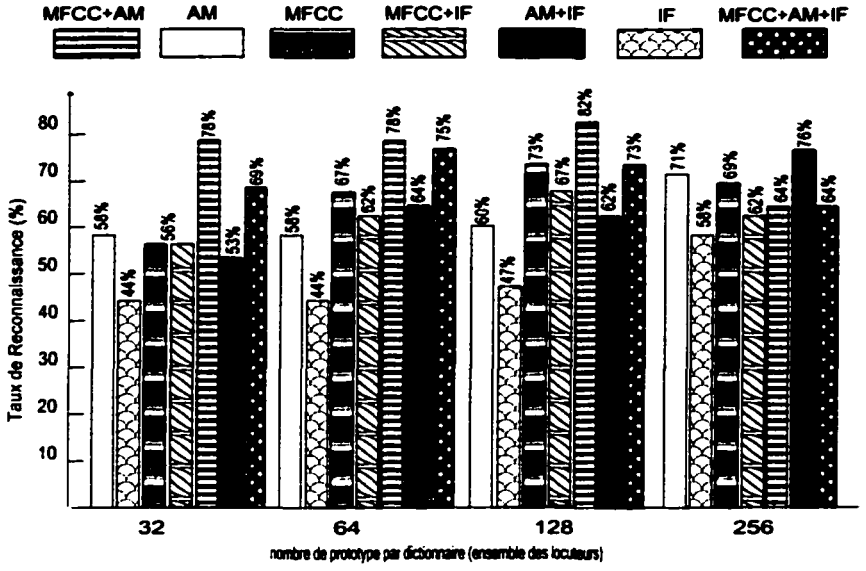

FIG. 2.8 - Taux de reconnaissance pour les paramètres AM, IF and $MFCC$, 45 locuteurs.

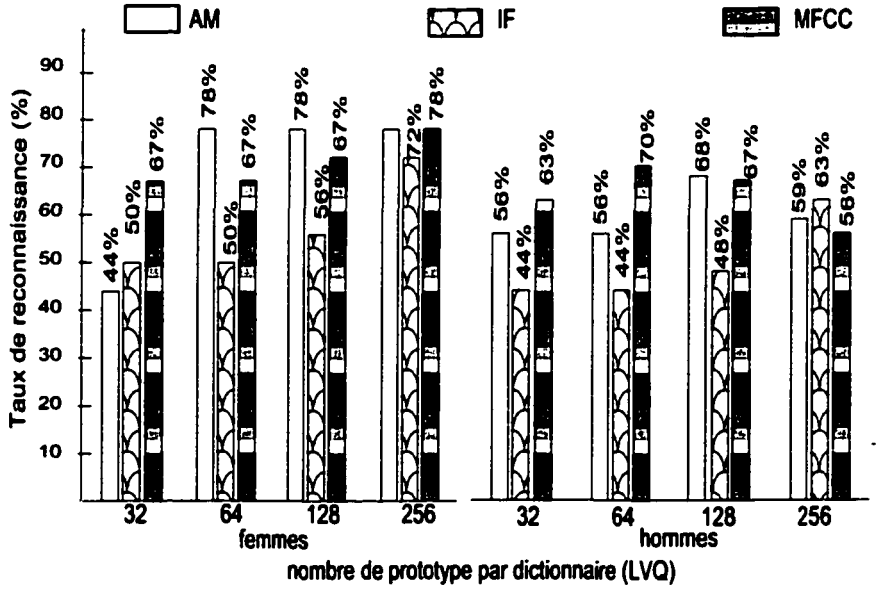

FIG. 2.9 - Taux de reconnaissance pour les paramètres AM, IF and $MFCC$ par sexe; les mêmes 45 locuteurs que la Fig. 2.8.

2.7 Résultats sur la comparaison entre auditeurs naïfs et de la machine en RAL

Pour mieux comprendre l'impact des combinés, on a cherché à évaluer les aptitudes de l'humain et la capacité d'une machine à reconnaître les locuteurs dans des conditions impliquant la variabilité des combinés téléphoniques.

La comparaison entre humain et machine en identification du locuteur a été traité par Doddington [29] dans les années 80. Cependant, son étude ne tient pas compte de l'implication de la variabilité des combinés téléphoniques. Dans cette section, on tentera de répondre partiellement à cette question même si l'étude qu'on propose reste un peu biaisée dans la mesure où les conditions d'expériences proposées ont avantage à favoriser plus l'humain que la machine. Dans la suite, on donnera un résumé sur les conditions expérimentales, les résultats et la discussion. Pour plus de détails, on invite le lecteur à lire l'article Ezzaidi et Rouat [35] joint en annexe.

2.7.1 Conditions expérimentales

Auditeurs naïfs

On a réalisé une étude d'écoute des fichiers de parole pour les 10 femmes les plus confondues entre elles de la base SPIDRE. Il n'est donc pas possible pour cette expérience de séparer les locuteurs sur la base de leur sexe. On a fait appel à 10 auditeurs (1 femme et 9 hommes) naïfs auxquels on a présenté des paires de segments de parole, sélectionnées de façon aléatoire. Les auditeurs parlent le français comme langue maternelle et la plupart ne comprennent pas bien l'anglais. Pour chaque paire écoutée, l'auditeur doit faire un des choix suivant :

- a : Certainement le même locuteur ;

- b : Probablement le même locuteur ;
- c : Certainement des locuteurs différents ;
- d : Probablement des locuteurs différents.

Paramétrisation et modèles pour la machine

Le vecteur de paramètres utilisé est composé de

- 12 coefficients MFCC (statiques) ;
- 12 delta coefficients MFCC (dynamiques) ;
- 1 coefficient pour l'énergie ;
- 1 coefficient pour delta-énergie (différence).

Une fenêtre d'analyse de 32 ms glissante à tous les 10 ms a été considérée pour calculer le vecteur des paramètres. Deux techniques de classification ont été proposées. La première est la classification non paramétrique qui consiste en une quantification vectorielle hybride (LVQ-SLP) telle que proposée par He et al [52]. La deuxième classification est une modélisation paramétrique utilisant un modèle à mélange de gaussiennes (GMM).

2.7.2 Résultats

Pour les auditeurs naïfs la table 2.3 illustre dans les colonnes 2 et 3 les scores mettant en évidence la variabilité intralocuteur et dans la colonne 4 les scores correspondant à la variabilité interlocuteur. Les auditeurs sont reportés dans la première colonne.

En utilisant le même combiné téléphonique et pour le même locuteur, on obtient un score moyen de 81% avec une variance de l'ordre de 16%. Cette fluctuation est due principalement aux 3 auditeurs L1 et L6. Si on ne prend pas ces 2 auditeurs dans

TAB. 2.3 - Les scores moyens pour les 10 auditeurs. La dernière colonne correspond à une paire de conversations impliquant deux locuteurs différents, utilisant des combinés inconnus. Similaire : même combiné et même locuteur, différent : combiné différent et même locuteur.

Auditeurs	même locuteur dans le test		Different
	(similaire)	(différents)	locuteurs
L1	60%	68%	67%
L2	90%	74%	77%
L3	90%	72%	75%
L4	70%	72%	72%
L5	90%	72%	75%
L6	50%	64%	65%
L7	90%	71%	92%
L8	83%	62%	75%
L9	100%	72%	62%
L10	90%	72%	75%
μ	81.3%	73.5%	69.9%
σ	15.95	3.95	8.24

TAB. 2.4 Scores de reconnaissance obtenus par la machine

Combinés	Taille des dictionnaires (LVQ-SLP)			32
conditions	512	256	128	GMM
Similaires	90%	90%	90%	-
Différents	60%	60%	60%	90%

statistiques, le score moyen pourrait atteindre un pourcentage de 90%.

Avec la variabilité des combinés, les scores chutent en moyenne de 11% avec une faible variance. Les tests impliquant la variabilité interlocuteur présentés dans la colonne 4, donnent des scores compatibles avec les résultats de la colonne 3 (variabilité des combinés pour le même locuteur).

On peut déduire que l'influence du combiné a tendance à dominer l'empreinte vocale du locuteur lors des tests d'écoutes présentés aux auditeurs naïfs. Par conséquent, les caractéristiques acoustiques du locuteur se retrouvent largement dégradées par

l'effet du combiné.

Les résultats pour la machine sont illustrés dans la table 2.4. Dans le cas de la classification non paramétrique et pour toutes les tailles de dictionnaires testés, un score de 90% a été obtenu pour les conditions impliquant des combinés similaires. Avec des combinés différents les scores chutent à 60%, toutefois avec les modèles on obtient le meilleur score (90%) et devance dans ce cas l'aptitude humaine et dépasse significativement les performances humaines.

2.7.3 Discussion

Dans tous les cas, l'utilisation des combinés entre les sessions d'entraînement et de test génère une dégradation en performance qui est la plus faible pour le cas de la machine. Le succès de la performance de la machine sur l'homme est peut-être attribué au modèle, à la qualité des paramètres et au prétraitement effectué (préaccentuation, normalisation, liftering,...). Bien que cette expérience semble simple avec des données restreintes, néanmoins la performance des auditeurs naïfs nous donne une idée sur l'impact des combinés et la complexité de la base de données SPIDRE.

2.8 Conclusion

Dans ce chapitre, on s'est intéressé à examiner principalement différentes techniques en paramétrisation en vue d'améliorer les performances des systèmes actuels en RAL. Le système de référence retenu pour nos expérience se base sur les coefficients MFCC. La base de données utilisée est le Corpus SPIDRE. Les paramètres proposés sont dérivés de l'enveloppe et de la fréquence instantanée, estimés à la sortie d'un banc de filtres cochléaires. Plusieurs études préliminaires ont été réalisées

dans l'unique but de cibler une technique en paramétrisation robuste aux conditions téléphoniques et impliquant plusieurs combinés. On a exploré la caractérisation en se basant sur :

- le déphasage entre les canaux cochléaires,
- l'énergie des lobes secondaires,
- la différence des canaux adjacents (pour MA et FI),
- l'information à l'intérieur de chaque impulsion glottale (MA et FI).

L'exploitation et l'extraction des paramètres synchrones à la glotte nous paraît une des techniques qui peut être prometteuse et intéressante. Dans cette mesure, différents paramètres ont été proposés et analysés. Les performances obtenues sont très proches des performances du système de référence. Un gain significatif a été obtenu lorsque le système se base à la fois sur les coefficients MFCC et les paramètres d'enveloppe et/ou la fréquence instantanée. Une comparaison entre la perception humaine et la capacité d'une machine en caractérisation du locuteur a été également effectuée, notamment pour examiner l'impact des combinés téléphoniques. Il s'est avéré que l'impact du combiné affecte grandement la perception humaine comparativement à la capacité de la machine.

Pour achever ce chapitre, on recommande d'envisager ultérieurement d'autres modifications aux paramètres proposés qui se rapportent principalement à :

- rechercher une transformation mathématique pour mieux décorréler les composantes des vecteurs de paramètres proposés.
- envisager une procédure de sélection de canaux non perturbés par le bruit.
- pondérer les composantes en fonction de l'énergie du canal.

- proposer un modèle hybride pour une combinaison adéquate et une fusion optimale de différents paramètres.

Chapitre 3

Modélisation de la source et du conduit vocal par loi de probabilité conjointe

Dans le précédent chapitre, on s'est intéressé particulièrement à analyser et à comparer l'utilisation de nouveaux paramètres dérivés de l'enveloppe et de la fréquence instantanée. On s'intéressait à l'identification du locuteur en mode téléphonique, en exploitant indirectement l'information de la source. Donc, dans l'architecture globale du système d'identification du locuteur, on était situé exactement au niveau du module de paramétrisation. Dans ce chapitre, on essaye d'exploiter l'information de la source mais cette fois-ci on se place au niveau du module de modélisation. Précisément, on introduit un nouveau modèle probabiliste qui prend en compte la dépendance entre la source et le conduit vocal.

3.1 Introduction

Les mesures statistiques de la fréquence fondamentale Fo à long terme, ont été utilisées avec succès dans les systèmes d'identification du locuteur comme paramètres

prosodiques. Elles ont démontré une meilleure robustesse aux bruits et aux distorsions introduites par les canaux de transmission. La valeur moyenne de Fo a été considérée dans les travaux de Sönmez et al [105] comme une mesure descriptive au niveau perceptive. Sönmez et al [104] ont démontré mathématiquement que pour que la distribution du fondamental suive une forme normale, il faut prendre le logarithme naturel du fondamental $log(fréquence)$.

En effet, la majorité des systèmes de RAL utilisant la prosodie supposent une indépendance à court terme entre le conduit vocal et la source. Cependant, le rehaussement du taux de reconnaissance obtenu par ces systèmes reste très faible si on les compare aux systèmes qui exploitent uniquement les coefficients caractéristiques de conduit vocal. Bien que la contribution des paramètres prosodiques s'avère fragile dans les systèmes de RAL, il reste que le mécanisme de production est assez complexe et implique une nette dépendance de l'articulation, de la vibration des cordes vocales et du conduit vocal. L'insuccès à l'exploitation des paramètres prosodiques peut être dû au fait que les techniques ou les modèles utilisés ne collent pas exactement aux distributions statistiques de ces paramètres.

Dans cette perspective, on propose une nouvelle approche qui tient compte de la dépendance entre la glotte et le conduit vocal. On étudie l'influence de cette dernière dans le contexte de l'identification du locuteur indépendant du texte. Particulièrement, on suppose que le conduit vocal et la source sont deux processus aléatoires X et Y régis par une loi de probabilité conjointe $f(X,Y)$.

La motivation, le formalisme mathématique, la paramétrisation, les modèles et les critères de reconnaissance utilisés pour cette approche sont décrits en détails dans les articles [33] ou [34]. Toutefois, dans ce chapitre on fera un bref rappel sur la motivation

et le formalisme mathématique. Ensuite on présentera en détail les résultats obtenus avec les discussions et les conclusions appropriées.

3.2 Motivation

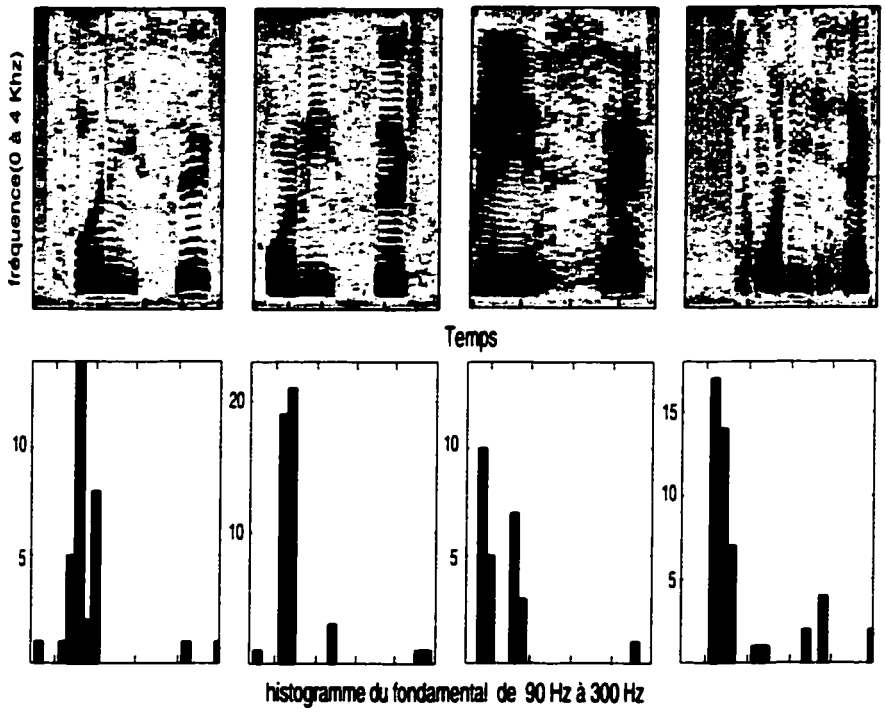

FIG. 3.1 - Spectrogrammes et histogrammes pour 4 locuteurs de sexe masculin

L'importance de la dépendance entre la source et le conduit vocal a été maintenue et analysée dans un premier temps sur la base d'une série d'analyses graphiques. La figure 3.1 illustre quatre spectres et quatre histogrammes de la fréquence glottale. Chaque colonne de la figure correspond à un locuteur différent de sexe masculin. Les locuteurs font partie de la base de données YOHO. Tous les locuteurs ont prononcé le nombre "26" à travers un support téléphonique. Les spectres des quatre locuteurs

montrent une certaine similarité de la distribution des formants. En effet, cette distribution spatiale des formants représente la variabilité interlocuteur telle qu'elle a été discutée dans les travaux de Doddington [29]. Cependant, les histogrammes de la fréquence de la glotte sont différents et varient d'un locuteur à un autre et ceci pour le même contexte prononcé. Si on fait une comparaison en prenant les amplitudes comme critère, on remarque que l'histogramme de la fréquence de la glotte du locuteur en colonne deux et celui en colonne quatre présentent une forte similarité. Cependant, les histogrammes du premier locuteur et du troisième locuteur sont plutôt caractérisés par des histogrammes complètement différents. Par conséquent, si on tient compte de l'information du fondamental, la variabilité inter-locuteur se trouve réduite aux locuteurs ayant une grande similarité de la vibration de leurs cordes vocales. Alors que les deux autres locuteurs peuvent être classés directement comme étant des locuteurs différents. Pour les locuteurs à distribution du fondamental similaire, on peut poursuivre la classification sur un deuxième niveau en se basant sur leurs caractéristiques spectrales.

En bref, l'information du fondamental et les caractéristiques spectrales du conduit vocal, peuvent être exploités conjointement dans le but de développer des systèmes d'identification du locuteur plus robustes.

3.3 Formalisme théorique

Sans reprendre les détails du formalisme théorique exposé dans l'article [33], on rappelle simplement que la source et le conduit vocal ont été supposés être caractérisés respectivement par deux variables aléatoires $X(t)$ et $Y(t)$. Dans le cas discret, ces variables aléatoires peuvent être approximées par $\widehat{X(n)}$ et $\widehat{Y(n)}$ à l'instant $n\Delta t$.

La variable aléatoire $\widehat{X(n)}$ peut être considérée comme une application qui associe à chaque état de la source (épreuve) une valeur entière de la fréquence de la glotte, appartenant au domaine $D = [63\ Hz, 600\ Hz]$ avec $D \in N$. La variable aléatoire $\widehat{Y(n)}$ est une application qui associe à chaque configuration du conduit vocal (épreuve) un vecteur $\overrightarrow{y_j}$ de dimension l. Ce vecteur $\overrightarrow{y_j}$ peut être estimé par n'importe laquelle des méthodes qui caractérisent les propriétés du conduit vocal telles que les méthodes LPC et $MFCC$.

Pour prendre en compte la dépendance entre la source et le conduit vocal, on a supposé que les deux variables aléatoires sont régies par une fonction de probabilité conjointe :

$$f(x_i, \overrightarrow{y_j}) = P(\widehat{X} = x_i, \widehat{Y} = \overrightarrow{y_j}) \text{ avec} \tag{3.3.1}$$

$$0 \leq f(x_i, \overrightarrow{y_j}) \leq 1 \text{ et } \sum_{i=1}^{n}\sum_{j=1}^{m} f(x_i, \overrightarrow{y_j}) = 1. \tag{3.3.2}$$

Les fonctions de probabilités marginales respectives sont données par les équations suivantes :

$$f(x_i) = \sum_{j=1}^{m} f(x_i, \overrightarrow{y_j}) \text{ et } f(y_j) = \sum_{i=1}^{n} f(x_i, \overrightarrow{y_j}). \tag{3.3.3}$$

Chaque locuteur s est supposé défini par sa propre fonction de probabilité f_s qui tient compte du couplage entre la source $(\widehat{X})$ et le conduit vocal $(\widehat{Y})$:

$$f_s(x_i, \overrightarrow{y_j}) = P_s(\widehat{X} = x_i, \widehat{Y} = \overrightarrow{y_j}). \tag{3.3.4}$$

D'après la relation de Bayes, on peut déduire la probabilité conditionnelle :

$$f_s(x_i, \overrightarrow{y_j}) \ = \ f_s(\overrightarrow{y_j}/x_i) f_s(x_i). \tag{3.3.5}$$

$f_s(x_i)$ est la probabilité à priori d'observer la valeur x_i et $f_s(\overrightarrow{y_j})$ est la probabilité à posteriori d'observer $\overrightarrow{y_j}/x_i$ sachant que l'événement x_i a été réalisé.

En théorie, pour chaque x_i est associé un nombre fini de vecteurs $\overrightarrow{y_j}$. Ce qui veut dire, qu'on devra entraîner un modèle de classification pour chaque valeur de x_i. En pratique, ce concept est très lourd et surtout il ne présente aucun intérêt pour les applications commerciales. On a donc passé par certaines restrictions afin de faciliter le traitement. On a procédé par fragmentation de l'espace $(x,\overrightarrow{y_j})$ en sous-espaces adjacents notés H_k. Chaque sous-espace est caractérisé par un sous-intervalle de fréquence I_k avec $k = 1, \ldots, N$. Les N sous-intervalles sont répartis sur la plage de la fréquence de la glotte allant de $x_1 = 60\ Hz$ à $x_n = 600\ Hz$. La probabilité conditionnelle $f_s(\overrightarrow{y_j}/x_i)$ est supposée localement indépendante de la valeur du fondamental à l'intérieur de chaque sous-espace. On peut donc écrire :

$$f_s(\overrightarrow{y_j}/x_i) = P(\widehat{Y} = \overrightarrow{y_j}/I_k,\ \text{locuteur } = s \text{ avec } x_i \in I_k) \tag{3.3.6}$$

En subdivisant l'espace, on réduit le nombre de modèles théoriques à N/n . La figure 3.2 illustre la notion des sous-espaces et leurs modèles pour les N densités de probabilités $f_s(\overrightarrow{y_j}/x_i)$. La largeur des intervalles I_k a été déterminée en se basant sur la forme de l'histogramme du fondamental.

3.3.1 Méthodologie proposée

Pour nos expériences, on a eu recours à trois systèmes différents pour des fins de comparaison. Le premier système suggéré correspond au système de référence qui en principe extrait les coefficients caractéristiques à partir des segments voisés et non-voisés. Le deuxième système ne traite que les segments voisés en ignorant

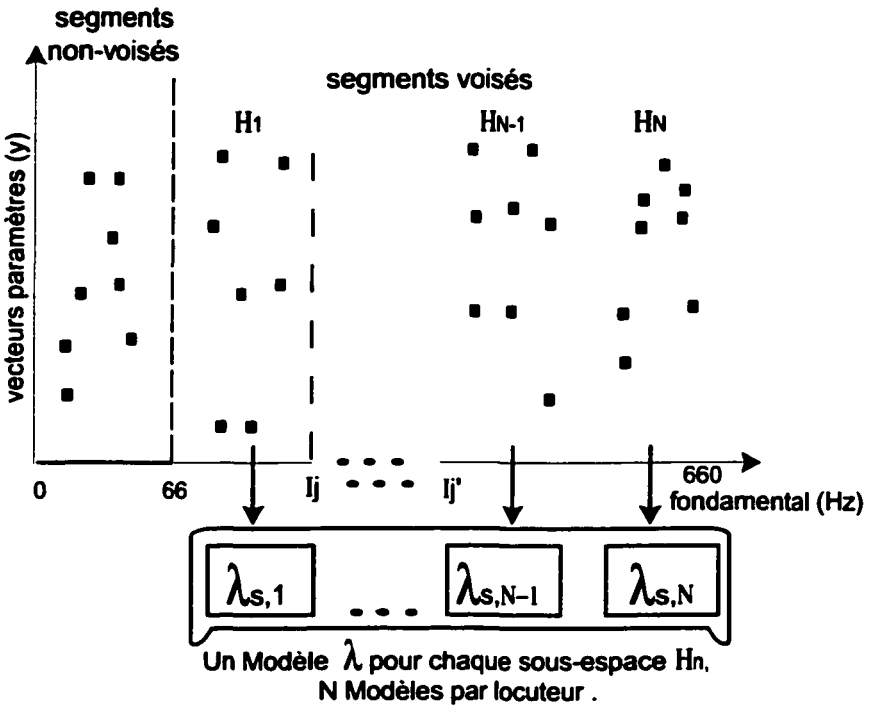

FIG. 3.2 - Illustration de la fragmentation de l'espace des paramètres et l'accroissement du nombre de modèles par locuteur.

tous les segments de parole non-voisés pendant le traitement. Pour chacun des deux systèmes précédents, un seul modèle paramétrique a été utilisé pendant la phase d'entraînement.

Le troisième système est le système proposé qu'on désignera par la suite sous le nom : de multi_modèles. En principe, on commence par estimer le fondamental à toutes les 10 ms sur une fenêtre d'analyse de 32 ms dont l'objectif est de sélectionner les segments voisés. Ensuite, on extrait les coefficients caractéristiques du conduit à partir des segments voisés, et on les distribue selon l'appartenance de leurs fréquences de glotte aux sous-intervalles I_k, prédéfinis à priori, en N sous-classes H_k. Les données de chaque sous-classe, serviront à entraîner un modèle représentant cette sous-classe (ou sous-intervalle). Donc avec ce système multi_modèles, on se retrouve avec autant de modèles que de sous-classes.

Dans cette expérience, tous les modèles proposés pour caractériser la contribution du conduit vocal sont composés d'un mélange de 32 Gaussiennes pondérées. Les paramètres utilisés sont les 12 coefficients statiques des MFCC. Le nombre de modèles (ou intervalles de fréquence) pour le troisième système a été fixé à quatre. Un tel nombre a été considéré suffisant pour alléger le déroulement des tests.

Les intervalles de fréquence en Hz prédéfinis dans nos expériences sont les suivants :

- Pour les femmes :
  - $I_1 = [150, 180]$;
  - $I_2 = [170, 200]$;
  - $I_3 = [190, 220]$;
  - $I_4 = [63, 150] \cup [220, 600]$.
- Pour les hommes :

- $I_1 = [90, 120]$;
- $I_2 = [110, 130]$;
- $I_3 = [120, 150]$;
- $I_4 = [63, 90] \cup [150, 600]$.

Le chevauchement entre intervalles adjacents a été introduit volontairement pour pallier aux erreurs qui peuvent être introduites pendant l'estimation du fondamental et/ou par l'utilisation des supports de communication.

3.3.2 Résultats

Dans cette première partie, on a supposé que la probabilité à priori est localement uniforme sur chaque intervalle de fréquence I_k, et l'équation 3.3.5 reprendra une forme plus simple :

$$f_s(x_i, \overrightarrow{y_j}) \;=\; f_s(\overrightarrow{y_j}/x_i). \tag{3.3.7}$$

Par conséquent, le fondamental ne sert plus qu'à sélectionner un modèle précis des N modèles proposés pour estimer le score en reconnaissance par la méthode du maximum de vraisemblance. Tel qu'illustré par l'équation 3.3.7, le fondamental ne contribue pas entièrement dans l'estimation des scores.

Dans la table 3.1, on reporte les résultats observés pour les trois systèmes proposés, sur la base de dix-huit (18) femmes de la base de données SPIDRE. Dans ces expériences seules les conditions avec la variabilité des combinés sont prises en considération. La première colonne indique la durée du test en secondes.

En comparaison, on a obtenu des taux de reconnaissance plus bas pour toutes les durées testées avec le système de référence. Le système basé sur les sons voisés

a enregistré un gain allant de 1% à 4.5%. Cependant, le système proposé donne les meilleurs scores avec un gain allant de 2.5% à 6%.

Comme explication, on peut possiblement attribuer la perte en performance du système de référence à la présence d'un nombre réduit des segments voisés sur une durée de test fixée.

Pour les grandes durées de tests, les gains en performance de façon globale ont tendance à converger. En conséquence, le système proposé peut être jugé intéressant seulement pour les applications qui nécessitent une durée de test très petite.

Cependant, le système multi_modèles ne peut pas représenter pour l'instant un succès, si on tient compte des facteurs additionnels suivants en comparaison au système de référence : les opérations de calculs additionnelles et le gain en performance apportés par rapport au système de référence.

Plusieurs restrictions et approximations ont été effectuées dans notre travail. Parmi d'autres, la probabilité à priori $f_s(x_i)$ a été considérée constante d'après l'équation 3.3.5 sous l'hypothèse qu'on travaille sur des données provenant des locuteurs de même sexe. Cette approximation, n'est pas toujours vraie et il faut envisager à reprendre en considération le densité de probabilité $f_s(x_i)$. Ceci constitue l'objet de la prochaine partie des résultats.

3.3.3 Modélisation de la source

Ici, on s'intéresse à supprimer l'une des restrictions imposées au système multi_modèles proposé précédemment. En effet, on remplacera la densité de probabilité $f_s(x_i)$ considérée

TAB. 3.1 - GMM : Taux de Reconnaissance pour les 18 femmes.

Temps(secondes)	Référence(%)	Voisé(%)	Multi_modèle(%)
0.1	36.8	37.7	40.5
0.5	63.4	66.7	69.4
1	75.4	79.9	80.8
2	84.2	87.9	88.0
3	88.0	90.6	90.5
4	90.0	93.9	93.3
5	91.4	95.4	94.7
6	92.7	95.3	95.2

localement uniforme dans les résultats précédents, par une densité de probabilité non-uniforme à l'intérieur de chaque intervalle I_k. La distribution statistique du fondamental est modélisée par un modèle paramétrique à mélange de Gaussiennes, dans la même perspective que les modèles pour les coefficients MFCC. La base de données a été augmentée à quarante-cinq (45) locuteurs de la base de données SPIDRE.

Modèles proposés pour la contribution de la source

On a utilisé un modèle paramétrique à mélange de quatre (4) Gaussiennes pondérées, pour chacun des quarante-cinq (45) locuteurs et ceci pour chacun des sous-intervalles de fréquence I_k. Après une série d'analyse sur les histogrammes du fondamental pour différents locuteurs, le choix de quatre (4) Gaussiennes par modèle s'avère à la limite suffisant.

La procédure d'entraînement des modèles pour le fondamental s'est faite de la même manière que celle décrite dans la section précédente pour les sous-modèles des coefficients MFCC mais selon deux approches. Dans la première approche, on s'est inspiré des travaux de Sönmez et al [105] pour appliquer une transformation logarithmique sur les fréquences estimées de la glotte. Ce système est désigné dans

la suite par logFo-multi_modèles. Dans la deuxième approche, on prend directement les fréquences sans apporter aucune modification et le système cette fois-ci portera le nom de Fo-multi_modèles. Les modèles entraînés par les deux approches ont été expérimentés. Dans les tests, on a utilisé les données de la 4 *ième* conversation avec un combiné différent de ceux utilisés dans l'entraînement.

Un autre système désigné par max-multi_modèles a été ajouté afin de vérifier si tout simplement l'alternative qui consiste à faire éclater un seul modèle en plusieurs modèles sans inclure la contribution du fondamental dans les scores, peut en soi apporter un gain et un intérêt comme justification pour le système proposé. En effet, ce système opère sans présélection des modèles par le fondamental et consiste à évaluer le score, à tour de rôle, pour tous les modèles d'un locuteur, pour enfin maintenir le score jugé maximal. En principe, on doit comparer le système max-multi_modèles au système voisé puisqu'ils utilisent les mêmes données des tests. En effet, le modèle du système voisé n'est qu'une généralité des modèles du système max-multi_modèles.

Les résultats sont illustrés par les tables 3.2, 3.3 et 3.4 respectivement pour l'ensemble des locuteurs, les femmes et les hommes. Dans toutes les tables, le signe "+" reporte un gain en performance (en %) par rapport aux performances du système de référence. Alors qu'avec le signe "-", on indique une perte en performance (en %) par rapport au performance du système de référence. En premières colonnes, on donne différentes unités de temps utilisées pour réaliser les tests. Dans la ligne référence, on donne les scores pour le système de référence. Dans la ligne étiquettée voisé on donne les performances en gain pour le système qui traite seulement les segments voisés.

On a gardé la même paramétrisation (12 MFCC statiques) et les mêmes modèles

(32-GMM) présentés dans la première partie des résultats pour modéliser les paramètres caractérisant la contribution du conduit vocal.

Comme leurs noms l'indiquent, les autres lignes donnent les résultats pour les systèmes désignés par *Fo-multi_modèles*, *logFo-multi_modèles* et *max-multi_modèles*. Pour chaque durée test, le locuteur ayant le modèle qui donne le score le plus élevé, est désigné comme étant le locuteur reconnu par le système. Un point à éclaircir est celui où le fondamental se retrouve à l'intérieur de deux intervalles de fréquence entrelacés. Dans ce cas, on teste les modèles correspondants aux deux intervalles en questions pour évaluer les scores et garder le maximum des deux.

TAB. 3.2 - Performance sur tous les locuteurs (hommes+femme)

temps de tests	100 ms	500 ms	1 s	2 s	3 s	4 s	5 s	6 s
référence	29%	58%	71%	81%	85%	88%	90%	91%
voisé	+1.48	+1.38	+1.72	+3.11	+3.17	+3.97	+3.90	+3.01
logFo-multi_modèles	+9.32	+6.17	+3.04	+2.05	+0.14	-0.66	-1.16	-1.58
Fo-multi_modèles	+9.30	+6.10	+3.06	+2.34	+0.28	-0.53	-0.88	-1.34
max-multi_modèles	+4.09	+5.85	+5.77	+4.81	+3.25	+2.33	+2.08	+0.82

TAB. 3.3 - Performance sur les femmes

temps de tests	100 ms	500 ms	1 s	2 s	3 s	4 s	5 s	6 s
référence	26%	56%	68%	79%	84%	87%	88%	90%
voisé	+0.91	+0.69	+1.84	+3.70	+3.23	+4.01	+4.90	+4.13
logFo-multi_modèles	+11.84	+9.42	+6.82	+4.74	+0.76	-0.04	+0.01	-0.22
Fo-multi_modèles	+11.87	+9.09	+6.63	+4.83	+0.76	-0.02	+0.26	-0.08
max-multi_modèles	+7.15	+7.03	+8.15	+6.21	+3.54	+2.21	+2.53	+1.68

D'après les résultats, on remarque que les systèmes "voisé" et "max-multi_modèle" enregistrent un gain en performance pour toutes les unités de temps testées. Ce gain diminue au fur et à mesure que la durée de test augmente. On peut en déduire

TAB. 3.4 – Performance sur les hommes

temps de tests	100 ms	500 ms	1 s	2 s	3 s	4 s	5 s	6 s
référence	31%	61%	73%	82%	86%	88%	90%	92%
voisé	+5.71	+4.58	+3.39	+3.61	+3.73	+4.28	+3.35	+2.20
logFo-multi_modèles	+10.12	+4.80	+0.55	+0.07	-0.34	-1.14	-2.06	-2.63
Fo-multi_modèles	+10.08	+4.87	+0.70	+0.49	-0.09	-0.92	-1.76	-2.31
max-multi_modèles	+6.64	+6.84	+4.90	+3.97	+3.02	+2.42	+1.74	+0.16

intuitivement que le système de référence également se base sur les segments voisés en caractérisation du locuteur. Avec une durée de test assez large, le système de référence se retrouve en possession d'un nombre important de segments voisés qui lui permette de se rapprocher des performances des autres systèmes. On peut dire qu'on a tendance à atteindre une certaine convergence.

Il est à noter aussi que les deux systèmes Fo-multi_model et logFo-multi_model enregistrent des scores similaires. Par conséquent, la modélisation après avoir effectué une transformation logarithmique n'apporte pas un intérêt particulier dans ce cas, au contraire elle augmente les opérations en calcul.

Par ailleurs, les performances du système "Fo-multi_model" par rapport au système de référence, sont importantes pour les petites durées de test, similaires pour les moyennes durées de test et minimes pour les longues durées de test. Pourtant, dans les résultats de l'article [33] avec le système proposé on avait obtenu des meilleurs performances que le système de "voisé" et le système de référence. La seule différence est qu'on avait pas pris en compte le facteur $f_s(x_i)$ ce qui met de nouveau en question l'homogénéité des scores. En effet, la probabilité à priori $f_s(x_i)$ est significativement plus élevée que la probabilité à posteriori $f_s(\frac{\vec{y_i}}{x_i})$. Cette différence de probabilité se trouve affecter faiblement les scores sur de courte durée alors qu'elle influence fortement les résultats pour de longue durée de test. L'origine de cette différence est

liée aux deux formes des modèles proposés pour la source et le conduit vocal qui se distinguent particulièrement par le nombre de Gaussiennes, la dimension et le type de paramètres. Il faut donc penser à une forme d'homogénéisation de ces deux types de paramètres, caractérisant la source et le conduit vocal. Il y aura lieu par exemple de normaliser chaque vecteur paramètre par une variance calculée sur tous les paramètres peu importe leurs types, ou les diviser par le produit de la variance de chaque type de paramètre ou d'essayer tout simplement de pondérer les deux facteurs $f_s(x_i)$ et $f_s(\frac{\vec{y_j}}{x_i})$. En ce qui concerne la différence dans les résultats entre les sexes, on peut l'associer à leurs dissimilitudes des propriétés spectrales.

Récemment, Arcienega et al [3] se sont intéressés à la même problématique tout en envisageant une démarche presque similaire à ce qu'on a proposé dans ce chapitre. Particulièrement, à chaque locuteur est associé trois modèles GMM. Le premier modèle est entraîné exclusivement à partir de données extraites des segments non-voisés. Le deuxième modèle est entraîné exclusivement à partir de données extraites des segments voisés. Le troisième modèle consiste à modéliser paramétriquement la distribution statistique du fondamental. Ce dernier modèle agit comme un facteur de pondération sur les scores évalués par le modèle des voisés (deuxième modèle). En utilisant la base de données $TIMIT$, les auteurs confirment avoir obtenu avec le système proposé un gain en performance de 11% plus élevé que le système de référence. Ce dernier correspond au même système de référence qu'on utilise. Cependant dans leur expérience, ils ont compilés les résultats seulement pour la durée de 3 ms et avec des données propres. Il n'est donc pas possible de faire une comparaison avec notre système qui ne tient pas compte de la contribution des segments non-voisés.

3.4 Conclusion

Une nouvelle méthode qui maintient la dépendance entre la source et le conduit vocal a été proposée. Les expériences qui calculent la probabilité à posteriori d'observer le vecteur de paramètres MFCC étant donné que la fréquence du fondamental correspondante est connue, ont été reportées et discutées. Elles ont été comparées au système du référence opérant sur les segments voisés et non voisés. Un deuxième système qui exploite seulement les segments voisés a été aussi considéré. Les expériences ont été réalisées avec les quarante-cinq (45) locuteurs de la base Spidre.

On a obtenu un gain en performance significatif quand la durée du test est inférieure à 500 ms. Lorsqu'on augmente la durée, les performances entre les différents systèmes ne demeurent pas en général assez importante.

En effet, plusieurs restrictions et hypothèses ont été prises pour la réalisation du système. On a supposé que le détecteur et l'estimateur du fondamental sont fiables. On a utilisé la même subdivision sur l'intervalle de fréquence pour générer les sous-modèles pour l'ensemble des locuteurs, alors qu'en réalité des locuteurs de même sexe peuvent avoir des histogrammes de fréquences différentes. On a supposé que tous les sous-modèles (modèle par classe) ont été bien entraînés alors qu'on disposait de moins de données dans certains cas.

L'estimation des paramètres des modèles probabilistes que ce soit pour la source ou le conduit vocal a été faite sans porter attention à tenir compte de la variance des paramètres sur un aspect global. Il s'est avéré qu'il fallait appliquer une normalisation sur la base des paramètres pour rendre les probabilités homogènes ou plutôt procéder par une pondération entre les décisions de la source et du conduit vocal. Une situation, qui peut nous expliquer pourquoi le système proposé n'excède pas le système de

référence significativement pour de longues durées de test.

On suggère pour les prochains travaux, de reprendre les restrictions proposées et de pousser leurs études plus en profondeur. Particulièrement, la normalisation des paramètres de la source et du conduit vocal risque d'influencer beaucoup les performances du système. Il y aura lieu de caractériser chaque locuteur par ses propres intervalles de fréquence au lieu de les fixer de façon absolue. Également, il faut tenir compte des segments non-voisés en modélisation, ainsi les scores seront évalués à chaque trame.

Chapitre 4

Discrimination Parole/Musique

4.1 Introduction

Le présent chapitre porte particulièrement sur la phase de discrimination Parole/Musique. L'effort sera donc focalisé ici, sur la faculté à mieux identifier les séquences de la parole, de la musique et de la musique chantée. On fera en sorte que le système soit robuste et indépendant de l'énergie du signal et des conditions d'environnement hostiles.

En partant de cette problématique, on proposera deux approches pour la discrimination de Parole/Musique, différentes en quelque sorte de ce qu'on trouve dans les publications scientifiques.

Comme première approche, on présentera une technique qui ne nécessite aucun apprentissage et qui repose simplement sur des manipulations d'ordre statistique des vecteurs paramètres. La simplicité de ce traitement repose sur l'utilisation d'un seuil comme critère pour la tâche de classification et sur sa rapidité en temps de calcul. Le seuil peut être facilement adapté à d'autres conditions d'utilisation et ajusté selon les besoins et perspectives de chaque application.

Dans la deuxième approche, on propose une modélisation paramétrique à mixture de *Gaussiennes*. Classiquement, on entraîne deux modèles respectivement pour la parole et la musique. Dans notre cas, on introduit un troisième modèle dédié à la musique chantée. Une telle approche, nous a permis d'obtenir de meilleures performances comparativement à l'approche classique. L'article en annexe de H. Ezzaidi et J. Rouat [32], décrit sommairement le principe et les expériences portant sur ces deux approches.

4.2 Bibliographie

Récemment, les systèmes en discrimination Parole/Musique ont suscité un grand intérêt pour plusieurs applications. En multimédia, ces systèmes peuvent être utilisés pour les tâches de classification : indexage, filtrage, archivage et récupération de données sonores ou vocales. La discrimination Parole/Musique joue un rôle primordial dans les systèmes de reconnaissance automatique de la parole (ou du locuteur) dans le but d'éviter la prise de mauvaises décisions dues à la confusion entre les segments de la parole et de la musique. De telles confusions risquent d'être catastrophiques pour certaines situations (exemple : un système de contrôle par commande vocale). Par ailleurs, le codage correspond à un autre genre d'applications où sa qualité est directement liée aux performances réalisées en discrimination Parole/Musique.

L'évolution de la dynamique des formants durant le phénomène de la production de parole, est l'une des propriétés les plus exploitées en discrimination de la Parole/Musique dans la plupart des publications. Son succès est attribué à l'alternance entre consonnes, voyelles et silences qui engendrent une irrégularité du signal au sens

rythmique. Généralement, les techniques proposées sont classiques et simples dans l'optique d'être applicables à la majorité des styles de musique (musique chantée non incluse) et être indépendantes du locuteur et du contexte linguistique. Cependant avec la quantité de données utilisées pour entraîner les modèles, on ne croit pas qu'elles le soient vraiment.

Le premier système de ce genre proposé par Saunders [99], a été implémenté au niveau d'un récepteur de radio. Son objectif est de prendre des mesures statistiques sur les programmes diffusés par une station donnée. Le système fonctionne en temps réel et estime les paramètres sur une longue fenêtre d'analyse, en se basant sur le taux de passage par zéro et sur l'énergie.

La moyenne des coefficients cepstraux, sur plusieurs trames, a été proposée par Spina et Zue [107]. Ils ont observé une bonne performance lorsque la parole occupe une largeur de bande de 4 KHz et la musique une largeur de bande de 16 KHz. Dans la situation où les deux signaux occupent la même largeur de bande, les performances chutent significativement. En effet, le sous-échantillonnage a plus d'impact sur le signal de la parole que sur de la musique, il réduit la durée originale des unités linguistiques véhiculées par le signal vocal. Par conséquent, l'irrégularité attribuée aux séquences linguistiques s'accentue et devient plus apparente. On pense que l'effet du sous-échantillonnage peut être exploité indirectement en augmentant simplement la taille de la fenêtre d'analyse et le pas d'entrelacement.

Carey et al [24] ont tenté d'examiner la différence entre les propriétés de la parole et de la musique au niveau perceptif. Ils supposaient que les paramètres prosodiques seraient de bons candidats. La fréquence fondamentale et l'énergie ont donc été proposées comme paramètres prosodiques et ont été comparées aux coefficients

MFCC pris pour paramètres de référence. Comme résultat, il s'est avéré que les coefficients *MFCC* (statiques+dynamiques) entraînés sur un modèle à mélanges de 64 *Gaussiennes* pondérées enregistrent la meilleure performance. Les données d'entraînement dans leur expérience, représentent une durée de 6 heures pour la parole extraites de différentes langues et une durée de 4 heures pour la musique. On tient ici à insister sur le fait que les coefficients *MFCC* utilisés ont été normalisés, un point qui sera d'ailleurs remis en question dans la cadre de cette thèse.

Williams et Ellis [117], ont proposé un réseau de neurone qui estime d'abord la probabilité à posteriori pour classer chaque trame dans l'une des 50 classes de phonèmes. Ensuite, ils ont traité l'information de la probabilité à posteriori via 4 formules statistiques pour fournir une décision finale sur le contenu réel de la trame (parole ou musique). Les formules statistiques sont inspirées de la théorie de l'information (entropie), de l'énergie et de leurs dérivées.

En se basant sur un groupe de paramètres, Zhang et Kuo [123] proposent un système hiérarchique à trois niveaux pour la classification, l'archivage et la localisation des données en parole, musique ou silence. Les auteurs affirment avoir obtenu de bons résultats.

Aarts et Ellis [1] ont proposé une segmentation du signal acoustique de types parole, musique, parole+musique et parole+bruit. Pour l'identification des segments de la parole corrompue par de la musique, ils estiment d'abord les pôles sur une longue durée par la méthode LPC. Ensuite ils procèdent par un filtrage inverse afin d'éliminer la contribution de la musique. En effet, le principe de leur démarche repose sur l'hypothèse que les pôles sont stables et réguliers sur une longue durée du signal musical.

Maes et al [64] ont décrit un circuit pour les applications en temps réel, composé de 3 modules de traitement pour discriminer les séquences de la parole et de la musique. La fonctionnalité respective des trois modules est la suivante :

- filtrage et normalisation ;
- extraction des paramètres tels que la montée, la descente, la durée d'une impulsion glottale et d'autres composantes spectrales ;
- combinaison par la logique floue (fuzzy logic).

Samouelian et al [98] de leurs côtés se sont orientés vers la transcription des programmes de télévision par leurs contenus en silence, bruit, parole et musique.

La nouvelle génération technologique des compresseurs et des codeurs à faible débit ont besoin de reconnaître la nature de l'information convoyée par le signal (musique ou parole) avant de la coder afin de garantir une meilleure qualité en codage. Précisément, au niveau de l'architecture, on trouve l'implémentation simultanée de deux codeurs, le premier est dédié au codage de la parole et le deuxième est dédié au codage de la musique [108].

Scheirer et Slaney [100] ont évalué de leurs cotés, 13 paramètres censés capturer certaines propriétés en caractérisation de la parole et de la musique. Les 13 paramètres sont essentiellement basés sur la structure fine du spectre, le taux de passage par zéro et l'énergie. L'originalité de leur travail, selon notre point de vue, réside dans l'étude portant sur la comparaison de différents classificateurs. Ils ont étudié la classification par la distance de *Mahalanobis*, par les modèles paramétriques à mélange de M *Gaussiennes* avec ($M = 1, 5, 20$ et 50), par l'approche dite des plus proches voisins notée $K-NN$ avec ($K = 1, 5, 11$ et 25) et par l'algorithme en arbre le $K-d$. D'après les résultats obtenus, les auteurs déduisent que :

- la différence en performance entre les différents classificateurs est minime, donc le type de classificateur utilisé reste sans importance pour la discrimination Parole/Musique ;
- l'augmentation du nombre de *Gaussiennes*, pour les GMM, des proches voisins, pour la méthode $K-NN$, ou des feuilles de l'arbre, pour la méthode $K-d$, n'ont aucune influence sur les performances ;
- il est plus facile de discriminer la parole que la musique par les modèles GMM, par la distance de *Mahalanobis* et par la méthode des plus proches voisins alors qu'avec la méthode spatiale $k-d$, les performances demeurent contrebalancées.

Les modèles paramétriques sont entraînés respectivement sur 20 mn de parole et 20 mn de musique avec une qualité radio (Fe=22.05 kHz). Toutefois, les auteurs ne prennent pas en considération l'enregistrement des données dans des conditions téléphoniques ni l'utilisation de combinés différents. La caractérisation différenciée en musique et en musique chantée comme étant deux entités distinctes, ne faisait pas non plus l'objet de leur étude, ce qui a probablement rendu (selon leurs résultats), la tâche de reconnaissance plus difficile pour la musique.

Ludovic et al [63] ont proposé une étude d'un système d'indexation qui se base sur une modulation différenciée pour chacune des classes : parole et musique. L'approche est mise en oeuvre à partir des modèles GMM (32 *Gaussiennes* pour la parole et 10 *Gaussiennes* pour la musique). Avec une analyse spectrale non-linéaire les scores obtenus sont de 95% pour la parole et 81.5% pour la musique.

Ces résultats sont cohérents avec les observations présentées dans l'étude [100] montrant qu'il est plus difficile d'identifier la musique que la parole. En procédant par une analyse spectrale linéaire (tous les filtres ont la même largeur de bande) pour

la musique tout en gardant une analyse spectrale non-linéaire pour la parole (échelle Mel), le taux de reconnaissance de la musique a été amélioré de 81.5% à 93%. La taille de la fenêtre d'analyse considérée est de 200 ms.

Récemment, El-Maleh et al [31] ont proposé la méthode de LSP (Line Spectral Frequency) et le taux de passage par zéro comme techniques de paramétrisation. Les tests se sont déroulés avec une fenêtre d'analyse de durée 20 ms, glissante à tous les 10 ms.

En résumé, les auteurs utilisent des coefficients estimés sur de longues durées temporelles (0,5 à 5 secondes), à l'exception de El-Maleh et al [31] qui ont utilisé une fenêtre de taille 20 ms. Cependant, les auteurs se sont aperçus d'une baisse de performance et ont été contraints d'augmenter la durée de la fenêtre d'analyse à une seconde pour atteindre les performances obtenues en moyenne par la plupart des systèmes ($\approx$ 92%). Les paramètres les plus utilisés généralement découlent de l'énergie, du taux de passage par zéro et de l'analyse spectrale. Les données utilisées sont en général de bonnes qualités.

Après ce tour d'horizon autour des revues bibliographiques en discrimination Parole/Musique, on retient cependant certaines lacunes dans le domaine. D'abord dans la base de données musicales, on ne trouve pas de la musique chantée. Ensuite, les expériences présentées en mode téléphonique ne tiennent pas compte de l'impact de la variabilité des combinées. Et finalement, les meilleures performances en discrimination Parole/Musique sont toujours observées avec des test de durée plus qu'une seconde. Dans la suite de chapitre, on s'intéresse à proposer de nouvelles approches en discrimination Parole/Musique (musique chantée afin de répondre aux limites associées à ces lacunes.

4.3 Établissement d'une base de données pour nos expériences

Comme nous n'avons pas accès à une base de données comprenant parole et musique, nous avons élaboré une base pour les besoins de la thèse.

4.3.1 Origine des données pour la musique

Les données de musique ont été obtenues à partir de différents sites internet de diffusion. On y trouve toutes sortes de qualités (qualité radio bande étroite ou qualité haute fidélité). Afin de simuler une transmission téléphonique, les signaux de musiques ont ensuite été retransmis sur une ligne téléphonique impliquant trois combinés différents. La largeur de bande est donc limitée avec l'introduction de l'effet du canal et la variabilité des combinés. À la réception, ils ont été réenregistrés et échantillonnés à 8000 Hz. Les enregistrements ont été effectués sur plusieurs sessions avec divers types de musique. On y trouve de la musique Classique (Beethoven, etc.), du Jazz, du Populaire, du Rock, de la musique Country, du Rap, du Western, de la musique extraite de quelques long-métrages cinématographiques (situations de suspense, romantique, parole+musique et parole d'acteur). De la musique provenant du standard téléphonique de l'université du Québec à Chicoutimi a aussi été utilisée.

4.3.2 Données pour la parole

Toutes les données de la parole sont extraites en partie de la base de données SPIDRE. On y trouve 45 locuteurs avec 4 conversations pour chaque locuteur enregistrées via trois combinés différents.

4.3.3 Mixage Musique/Parole proposé

La procédure de mixage adoptée pour simuler les conversations de type conférence avec alternance de musique, consiste à générer des fichiers contenant chacun 3 segments de parole et 3 segments de musique. Chaque segment a une durée de 1 seconde suivie de 10 ms de silence. Les segments sont répartis en alternance : parole, musique, parole, musique, etc. Les segments de parole sont extraits aléatoirement de la base SPIDRE qui comprend 45 locuteurs ayant chacun 4 conversations. Les segments de musique sont extraits aléatoirement de la base de musique qu'on a préparée. Dans chaque conversation de type conférence, les segments de parole et musique proviennent de différents locuteurs et de plusieurs styles de musique. Aucun calibrage, pondération ou normalisation n'a été appliqué sur les segments extraits que ce soit pour la musique ou la parole.

4.3.4 Critère de reconnaissance

Le score S de reconnaissance est défini comme étant le rapport du nombre de trames correctement classifiées C par le nombre total de trames testées T :

$$S = \frac{C}{T}. \tag{4.3.1}$$

4.3.5 Système de référence

En s'inspirant de la bibliographie présentée auparavant, le système de référence retenue pour fins de comparaison, se base sur les paramètres MFCC avec une normalisation du canal et sur les modèles GMM.

4.4 Caractéristiques distinctives entre la parole et la musique

La parole se distingue des autres sons par les propriétés acoustiques dont les origines proviennent du mécanisme de la production vocale. Pour chaque position articulatoire, le spectre des sons produits dépend des propriétés résonantes particulières du conduit vocal, des caractéristiques du rayonnement des lèvres et du processus vibratoire de la glotte. Notamment, la parole se caractérise par une structure formantique et non stationnaire qui reflète la résonance du conduit vocal en réponse à l'excitation de la source glottale. L'alternance des sons voisés, non-voisés et surtout le silence apporte un autre degré à la parole et lui permet de se distinguer davantage des propriétés des autres sons.

La musique à son tour, est caractérisée en général par une structure harmonique et stationnaire, un rythme répétitif, une absence de silence en général et une mélodie dépendante du style de musique.

On trouve judicieux de développer un système de discrimination Parole/Musique qui exploite les traits caractéristiques liés à la parole plutôt qu'à ceux de la musique. La raison motivante derrière ce choix repose essentiellement sur deux points. D'une part, on dispose d'outils mathématiques assez sophistiqués pour le traitement du signal vocal. D'autre part, on peut développer un système qui prend des décisions sur de courte durée puisque la structure du conduit vocal reste figée (forme stable) sur des durées n'excédant pas 200 ms. Il arrive que les durées atteignent un maximum allant jusqu'à 500 ms, mais de telles situations se produisent très rarement. On a donc orienté la paramétrisation vers la caractérisation liée à l'évolution des formants

du conduit vocal ce qui constitue l'objet des prochaines sections.

4.5 Stratégie à base de seuils

Cette section décrit de façon chronologique, les démarches suivies pour aboutir à la proposition d'une technique en caractérisation de la Parole/Musique. Cette technique ne nécessite aucune phase d'entraînement et la classification est effectuée en se basant sur la valeur d'un seuil pré-défini.

4.5.1 Paramétrisation et analyse

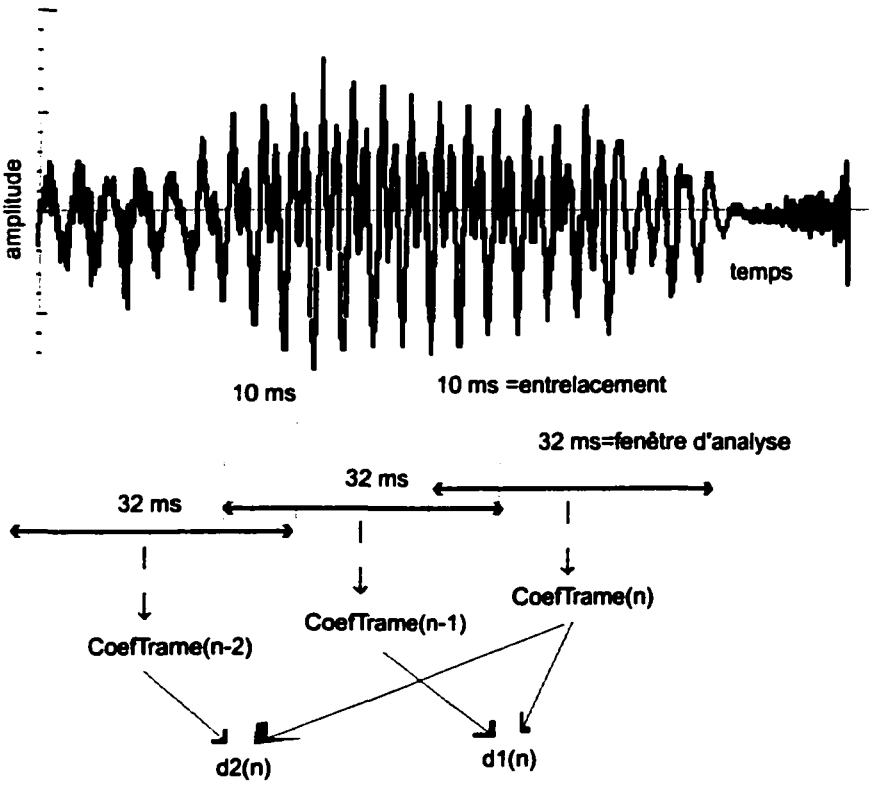

FIG. 4.1 – Calcul des distances métriques

On a débuté par l'extraction du vecteur paramètre $CoefTrame(n)$ correspondant à la n *ième* trame (pouvant être des $MFCC$ ou des $\Delta MFCC$). Pour chaque trame, on a pris une fenêtre d'analyse de durée 32 ms, glissante à tous les 10 ms. On a calculé une distance métrique $d_i(n)$, à l'instant n pour $i = 1$ et $i = 2$, à partir des coefficients $CoefTrame(n)$ de trois fenêtres adjacentes de n à à $n - 2$ (voir la figure 4.1). On a définie les distances métriques $d_1(n)$ et $d_2(n)$ par :

$$d_1(n) = DISTANCE(\Delta MFCC_i(n), \Delta MFCC_i(n-1)) \tag{4.5.1}$$

et

$$d_2(n) = DISTANCE(\Delta MFCC_i(n), \Delta MFCC_i(n-2)). \tag{4.5.2}$$

Les coefficients peuvent être n'importe quels types de paramètres sauf qu'ils doivent caractériser la contribution du conduit vocal, c'est à dire les formants. Dans notre cas, on a extrait les coefficients cepstraux c_1 à c_{12} respectivement par la méthode LPC, la méthode $MFCC$ et leurs dérivées (vitesse et accélération). Pour la distance métrique $DISTANCE()$, on a utilisé respectivement la distance *Euclidienne* pour les coefficients $MFCC$ et la distance d'*Itakura* pour les coefficients LPC.

Sur une base de données restreinte, on a calculé les distances d_1 et d_2 respectivement avec les coefficients $MFCC$ et les coefficients LPC. En s'appuyant sur des interprétations graphiques, on a observé que les coefficients $MFCC$ apportent une meilleure discrimination que les coefficients LPC. De plus, les coefficients $MFCC$ sont utilisés dans presque tous les systèmes impliquant un traitement du signal vocal et par conséquent leur exploitation ne requiert aucun calcul supplémentaire. Leur succès peut être attribué au fait qu'ils sont supposés être corrélés à la forme du conduit vocal et statistiquement modélisés convenablement par les GMM.

Dans la figure 4.2(a), on donne à titre d'exemple un fichier de données mixtes, contenant la conversation de deux femmes, deux hommes et six styles de musique. La durée totale du fichier est de 20 secondes répartie uniformément sur l'ensemble des signaux.

Les distances d_1 et d_2 superposées, sont illustrées dans la figure 4.2(b). La courbe en trait continu indique toutes les 10 ms la similarité entre la trame n et la trame $n-1$. La courbe en trait discontinu indique, toutes les 10 ms, la similarité entre la trame n et la trame $n-2$. On en déduit d'après le graphique que les deux courbes sont très proches et stables avec un niveau d'amplitude faible dans le cas de la musique. Pour la parole, on observe de très fortes fluctuations en amplitude et les deux courbes sont plus distantes entre elles. En fait, ces constatations sont liées directement aux propriétés de la stationnarité et de la rythmicité pour la musique et la non-stationnarité et l'irrégularité pour la parole.

D'après la figure 4.2(b), on se rend compte que la discrimination ne peut être effectuée de façon optimale, si on désire fixer une valeur d'amplitude comme seuil de classification. En effet, la taille de la fenêtre d'analyse utilisée pour évaluer les distances métriques est de l'ordre de 52 ms (32 ms +10 ms +10 ms), ce qui n'est pas en principe suffisant pour caractériser convenablement la variabilité phonétique. Comme solution évidente, on peut songer à augmenter simplement la taille de la fenêtre d'analyse, cependant avec un tel traitement on revient au principe déjà exploité par les méthodes classiques et qui est difficile à mettre en oeuvre pour le temps réel.

Dans notre cas, on a procédé par une analyse d'ordre statistique qui consiste à estimer les écart-types des distances $d_1(n)$ et $d_2(n)$ à partir d'un nombre de trames fixe. On suppose qu'une telle stratégie nous permettra de mieux faire ressortir la

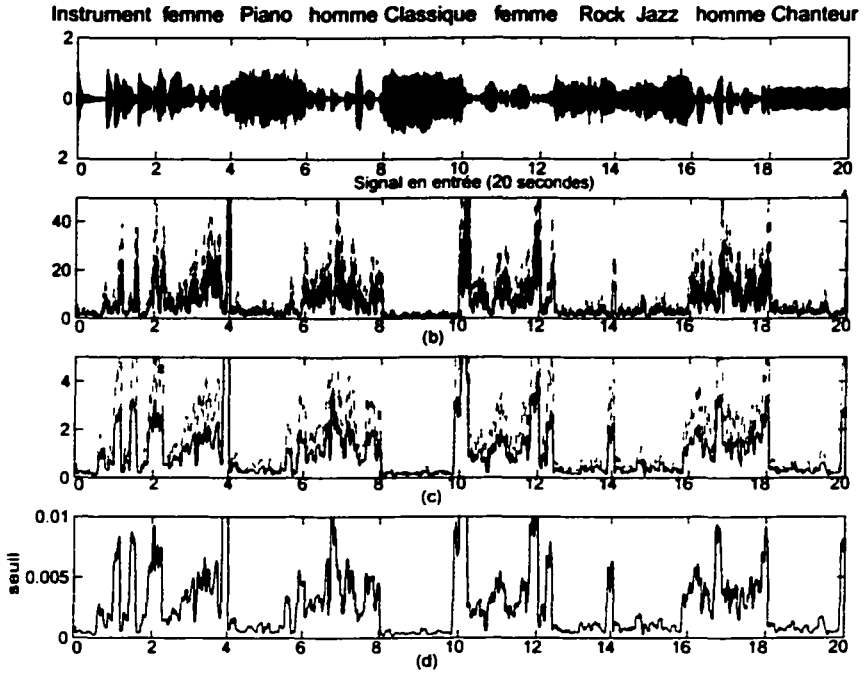

FIG. 4.2 - **a** : 20 secondes de Parole et de Musique séparées par des lignes discontinues, Chaque segment est de durée 2 secondes ; **b** : différence entre les coefficients $\Delta MFCC$ estimés à partir de segments adjacents à toutes les 10 ms tel que décrit par les équations : $d_1(n)$ et $d_2(n)$; **c** : écart-type des $d_1(n)$ et $d_2(n)$ sur une durée de 122 ms tel que décrit par les équations $\sigma_1(n)$ et $\sigma_2(n)$, **d** : application de la formule $P3_\sigma(n)$.

dynamique des formants tout en gardant une fenêtre d'analyse de petite taille. Les équations sont les suivantes :

$$\sigma_1(n) = \frac{\sqrt{\sum_{i=n-L}^{n}(d_1(n) - \overline{d_1})^2}}{L} \tag{4.5.3}$$

et

$$\sigma_2(n) = \frac{\sqrt{\sum_{i=n-L}^{n}(d_2(n) - \overline{d_2})^2}}{L}. \tag{4.5.4}$$

$\overline{d_1}$ et $\overline{d_2}$ sont les moyennes respectives de d_1 et d_2. Elles sont estimées sur le même intervalle que les $\sigma_1(n)$ et $\sigma_2(n)$. Le paramètre L correspond au nombre de trames

utilisées pour l'évaluation des variables statistiques.

Dans la figure 4.2(c), on illustre les courbes obtenues à partir des équations 4.5.3 et 4.5.4. Les équations sont calculées par rythme de 10 ms et pour un nombre de $L = 10$ trames ce qui correspond au total à une durée de 122 ms (32 ms+9x10 ms).

On remarque généralement que ces deux courbes sont plus distantes pour la parole et plus rapprochées entre elles dans le cas de la musique. De façon générale, les courbes de la variance suivent, dans leurs évolutions, les courbes des distances Euclidiennes, mais la discrimination en Parole/musique devient un peu plus évidente si on se fie au niveau d'amplitude.

D'un autre côté, quand l'enveloppe du signal de musique subit de grandes variations en modulation d'amplitude, les courbes ont tendance à avoir une allure qui colle plus aux caractéristiques de la parole. Cette situation est rarement rencontrée et on peut l'ignorer et se concentrer sur les caractéristiques plutôt à aspect global.

Si on veut expliquer le comportement des courbes $\sigma_1(n)$ et $\sigma_2(n)$ de façon analytique, il suffit d'analyser les équations 4.5.1 et 4.5.2 proposées à la source du traitement.

Avec l'équation 4.5.1, la différence entre les coefficients est estimée à partir de deux fenêtres adjacentes, décalées de 10 ms alors qu'avec la deuxième équation 4.5.2 les deux fenêtres adjacentes sont décalées de 20 ms. Or, si on tient compte du phénomène de la coarticulation pour la parole, on pourra présumer que la distance décrite par la première équation sera plus faible que celle décrite par la deuxième l'équation. Cependant, la contrainte de la rythmicité et de la stabilité pour le signal musical nous laisse prévoir que les distances devraient être très proches. En conséquence, les équations $\sigma_1(n)$ et $\sigma_2(n)$ consistent à étendre ces faits sur la base de plusieurs

trames. Par conséquent, on en déduit que la taille de la fenêtre d'analyse, le pas de chevauchement entre fenêtres et le nombre de trames considérés dans les statistiques sont des facteurs qui influencent directement les performances et il y aura lieu de les prendre en considération.

En s'appuyant sur les résultats graphiques, on a dérivé quatre nouvelles formules dans lesquelles, cette fois-ci, les informations liées aux écart-types $(\sigma_1(n),\sigma_2(n))$ et à leur différence sont combinées ensemble. Chaque formule se caractérise par une sorte de pondération typique des deux entités combinées. Les équations proposées sont les suivantes :

$$P1_\sigma(n) = (\sigma_1(n) - \sigma_2(n))^2 \tag{4.5.5}$$

ou

$$P2_\sigma(n) = (\sigma_1(n) - \sigma_2(n))^2 + max(\sigma_1(n), \sigma_2(n))^2 \tag{4.5.6}$$

ou

$$P3_\sigma(n) = (\sigma_1(n) - \sigma_2(n))^2 + max(\sigma_1(n), \sigma_2(n)) \tag{4.5.7}$$

ou

$$P4_\sigma(n) = (\sigma_1(n) - \sigma_2(n)) + max(\sigma_1(n), \sigma_2(n)). \tag{4.5.8}$$

L'objectif visé par l'application d'une pondération consiste à garder une variabilité inter-classes constante entre la musique et la parole, puisque chaque type de formule dépend différemment du rapport $(\sigma_1(n) - \sigma_2(n))/max(\sigma_1(n), \sigma_2(n))$. En effet, ce dernier dépend directement de la taille de la fenêtre d'analyse, du pas de chevauchement et du nombre de trames sur lesquelles sont estimées les mesures statistiques.

À titre d'exemple, on a reporté une illustration graphique de la formule $P3_\sigma(n)$ à la figure 4.2(d). On observe une petite amélioration sur la différence en amplitude entre

la parole et la musique. D'après les graphiques, il sera donc avantageux d'appliquer l'équation 4.5.8($P4_{\sigma}(n)$) pour rehausser significativement les performances.

Entre autres, on a repris les mêmes équations de 4.5.5 à 4.5.8 sous d'autres formes statistiques notées respectivement $P1_{\mu}(n)$, $P2_{\mu}(n)$, $P3_{\mu}(n)$ et $P4_{\mu}(n)$, en y remplaçant les écart-types $\sigma_1(n)$ et $\sigma_2(n)$ par des moyennes $\mu_1(n)$ et $\mu_2(n)$ qui sont estimées sur le même intervalle que les $\sigma_1(n)$ et $\sigma_2(n)$.

Il est temps maintenant de proposer une seule formule cohérente avec les conditions expérimentales qu'on utilise, avec une paramétrisation adéquate pour la mise oeuvre d'un système en discrimination Parole/Musique. Ceci fera l'objet du prochain paragraphe.

4.5.2 Optimisation avec données de simulations en mode conférence

Dans cette section, on présentera d'abord le critère statistique adopté pour la recherche d'une solution optimale pour notre problématique. Ensuite, on présentera une étude qui consistera à trouver la combinaison optimale impliquant différents vecteurs paramètres, méthodes de classification et valeurs de seuils dans le but de réaliser la meilleure classification possible entre parole et musique.

4.5.3 Critère utilisé pour la recherche de seuils optimaux

Le critère utilisé pour trouver le seuil optimal est la règle de Bayes dans le cas particulier de classification par erreur minimale. Cette règle consiste à minimiser le risque total ou à maximiser le produit de la vraisemblance d'une classe C_j, associée à une observation X, par la probabilité à priori de la classe C_j tel que :

$X \in C_i$ si $P(X/C_i)P(C_i) > P(X/C_j)P(C_j)$ pour $\forall j \neq i$

$P(X/C_i)$ est la probabilité conditionnelle de X ;

$P(C_i)$ est la probabilité à priori.

Dans notre cas, on dispose de deux classes respectivement pour la musique et la parole. On a supposé que chaque classe est équiprobable ($P(C_i) = P(C_j)$).

L'observation X est considérée comme la valeur estimée par l'une des équations 4.5.5 à 4.5.8 proposées auparavant. La probabilité pour chaque observation X est donc définie par sa fréquence relative. Puisque le vecteur paramètre est de dimension 1, la recherche d'un seuil optimal par la règle de classification par erreur minimale revient simplement à trouver la frontière coïncidant entre les classes (musique et parole) où les probabilités conditionnelles associés à chaque classe sont identiques : $P(X/C_{parole}) = P(X/C_{musique})$.

4.5.4 Résultats

Dans une première expérience, on s'est intéressé à trouver la combinaison optimale entre les vecteurs paramètres utilisés et les méthodes de classification proposées. Comme vecteurs paramètres, on a expérimenté les coefficients $MFCC$ statiques notés (CC_Stat) et les coefficients $MFCC$ dynamiques notés (CC_Dyn). On a extrait 12 coefficients $MFCC$ allant de c_1 à c_{12}. Les méthodes de classification représentées par les 4 formules statistiques 4.5.5 à 4.5.8 ont été également examinées à tour de rôle.

Les données de cette première expérience sont des données mixtes et simulent des conversations en mode conférence décrites dans la sous-section 4.3.3. Dans chacun des 8 fichiers, on a 3 secondes de parole et 3 secondes de musique.

À la table 4.1, on donne les performances d'identification pour l'ensemble des locuteurs impliqués dans la simulation en mode conférence (c-à-d : ce sont uniquement les

scores compilés sur les sections de parole). À la table 4.2, on trouve les performances d'identification de musique compilées uniquement sur les sections de musique. Dans la première colonne, on spécifie le type de paramètre acoustique utilisé (ex : *CC_Stat* ou *CC_Dyn*) et le genre de formule de classification proposé (ex : $P2_\mu$, $P3_\sigma$, $P4_\mu$ ou $P4_\sigma$). Les colonnes de 1 à 8 correspondent aux scores obtenus sur chacun des 8 fichiers de simulation. Le seuil de discrimination a été fixé à 0.01 et les statistiques sont évaluées sur 19 trames (212 ms) entrelacées.

TAB. 4.1 - Taux de Reconnaissance (en %) des différents paramètres proposés basés sur un seuil fixe à partir des 8 fichiers simulant les conversations de type conférence. Le μ_{parole} est le taux moyen des 8 conversations. Ici, les scores sont compilés uniquement sur les portions de parole.

SimConf	1	2	3	4	5	6	7	8	μ_{parole}
$P4_\sigma(CC\_Stat)$	83.1	77.6	91.4	82.5	65.6	85.9	81.3	81.9	81.2
$P4_\mu(CC\_Stat)$	85.3	85.9	90.5	85.9	68.1	89.0	83.4	89.6	84.7
$P3_\sigma(CC\_Stat)$	82.8	78.2	92.0	82.5	68.7	86.8	80.4	83.4	81.9
$P3_\mu(CC\_Stat)$	82.5	80.1	89.6	83.7	56.1	78.8	74.5	81.9	78.4
$P2_\sigma(CC\_Stat)$	80.7	73.3	82.5	74.8	65.0	80.1	73.0	74.2	75.5
$P2_\mu(CC\_Stat)$	82.5	75.8	85.9	83.1	53.1	81.6	73.9	76.7	76.6
$P3_\sigma(CC\_Dyn)$	91.4	82.8	93.9	92.9	77.3	93.6	85.0	91.4	88.5
$P3_\mu(CC\_Dyn)$	90.5	92.9	92.0	88.3	78.2	92.3	89.9	91.7	89.5
$P4_\sigma(CC\_Dyn)$	90.2	79.8	93.6	92.6	76.1	93.6	83.1	89.9	87.3
$P4_\mu(CC\_Dyn)$	90.2	92.6	91.7	88.0	76.4	92.0	89.3	91.7	89.0

Dans la deuxième expérience, on s'est intéressé cette fois-ci à trouver la valeur du seuil optimal, toujours selon la règle de Bayes. Donc le type de vecteur de paramètres et la méthode de classification ont été fixés en se basant sur les résultats obtenus de la première expérience. Cependant les données pour cette expérience ont été largement augmentées et plusieurs variétés ont été ajoutées. La description sur les types et les durées des données utilisées dans cette expérience, figurent sur la table 4.4. On y retrouve une colonne pour les données d'entraînement et une autre pour les

TAB. 4.2 - Taux de Reconnaissance (en %) des différents paramètres proposés basés sur seuils fixes sur les 8 fichiers simulant les conversations de type conférence. Le $\mu_{musique}$ est le taux moyen sur les données de musiques seules.

SimConf	1	2	3	4	5	6	7	8	$\mu_{musique}$
$P4_{\sigma}(CC\_Stat)$	94.2	81.3	65.3	82.8	89.0	87.4	76.4	84.0	82.6
$P4_{\mu}(CC\_Stat)$	97.2	98.5	50.9	64.1	67.2	78.8	63.8	92.3	76.6
$P3_{\sigma}(CC\_Stat)$	94.2	81.0	63.8	82.5	89.0	87.4	76.4	83.1	82.2
$P3_{\mu}(CC\_Stat)$	100.0	99.7	58.9	72.7	78.8	89.0	66.3	96.3	82.7
$P2_{\sigma}(CC\_Stat)$	84.7	79.8	67.2	79.8	81.9	81.6	70.6	86.5	79.0
$P2_{\mu}(CC\_Sat)$	100.0	100.0	77.9	93.9	90.8	96.6	76.4	98.2	91.7
$P3_{\sigma}(CC\_Dyn)$	84.0	82.8	72.7	83.7	78.5	80.7	74.2	77.0	79.2
$P3_{\mu}(CC\_Dyn)$	79.4	89.3	71.5	88.3	64.4	86.5	77.6	83.1	80.0
$P4_{\sigma}(CC\_Dyn)$	85.0	83.1	72.7	85.0	81.3	81.6	75.2	76.7	80.1
$P4_{\mu}(CC\_Dyn)$	84.0	89.9	74.2	90.5	71.5	89.3	78.5	87.4	83.2

données de tests avec leur durée respective. On rappelle que les données d'apprentissage ne sont pas utilisées dans la phase des tests. Les combinés téléphoniques utilisés sont également différents entres les sessions d'apprentissage et de test (à quelques exceptions pour la musique). La table 4.3 et la figure 4.3 reportent les résultats pour différentes valeurs de seuils, testés sur divers styles de musique et de conversations provenant de plusieurs locuteurs et locutrices avec une variabilité de combinés téléphoniques importante.

Discussion

De façon générale, d'après les résultats des tables 4.1 et 4.2, si on fixe le type de paramètre acoustique et on compare uniquement les méthodes de classification, on trouve que les scores obtenus pour les différentes formulations de classification sont presque similaires. Donc, la préférence ou la sélection de l'une de ces formulations reste sans importance.

Par contre, si on analyse les scores avec un scénario inverse en comparant les performances selon le type de paramètres employés sur la base d'une classification commune. On trouve que les coefficients dynamiques $MFCC$ ont amélioré les performances de l'ordre de 8% en discrimination de la parole comparativement aux coefficients $MFCC$ statistiques. Dans le cas de la musique, aucune amélioration importante n'a été enregistrée. En effet, les coefficients dynamiques $MFCC$ en principe, sont extraits de manière à tenir compte de la dépendance temporelle des séquences prononcées, et donc nécessitent quelques trames de plus. Les coefficients statiques $MFCC$, au contraire encodent l'information du message contenue par chaque trame.

Si on effectue une comparaison par le contenu de chaque fichier de simulation, on trouve que les chutes de performance rencontrées pour le cas de la musique sont attribuées à la musique de type chanté. Pour la parole, le principal facteur de dégradation est dû à la présence du silence de durée assez importante, considéré dans l'étiquetage comme étant de la parole, alors que le système le classifie comme étant de la musique.

Dans la deuxième expérience à la table 4.3, on observe une très bonne discrimination de la parole vis-à-vis de la musique lorsque le seuil est bas. Pour des seuils élevés, le scénario est complètement renversé. Pour des valeurs de seuil moyennes, on obtient des scores témoignant d'une discrimination balancée entre parole et musique. On peut alors ajuster le seuil, selon les tolérances espérées et les besoins demandés. Cependant, on n'est jamais arrivé à effectuer une discrimination parfaite avec un taux de 100%. Il en découle que les espaces de représentation pour la musique et la parole ne sont pas complètement disjoints et il y aura lieu d'ajouter un traitement supplémentaire dans l'objectif de rehausser les scores. On pense que l'énergie ou le taux de passages

par zéro peuvent améliorer considérablement la robustesse du système.

En termes d'efficacité, on observe que pour les styles de musique caractérisés par un rythme régulier (classique, douce,..), les performances sont excellentes. Par contre pour la musique chantée, les performances chutent relativement au degré du couplage entre le chanteur et la musique.

La particularité de cette méthode est qu'elle est indépendante de l'énergie et en quelque sorte du combiné puisqu'on effectue une différence des paramètres entre segments adjacents.

Dans la figure 4.3, on reporte sous une forme graphique les mêmes résultats présentés à la table 4.3. Dans ce cas, la classification selon la règle de Bayes revient à trouver le lieu des frontières entres les classes, qui correspondent au seuil optimal. Puisque le vecteur paramètre $P3_\sigma(n)$ est de dimension 1, on peut considérer l'espace paramètres divisé en deux demi-plans respectivement pour la musique et la parole. La droite séparant les deux demi-plans en question correspond au seuil optimal.

Dans la partie 4.3(a), on donne pour différents seuils, les scores obtenus respectivement pour la parole (en trait continu) et pour la musique (en trait discontinu). La courbe en trait (-*-) dans les trois sous-figures, indique le taux moyen de classification. Le point d'intersection des 3 courbes est donc le seuil optimal de coordonnées 000.1 (valeur) et 90% (score). On remarque que l'allure de la courbe pour la parole décroît de façon linéaire alors que celle de la musique suit une courbe quadratique. On peut déduire que les variances des classes ne sont pas similaires. Cependant, si on veut introduire la notion du coût selon Bayes en favorisant l'identification d'une classe au profit d'une autre, il suffit de déplacer le seuil selon les perspectives souhaitées.

Dans la partie 4.3(b), on reproduit la même expérience mais cette fois-ci, avec

les données de la parole et les données de la musique chantée. On s'aperçoit que le seuil optimal est situé à l'extérieur de l'intervalle des valeurs proposées. D'après la figure, il semble que la performance associée au seuil a baissé à 80%. Cela montre que le chevauchement entre les classes parole et musique chantée est plus important que dans le premier cas 4.3(a).

Dans la partie 4.3(c), on a de nouveau repris la même expérience avec les données de la classe parole et les données de la classe incluant à la fois la musique et la musique chantée. Le seuil optimal (sécante des courbes) est situé à la valeur 0.0014 avec un score optimal de 84%.

En fait, le choix de seuil reste relatif aux besoins de chaque application et peut être aisément déterminé à partir des graphiques. Cependant, les valeurs de seuil illustrées ici restent valables uniquement pour les conditions de paramétrisation mentionnées auparavant. Le cas éventuel, où une des conditions a été modifiée, il y aura lieu de réoptimiser les valeurs seuils.

L'intérêt de l'approche par seuil, est qu'elle utilise directement les coefficients $\Delta MFCC$ pour effectuer la classification. Elle nécessite un temps de calcul négligeable. Les performances obtenues sur une fenêtre d'analyse de 100 ms sont comparables à celles obtenues dans d'autres publication et qui opèrent avec une fenêtre d'analyse de durée 1000 ms. En plus, les données d'expériences sont enregistrées dans les conditions téléphoniques.

4.6 Stratégie par modélisation paramétrique

Dans cette section, on exposera les grandes lignes de la deuxième approche pour la discrimination en Parole/Musique. Comme stratégie de classification, on proposera

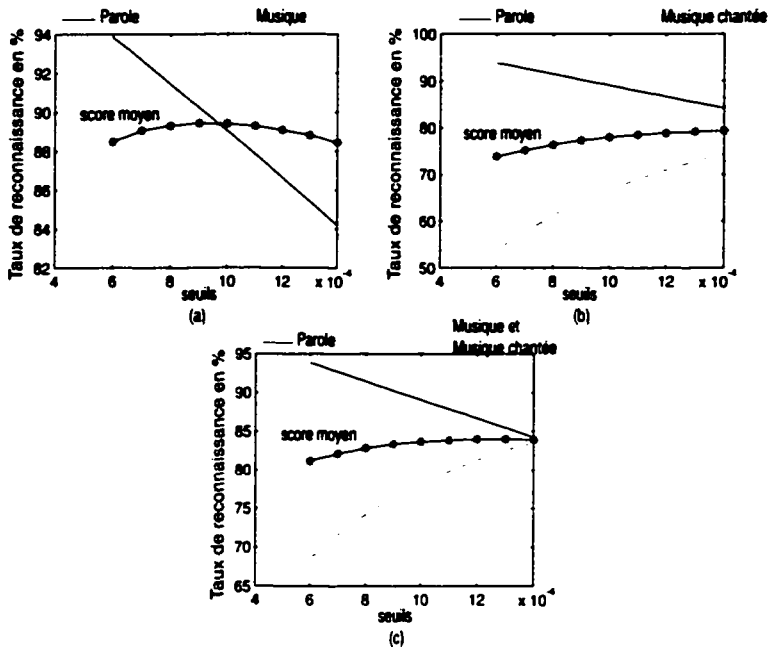

FIG. 4.3 - **Relations entre frontières de classes et critère de Bayes par minimisation d'erreur.**

une modélisation différenciée à base de GMM. On montrera que le découpage de l'espace de la musique en deux sous-espaces en musique et musique chantée entraînera l'obtention de meilleurs scores en comparaison à la démarche classique utilisant un seul modèle sur l'espace entier de la musique. Dans le cas de la musique chantée, on a trouvé que la normalisation des coefficients affecte considérablement les performances comparativement à la situation où on procède sans normalisation.

4.6.1 Classification et modèles

Comme technique de classification, on a proposé ici les modèles à mixtures de *Gaussiennes* pondérés dont la matrice de covariance est diagonale. Tous les modèles utilisés dans cette expérience sont seulement composés de 8 *Gaussiennes*. En effet,

avec 8 *Gaussiennes*, on a supposé que pendant la phase d'entraînement les modèles n'apprennent pas les caractéristiques spécifiques liées à la variabilité interlocuteurs, intralocuteurs et les différences entre les styles de musique. On a pu expérimenter cette technique en passant par l'examen de deux approches. Dans la première approche, on a considéré deux modèles GMM, le premier est entraîné avec des données vocales et le deuxième avec des données musicales. On s'attend à ce que les segments de la musique chantée soient absorbés par l'un des deux modèles suivant le degré du couplage entre le son des instruments et la voix du chanteur. Comme deuxième approche, on a entraîné 3 modèles respectivement pour la parole, la musique et la musique chantée. Chaque modèle est entraîné sur son propre espace de données. Dans tous les traitements de cette approche, les paramètres utilisés sont les coefficients statiques $c1$ à $c12$ ($MFCC$).

La description sur les types et les durées des données utilisées dans cette expérience, figure sur la table 4.4. On y retrouve une colonne pour les données d'entraînement et une autre pour les données de tests. On rappelle que les données d'apprentissage ne sont pas utilisées dans la phase des tests. Les combinés téléphoniques utilisés sont également différents entres les sessions d'apprentissage et du test (à quelques exceptions pour la musique). On remarque que la quantité des données d'apprentissage est presque 8 fois plus élevée que les données de test.

4.6.2 Résultats

Dans la table 4.5, on reporte les résultats pour la première perspective impliquant l'utilisation de deux modèles. Les paramètres n'ont pas été normalisés dans l'optique d'éliminer ou de réduire l'effet du canal. Dans la table 4.6, on reprend la même expérience mais cette fois-ci avec la normalisation des paramètres.

Dans chaque table, on trouve à la première colonne le nom de la catégorie de musique utilisée. Dans la deuxième et la troisième colonne, on trouve les scores obtenus respectivement pour les durées de 100 ms et 200 ms (10 ou 20 fenêtres d'analyses de 32 ms, glissantes à tous les 10 ms). Dans la dernière colonne, on trouve la durée totale du test pour chaque catégorie.

Dans la table 4.7, on donne les résultats pour la deuxième perspective impliquant l'utilisation de 3 modèles. Les conditions expérimentales sont exactement identiques à celles utilisées pour les deux tables précédentes 4.5 et 4.6.

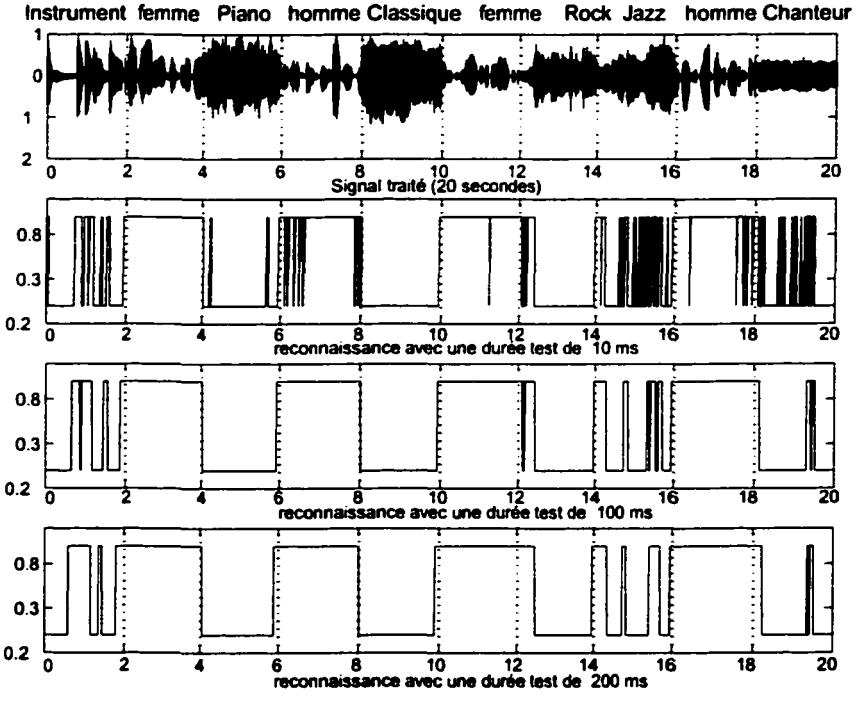

FIG. 4.4 – Reconnaissance avec deux modèles GMM, le premier est entraîné avec de la parole et le deuxième est entraîné avec la musique sans prendre en compte la musique chantée (Parole=1 and Musique=0).

En comparant les tables 4.5 et 4.6, on note que la performance de la parole n'a pas

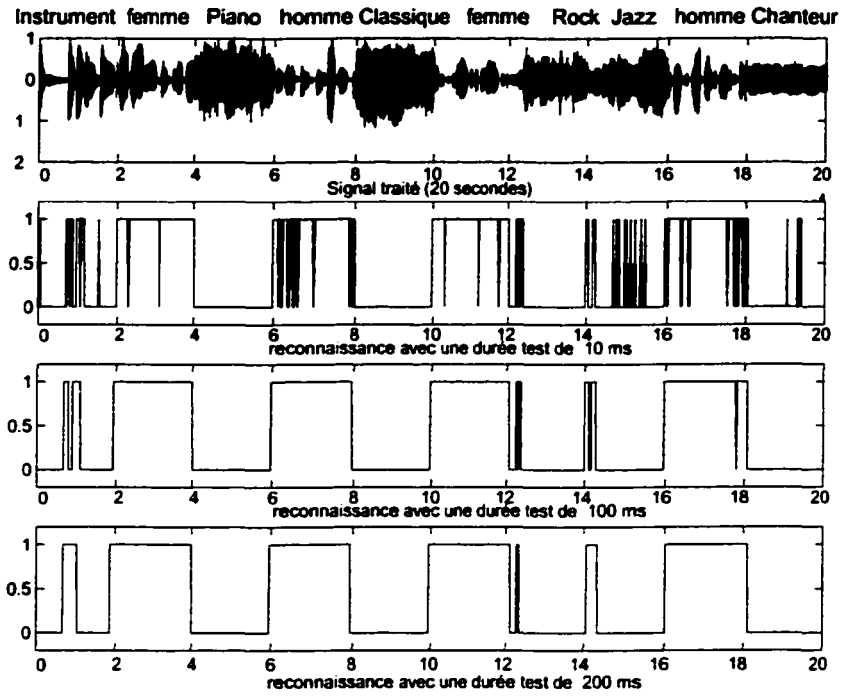

FIG. 4.5 – **Reconnaissance avec trois modèles GMM, le premier est entraîné avec de la parole, le deuxième avec de la musique et le troisième avec de la musique chantée (Parole=1 and Musique=0).**

été affectée par la procédure de normalisation du vecteur paramètres. Par normalisation, on entend la soustraction de la moyenne de l'ensemble des vecteurs paramètres de tous les vecteurs paramètres. Cependant, les performances ont chuté de 10% pour la musique et de 20% pour la musique chantée lorsque aucune normalisation des paramètres, n'a été effectuée. Ce dernier point peut être expliqué par les propriétés stables de la rythmicité et stationnarité caractérisant à long terme la musique et qui sont éliminées par la procédure de normalisation.

Avec la table 4.7, les plus hauts scores sont obtenus ce qui confirme que l'approche qui considère que la musique et la musique chantée sont deux classes différentes constitue une bonne stratégie. Avec une durée test de 100 ms, les scores sont comparables

aux meilleurs travaux publiés dans ce domaine alors que nos conditions expérimentales sont plus hostiles et difficiles. En comparant les performances obtenues pour les durées respectives de 100 ms et 200 ms, on peut dire qu'elles sont très proches. Donc, si on réduit de moitié la durée à 50 ms, on peut s'attendre à obtenir des performances similaires pour les femmes.

Dans la figure 4.4, on illustre graphiquement les décisions prises par le système qui se caractérise par deux modèles GMM sans effectuer la normalisation au niveau des paramètres. Et dans la figure 4.5, on illustre les décisions prises par le système qui se caractérise par trois modèles GMM sans effectuer la normalisation au niveau des paramètres en présentant les mêmes données. Les décisions fournies sont calculées respectivement sur des durées de tests de 10 ms, 100 ms et 200 ms, avec une fenêtre d'analyse de 32 ms, glissante à tous les 10 ms. Particulièrement, on trouve que la majorité des mauvaises décisions prises par le premier système à deux modèles GMM, a été correctement identifiée par le système à trois GMM.

4.6.3 Discussion

À priori, en modélisation paramétrique, les données d'entraînements doivent être présentées de façon exhaustive dans toute session d'apprentissage. Ainsi, pendant le processus d'entraînement, les paramètres du modèle seront mieux ajustés pour représenter analytiquement et de façon convenable, la distribution statistique des données. De plus, dans le cas de la modélisation par mélange à mixture de Gaussiennes, le nombre de Gaussiennes à utiliser est directement proportionnel à la variance inter-classes et surtout à la variance intra-classes. Pour notre problématique, une telle situation peut donc être très coûteuse en temps de calcul si on tient compte

du nombre élevé de styles de musique et du nombre de propriétés caractéristiques du language et du locuteur. On risquera donc de dépasser les limites permises du matériel de la machine. Comme solution alternative, on peut songer à caractériser chaque style de musique et chaque phonème par un modèle à plusieurs états, cependant le temps d'exécution sera considérablement long et le système ne pourra peut être pas fonctionner pour les applications en temps réel.

De plus, on pense qu'une telle démarche, a tendance à modéliser conjointement la variabilité intra-classes et la variabilité inter-classes. Or, ce qui nous intéresse particulièrement est la modélisation des propriétés distinctives (aspect général) entre la musique et la parole.

Pour cette raison, on a limité d'une part le nombre des *Gaussiennes* des modèles proposés à 8 et d'autre part on a réduit significativement la quantité de données pendant la phase d'entraînement. En effet, en moyenne une durée d'une seconde a été extraite d'un certain nombre de styles de musique. On a procédé de la même manière dans le cas de la parole en considérant 6 locuteurs dans le processus d'entraînement. Pour chaque locuteur, on a présenté 1 minute de parole pour l'ajustement des paramètres du modèle de la parole. Bien que dans d'autres travaux, un nombre de 64 composantes GMM ait été utilisé pour la même problématique et une quantité énorme de données d'entraînements, les résultats obtenus sont très convaincants. Évidemment, on peut remettre en question la notion de comparaison puisqu'on ne travaille pas avec la même base de données. Or, pour les systèmes de reconnaissance de parole ou du locuteur les mauvaises performances obtenues sont réalisées dans des conditions expérimentales similaires à celles qu'on utilise, ce qui devra probablement constituer un atout important pour nos résultats.

Avant de terminer cette section, il y a un dernier point qu'on doit éclaircir pour éviter toute confusion et interprétation. Le fait que toutes les données de musique sont extraites de façons différentes de celles de la parole, laisse croire que la discrimination peut être influencée par les aspects d'enregistrement et non pas par les caractéristiques des signaux sonores. En admettant que ceci est vrai, cela ne dévalorise pas les performances du système puisque l'objectif visé est de réaliser une discrimination en mode conférence (différents supports et combinés). Au contraire, l'impact des supports redevient un avantage. Cependant, dans l'expérience utilisant une modélisation à 2 GMM, les segments de la musique chantée ont été souvent confondus par la parole, alors qu'en réalité on traitait des données d'un même support d'enregistrement. Pour conclure, on peut admettre l'influence du canal dans la discrimination, mais cette influence reste toutefois faible comparativement à celle des signaux sonores.

4.7 Conclusion

Dans ce chapitre, on s'est intéressé à la discrimination Parole/Musique en proposant deux approches qui exploitent, en principe, la structure irrégulière des formants. La première approche se base sur un seuil fixe et ne requiert aucune session d'entraînement. Le seuil peut être estimé à partir de 4 formules proposées, d'ordre statistique. Le choix de l'une des formules est implicitement liée à la taille de la fenêtre d'analyse et au nombre de trames considérées.

Différents paramètres ont été analysés dans cette perspective. Les coefficients dérivées des $MFCC$ donnaient les meilleures performances en caractérisation de la musique/parole, suivis des coefficients cepstraux dérivés du LPC et finalement des coefficients LPC. Afin de simplifier le compte–rendu, nous n'avons présenté que

l'étude et les expériences mettant en jeu uniquement les coefficients $MFCC$. Particulièrement, les coefficients statiques ou dynamiques apportent des taux de reconnaissance similaires avec la musique. Cependant, les coefficients dynamiques apportent de meilleures performances dans le cas de la parole. En moyenne les scores sont de l'ordre de 90% et sont grandement affectés par les données de la musique chantée. Cependant le système est simple et très rapide en temps de calcul.

Comme deuxième approche, on a utilisé une modélisation paramétrique à base de GMM. D'abord, on a montré que la normalisation des paramètres a fait chuter significativement le taux de reconnaissance de la musique. Dans le cas de la parole le taux de reconnaissance est resté presque inchangé. De plus, la chute des scores est deux fois plus importante pour la musique chantée que pour la musique non chantée.

Ensuite, on a montré qu'avec l'utilisation de 3 modèles GMM entraînés sur la parole, la musique et la musique chantée, on a obtenu les meilleurs résultats. Avec une durée de test de 200 ms, les scores obtenus sont de 97.93% pour la parole, 98.55% pour la musique et 90.77% pour la musique chantée. Le système de référence utilisait seulement deux modèles GMM pour la musique et pour la parole. Les scores obtenus pour ce dernier sont de 96.62% pour la parole, 92.58% pour la musique et 69.31% pour la musique chantée.

Les scores obtenus pour les durées de tests de 100 ms et 200 ms, sont très proches, donc on suppose que la tendance peut se maintenir si on prend une durée test de 50 ms dans le cas de voix de femmes.

Avec ce qu'on a présenté jusqu'ici, on suppose qu'on a répondu à la première tâche que doit réaliser un système de reconnaissance de locuteur en mode conférence. Cette tâche consiste tout simplement à discriminer les segments de musique des segments

de parole. Les segments de musique seront rejetés ou ignorés et aucune décision ne sera prise par le système. Par contre, une fois qu'un segment acoustique est classé comme étant un segment de parole, il sera envoyé à un deuxième module, plus bas au sens architectural, dédié pour la tâche de la reconnaissance de l'identité de locuteur. Le traitement de ce dernier module a fait l'objet des chapitres précédents où on a proposé de nouveaux paramètres et modèles dans l'optique de concevoir un système de reconnaissance plus robuste aux distorsions introduites par le support de transmissions téléphoniques et la variabilité des combinés.

Seuil	0.0006	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012	0.0013	0.0014
Classique	82.63	84.97	86.68	88.09	89.23	90.13	90.92	91.63	92.63
Standard	93.80	94.93	95.73	96.39	96.90	97.18	97.43	97.63	97.77
Jazz	77.52	80.38	82.69	84.58	86.04	87.49	88.64	89.70	90.55
Rock	81.25	84.21	86.62	88.74	90.27	91.94	93.41	94.37	95.05
Metal	95.69	97.13	98.13	98.74	98.94	99.08	99.16	99.26	99.30
Western	70.31	73.67	75.90	77.82	79.37	80.49	81.52	82.38	83.16
Country	80.65	82.98	84.53	86.14	87.49	88.65	89.53	90.32	90.95
$\mu_{Musique}$	83.12	85.46	87.18	88.64	89.74	90.70	91.51	92.18	92.68
BluesChanté	60.10	64.07	67.96	71.31	74.19	76.55	78.43	80.19	81.58
CountryChanté	67.07	71.78	75.66	78.71	81.15	83.22	85.33	87.02	88.16
RapChanté	44.28	49.45	54.36	58.85	63.10	66.96	70.17	72.84	74.96
Films	79.47	82.17	84.37	85.96	86.02	87.08	88.07	88.96	89.73
RockChanté	21.80	25.68	29.05	32.80	36.08	39.17	41.89	44.40	46.71 7
VariousChanté	74.94	78.33	80.95	82.78	84.67	85.93	87.08	88.00	88.90
ReggaeChanté	29.51	32.78	36.25	39.64	42.12	44.37	46.49	49.02	50.92
$\mu_{MusiqueChanté}$	53.88	57.75	61.22	64.29	66.76	69.04	71.06	72.91	74.42
μ_{Parole}	93.83	92.67	91.42	90.23	89.05	87.84	86.64	85.43	84.20
μ_{Total}	76.94	78.63	79.94	81.05	81.85	82.53	83.07	83.51	83.77

TAB. 4.3 – Résultats en reconnaissance Parole/Musique basés sur un critère de seuil.

TAB. 4.4 - Description sur les durées de données utilisées dans les expériences (en min).

Style/temps(min)	apprentissage	Test
classique	1.77	11.6146
Standard	0.5	5.4354
Jazz	0.5	5.5942
Rock	0.04	0.3517
Metal	0.5	2.4975
Western	0.9	3.6269
Country	0.54	5.4231
μ_{Music}	4.75	34.5434
BluesChanté	0.5	1.7808
CountryChanté	1.5	3.5712
RapChanté	1.5	2.4879
RockChanté	1.0	2.4985
Films	0.62	9.4087
VariéesChanté	1.16	8.7456
Reggae	0.33	1.2098
$\mu_{MusiqueChanté}$	6.61	29.702
μ_{Parole}	6.00	74.2169
μ_{total}	17.3	138.4628

TAB. 4.5 - Résultats obtenus à partir de deux modèles (parole et musique) avec 8 GMM sans normalisation des paramètres.

Style	100ms	200ms	temps (min)
classique	93.35	94.52	11.6146
Standard	96.84	97.16	5.4354
Jazz	91.66	94.13	5.5942
Rock	91.31	93.13	0.3517
Metal	98.37	99.33	2.4975
Western	87.65	88.92	3.6269
Country	80.7660	80.8980	5.4231
$\mu_{Musique}$	91.42	92.58	34.5434
BluesChanté	54.80	52.33	1.7808
CountryChanté	78.19	79.33	3.5712
RapChanté	71.80	72.04	2.4879
RockChanté	61.84	55.18	2.4985
Films	80.07	80.17	9.4087
VariéesChanté	87.20	88.23	8.7456
Reggae	63.08	57.89	1.2098
$\mu_{MusiqueChanté}$	71.00	69.31	29.702
μ_{Parole}	92.46	96.62	74.2169
μ_{total}	80.78	80.79	138.4628

TAB. 4.6 - Résultats obtenus à partir de deux modèles (parole et musique) avec 8 GMM avec normalisation des paramètres.

Style	100ms	200ms	temps (min)
classique	86.03	88.51	11.6146
Standard	90.03	91.48	5.4354
Jazz	85.38	88.85	5.5942
Rock	84.73	85.76	0.3517
Metal	93.98	97.22	2.4975
Western	69.19	67.51	3.6269
Country	75.18	75.40	5.4231
$\mu_{Musique}$	83.50	84.96	34.5434
BluesChanté	55.75	51.34	1.7808
CountryChanté	61.29	57.72	3.5712
RapChanté	42.54	32.62	2.4879
RockChanté	37.06	25.88	2.4985
Films	74.23	74.02	9.4087
VariéesChanté	74.24	74.38	8.7456
Reggae	36.82	25.86	1.2098
$\mu_{MusiqueChanté}$	54.56	48.83	29.702
μ_{Parole}	89.96	96.52	74.2169
μ_{total}	70.67	69.04	138.4628

TAB. 4.7 - Résultats obtenus à partir de trois modèles (parole, musique et musique chantée) avec 8 GMM et sans normalisation des paramètres.

Style	100ms	200ms
classique	97.7981	98.5010
Standard	98.9667	98.9867
Jazz	98.5362	99.4512
Rock	98.0183	98.8867
Metal	99.2850	99.6275
Western	95.5200	96.7467
Country	95.4633	97.6833
$\mu_{Musique}$	97.65	98.55
BluesChanté	72.0688	71.3687
CountryChanté	94.3700	96.1160
RapChanté	95.0575	98.1100
RockChanté	85.0425	91.2675
Films	88.3111	88.3722
VariéesChanté	95.0686	96.1957
Reggae	92.2740	94.0100
$\mu_{MusiqueChanté}$	88.88	90.77
μ_{Parole}	94.2813	97.9393
μ_{total}	91.91	93.32

Chapitre 5

Conclusion

Le travail de cette thèse, a consisté principalement à une étude portant sur la reconnaissance du locuteur et sur la discrimination de la Parole/Musique dans le contexte de conférences téléphoniques.

Dans le chapitre 2, on s'est intéressé à la problématique en identification du locuteur en ensemble fermé. Particulièrement, on a mis l'accent sur la recherche de nouveaux paramètres censés être plus robustes aux conditions téléphoniques. On a donc proposé plusieurs paramètres potentiels extraits à partir de l'enveloppe (MA) et de la fréquence instantanée (FI) calculés à la sortie d'un banc de filtres cochléaires. En effet, la FI et la MA sont supposées tenir compte d'un phénomène non linéaire lié à l'interaction entre la source et le conduit vocal pendant le processus de production. Particulièrement, on a fait en sorte que tous les paramètres proposés soient calculés de façon synchrone à la glotte. Cela nous permet d'une part de travailler avec une fenêtre d'analyse dynamique et d'autre part de caractériser le locuteur à partir des variations sur de courtes durées. Seuls les segments voisés ont été considérés dans nos traitements. Ils ont été automatiquement localisés via un algorithme dédié à la détection de la fréquence glottale. Ce dernier a été renforcé par un traitement supplémentaire

qui exploite les pics des enveloppes (AM) pour avoir plus de précision et raffiner l'estimation sur la durée glottale.

D'après les résultats, ce type d'analyse apparaît prometteur et potentiel. Les performances obtenues sont comparables à celles obtenues par le système de référence utilisant les coefficients standards $MFCC$. Pourtant, on sait que les composantes $MFCC$ sont bien décorrélées entre elles, ce qui n'est pas le cas pour nos paramètres. On recommande donc d'envisager ultérieurement des modifications qui se rapportent principalement à la recherche d'une transformation mathématique adaptée pour mieux décorréler les composantes des vecteurs paramètres proposés. Tout laisse croire, qu'un traitement qui procède par sélection de canaux ne pourra que contribuer à rehausser les performances. Par sélection de canaux, on veut dire de ne prendre dans les décisions que celles provenant des canaux non fortement bruités. À retenir aussi, que dans notre traitement on a associé une pondération uniforme quant à la contribution des canaux. Cependant, cette pondération devra être proportionnelle à l'énergie transportée par chaque canal. Une démarche qui se retrouve justifiée pour être explorée ultérieurement. L'utilisation d'un modèle à architecture hybride ou parallèle où chaque sous-modèle exploite un espace de paramètres différents a été également étudiée dans un cadre plutôt exploratoire. En effet, on a expérimenté un système qui prend ces décisions en se basant conjointement sur les paramètres proposés et sur es coefficients MFCC. Avec une telle stratégie, on a obtenu les meilleurs performances. Ce gain peut être expliqué par une notion de complémentarité entre les paramètres AM, FM et MFCC.

Toutefois, les paramètres proposés ne représentent pour l'instant que l'ouverture d'une nouvelle porte en caractérisation du locuteur. On est loin encore d'une solution

optimale et on est conscient que d'autres travaux sont nécessaires pour mener à terme les perspectives de cette approche.

Une autre expérience visant à comparer les performances entre machines et auditeurs naïfs, a montré que la variabilité des combinés téléphoniques affecte grandement la perception humaine. En effet, les scores obtenus par la machine devance considérablement, ceux obtenus par les auditeurs naïfs.

Motivé par le fait que les coefficients MFCC se basent uniquement sur les propriétés résonantes du conduit vocal pendant le processus de reconnaissance, on a introduit (chapitre 3) une nouvelle approche en modélisation. Cette dernière consiste à combiner via un modèle probabiliste la dépendance liée au couplage du conduit vocal et de la glotte. Autrement dit, on tient compte de la dépendance temporelle de la source excitatrice et du conduit résonnant. Le système proposé a été comparé à deux systèmes opérant respectivement sur les segments voisés et non voisés. Les expériences ont été réalisées avec les quarante-cinq (45) locuteurs de la base Spidre.

Tous les modèles proposés pour caractériser la contribution du conduit vocal sont composés d'un mélange de 32 Gaussiennes pondérées. Les paramètres utilisés sont les 12 coefficients statiques des MFCC. Le nombre de modèles (ou intervalles de fréquence) pour système proposé a été fixé à quatre. On a obtenu un gain en performance significatif quand la durée du test est inférieure à 500 ms. Lorsqu'on augmente la durée, les performances entre les différents systèmes ne demeurent pas en général assez importantes.

En effet, plusieurs restrictions et hypothèses ont été prises pour la réalisation du système. Parmi autre, on a supposé que l'estimateur du fondamental est fiable. On a

aussi utilisé la même subdivision sur l'espace des paramètres pour l'ensemble de locuteurs, alors qu'en réalité il fallait procéder par une subdivision typique à chaque locuteur puisque la distribution de la fréquence varie selon le locuteur. Particulièrement, il sera intéressant d'analyser l'impact des espaces probabilistes des événements, associés respectivement à la source et au conduit vocal, sur la prise des décisions. On sera ainsi capable de savoir si on doit procéder par une normalisation sur l'espace des paramètres (source et conduit vocal). Finalement, on tient à recommander pour chaque locuteur l'incorporation d'un autre modèle qui caractérisera les segments de parole non-voisés.

Les données de musique incluant la musique chantée ont été obtenues à partir de différents sites Internet de diffusion. On y trouve toutes sortes de styles dont la largeur de bande a été limitée et dont la variabilité des combinés a été également introduite. Ceci nous place dans des conditions plus réalistes simulant l'utilisation dans un contexte téléphonique. Les données de la parole sont extraites de la base de données SPIDRE. On y trouve 45 locuteurs avec 4 conversations pour chaque locuteur enregistrées via trois combinés différents.

Dans le chapitre 4 cette fois,on s'est intéressé particulièrement à la problématique reliée à la discrimination Parole/Musique. Deux approches ont été proposées et expérimentées. Comme première approche, on a proposé des paramètres dérivés des coefficients MFCC après avoir appliqué certaines transformations d'ordre statistiques et aussi empiriques en s'inspirant des analyses graphiques. L'impact de ces transformations réside dans le fait de projeter l'espace des coefficients MFCC à N dimensions à un nouveau espace de dimension 1 où l'entrelacement entre les classes de musique et parole est considérablement réduite. Le processus de discrimination se base simplement sur

un seuil fixe pour effectuer la classification. En effet, le seuil a été déterminé selon la règle de Bayes qui consiste à trouver les frontières inter-classes en minimisant l'erreur totale de classification. Les transformations proposées restent toutefois implicitement liées à la taille de la fenêtre d'analyse et au nombre de trames considérées dans le traitement. Avec une durée de test de 100 ms, on a obtenu en moyenne des scores de 90% (parole et musique) et 84% (parole, musique et musique chanté). Ces scores sont en général comparables à ceux qu'on trouve publiés dans la littérature scientifique. Cependant dans notre cas, on utilise une durée de test 9 fois moins longue. L'avantage de cette technique repose sur sa simplicité, son temps de calcul très réduit (important pour les applications en temps réel) et la classification à base de seuil (pas d'entraînement exhaustif). On peut aussi songer à faire rehausser les performances obtenues en faisant, par exemple, incorporer d'autres modules comme le détecteur de silence, le calcul d'énergie et/ou le calcul du taux de passage par zéro. Cependant, avec l'incorporation de tels modules, on fera perdre les aspects de simplicité et de rapidité, considérées comme les points clefs de cette approche.

Comme deuxième approche, on a utilisé une modélisation paramètrique à base de GMM. On a pris deux modèles à 8 GMM respectivement pour la parole et la musique. On a montré que lorsque la normalisation du canal est appliquée (la soustraction du vecteur moyen de tous les vecteurs paramètres estimés), les performances chutent de 10% pour la musique et de 20% pour la musique chantée. Normalement, on utilise la normalisation du canal pour supprimer l'effet du canal dont les caractéristiques spectrales sont considérées stables. Or, la musique se caractérise aussi par une structure harmonique stable sur de longue durée qui se retrouve supprimer en grande partie par la normalisation du canal. On recommande donc de procéder sans effectuer la

normalisation du canal dans l'optique de réaliser une meilleure discrimination Parole/Musique. On rappelle que les données tests ici, sont enregistrées à partir d'un combiné différent de ceux présentés lors de la phase d'apprentissage.

Ensuite, on a montré qu'avec l'utilisation d'une modélisation différenciée basée sur 3 modèles GMM, entraînés respectivement sur la parole, la musique et la musique chantée, les meilleurs scores ont été obtenus. Avec une durée de test de 200 ms, les scores sont de 97.93% pour la parole, 98.55% pour la musique et 90.77% pour la musique chantée. Le système de référence utilisait seulement deux modèles GMM pour la musique et pour la parole. Les scores obtenus pour ce dernier sont de 96.62% pour la parole, 92.58% pour la musique et 69.31% pour la musique chantée. Les scores obtenus pour les durées de tests de 100 ms et 200 ms, sont très proches, on donc suppose que cette tendance peut se maintenir si on réduira la durée test à 50 ms (surtout pour les femmes).

Bibliographie

[1] Aarts R. M. and Dekkers R. T., A Real-time Speech-Music Discriminator, In *J. Audio Eng. Soc.*, Vol. 47, No. 9, September 1999.

[2] Walter D. Andrews, Mary A. Kohler, Joseph P. Campbell and John J. Godfrey, Phonetic, idiolectal and acoustic speaker recognition, In *A Speaker Odyssey, The Speaker Recognition Workshop*, pp. 55-63, June 2001, Crete, Greece.

[3] Arcienega Mijail and Drygajlo Andrzej, Pitch-Dependent GMMs for Text-Independent Speaker Recognition Systems, Proceedings of 7th European Conference on Speech Communication and Technology, In *Eurospeech*, Aalborg, Denmark, Sept. 3-7, 2001, pp 2821-2824.

[4] Artières T. and Gallinari P., Multi-state predictive neural networks for text-independent speaker recognition, In *Eurospeech*, pp. 633-636, September 1995, Madrid (Spain).

[5] Atal B. S., Automatic speaker recognition based on pitch contours, In *The Journal of the Acoustical Society of America*, pp. 1687-1697, Vol. 52, 1972.

[6] Atal B. S., Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, In *The Journal of the Acoustical Society of America*, pp. 1304-1312, Vol. 55, No. 4, June 1974.

[7] Atal B. S., Automatic recognition of speakers from their voices, In *Proc. IEEE*, pp. 460-475, Vol. 64, 1976.

[8] Auckenthaler Roland and Mason John S., Gaussian selection aplied to text-independent speaker verification, In *A Speaker Odyssey, The Speaker Recognition Workshop*, pp. 83-86, June 2001, Crete-Greece.

[9] Beaufays F. and Weintraub M., Model transformation for robust speaker recognition from telephone data, In *IEEE ICASSP*, pp. 1063-1066, 1997.

[10] Bennani Y., A Connectionist Approach for Automatic Speaker Identification, In *IEEE ICASSP*, pp. 265-268, 1990.

[11] Bennani Y., Approche Connexionniste pour la Reconnaissance Automatique du Locuteur : Modelisation et Identification, In *thesis : Paris XI*, 1992.

[12] Bennani Y., Speaker Identification Through a Modular Connectionist Architecture : Evaluation on the TIMIT database, In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 607-610, 1992.

[13] Bennani Y., Probabilistic Cooperation of the Connectionist Expert Modules : Validation on a speaker Identification Task, In *IEEE ICASSP*, pp. I-541-544, 1993.

[14] Besacier Laurent, Un Modèle Parallèle pour la Reconnaissance Automatique du Locuteur, In *Thèse : soutenue le 17 Avril 1998*, Université d'Avignon et des Pays Vaucluse, 1998.

[15] Bimbot F., Chollet G. and Paoloni A., Assessement Methodology for Speaker Identification and Verification, In *ESCA Workshop on Speaker Recognition, Identification, and Verification*, pp. 75-83, 1994.

[16] Bimbot F., Magrin-Chagnolleau Y., Mathan L. , Second-order statistical methods for text-independent speaker identification, In *Speech Communication*, Vol.17 (1-2), August, 1995.

[17] Bojan I., Kacic Z. and Horvat B., A study of harmonic features for the speaker recognition, In *Speech Communication*, Vol. 22, pp. 387-401, 1997.

[18] Bonastre J.F. and Meloni H., Inter- and Intra-Speaker Variability of French Phonemes ; Advantages of an Explicit Knowledge Based Approach, In *Esca Workshop on Speaker Recognition, Identification, and Verification*, pp. 157-160, 1994.

[19] Bonneau H. and Gauvain J., Vector quantization for speaker adaptation, In *IEEE ICASSP*, pp. 1437-1441, 1987.

[20] Bovik Alan C., Maragos Petros and Quatieri Thomas F., AM-FM Energy detection and separation in noise using multiband energy operators, In *IEEE Trans. on Signal Processing*, Vol. 41, No. 12, pp. 3245-3265, December 1993.

[21] Buck J., Burton D. and Shore J., Text-dependent speaker recognition using Vector Quantization, In *IEEE ICASSP*, pp. 391-394, 1985.

[22] Tubach J. P., Calliope : La parole et son traitement automatique, Collection Technique et Scientifique des télécommunications, MASSON, 1989.

[23] Campbell Joseph P., Speaker Recognition : A tutorial, In *Proceedings of the IEEE*, pp. 436-1462, Vol. 85, No. 9, 1997.

[24] Carey M. J., Parris E. S. and H. Lloyd-Thomas, A Comparison of Features for Speech, Music Discrimination, In *IEEE ICASSP* , Mars 1999, Phoenix, AZ.

[25] Delgutte B., Representation of speech-like sounds in the discharge patterns of auditory nerve fibers, In *JASA* , Vol. 68 ,

[26] Delgutte B. and Kiang N. Y., Speech coding in the auditory nerve :V. Vowels in background noise, In *Journal of the Acoustical Society of America*, Vol. 75, pp. 908-918, 1984.

[27] Demars C. and Gauvain J.L, Application des L-distributions à la reconnaissance de la parole, Dans *journal :14 ème JEP*, pp. 283-287, Juin 1985, Paris (ftp ://ftp.limsi.fr/Individu/chrd/RPBDM2000v1.ps.gz).

[28] Demars C., Représentation bidimensionnelles d'un signal de parole éléments de monographie, In *LIMSI, Orsay cedex,Paris, France*, 25 Novembre 1998.

[29] Doddington G., Speaker recognition-identifying people from their voices. In *Proc. IEEE*, pp. 1651-1664, Vol. 73 (11), 1985.

[30] Dunn R.B., Quarteri T.F., Reynolds D.A. and Campbell J.P, Speaker Recognition from Coded Speech in Matched and Mismatched Conditions, In *A speaker Odyssey, The Speaker Recognition Workshop* , pp. 115-120, June 2001, Crete, Greece.

[31] El-Maleh K., Klein M., Petrucci G. and Kabal P., Speech/Music Discrimination for multimedia applications , In *", Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (Istanbul),* pp. 2445-2448, June 2000.

[32] Ezzaidi Hassan and Rouat Jean, Speech, Music and Songs discrimination for real time Applications, In *International Conference on Spoken Language Processing (ICSLP)*, pp. 2013-2016,, 2002, Denver, USA.

[33] Ezzaidi Hassan, Rouat Jean and, Combining pitch and MFCC for speaker identification systems, In *The Speaker Recognition Workshop*, pp. 207-212, June 2001, Crete, Greece.

[34] Ezzaidi Hassan, Rouat Jean and O'Shaughnessy Douglas, Toward combining pitch and MFCC for speaker identification systems, In *Eurospeech*, 2001.

[35] Ezzaidi Hassan and Rouat Jean, Speaker Identification by Computer and Human Evaluated on The Spidre Corpus, In *Canadian Acoustics Journal*, Vol. 28, No. 3, pp. 156-157, 2000.

[36] Ezzaidi Hassan and Rouat Jean, Comparison of MFCC and Pitch Synchronous AM,FM Parameters for Speaker Identification, In *International Conference on Spoken Language Processing (ICSLP)*, CHINE, 2000.

[37] Foldvari R., Pitch frequency estimation based on instantaneous amplitude and frequency functions, In *IEEE ICASSP*, Annoncée dans IEEE-ASSP, October 1989.

[38] Furui S., An analysis of long term variation of feature parameters of speech and its application to talker recognition, In *In Trans. IECE, 57-A*, Vol. 12, pp. 880-887, 1974.

[39] Furui S., Cepstral analysis technique for automatic speaker verification, In *IEEE Trans. Acoust. Speech Signal Processing*, Vol. 29, No. 2, pp. 254-272, 1981.

[40] Furui S., An overview of speaker recognition technology, In *Workshop on Automatic Speaker Recognition and Verification*, pp. 1-9, April, 1994, Martigny, Switzerland.

[41] 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.

[42] Gauvain J.L, Lamel L.F and Prouts B., Experiments with Speaker Verification Over the Telephone, In *EUROSPEECH*, pp. 651-654, 1995.

[43] Geisler C. Daniel, From Sound to Synapse : physiology of the mammalian ear, In *Oxford* , 1998.

[44] Gish H., Karnofsky K., Krasner M., Roucos S., Schartz R. and Wolf J., Investigation of Text-Tndependent Speaker Identification over Telephone Chanels, In *IEEE ICASSP*, pp. 379-382, 1985.

[45] Gish H., Kraner M., Russel W. and Wolf J., Methods and experiments for text-independent speaker recognition over the telephone line, In *IEEE ICASSP*, pp. 865-868, 1986.

[46] Gish H., Robust Discrimination in Automatic Speaker Identification, In *IEEE ICASSP*, pp. 289-292, 1990.

[47] Gish H. and Schmidt M., Text-Independent Speaker Identification, In *IEEE Signal Processing Magazine*, pp. 18-32, 1985.

[48] Glasberg B.R. and Moore B.C.J., Derivation of auditory filter shapes from notched-noise data, In *Hearing Research*, Vol. 47, pp. 103-138, 1990.

[49] Grenier Y., Utilisation de la Prédiction Lineaire en Reconnaissance et Adaptation au Locuteur, In *XIth JEP*, pp. 163, 1980.

[50] Guia P. and Pisani R., Intra- and Inter-Speaker Variability, In *ESCA Workshop on Speaker Recognition, Identification, and Verification*, pp. 123-126, 1994.

[51] Hansen Eric G., Raymond E. Slyh and Timothy R. Anderson, Formant Fo Features for speaker verification, In *A speaker Odyssey, The Speaker Recognition Workshop*, pp. 25-29, June 2001, Crete, Greece.

[52] He J., Liu L. and Palm G., Speaker identification using hybrid LVQ-SLP networks, In *Proc. IEEE ICNN, Vol.4*, pp. 2051-2055, 1995, Perth, Australia.

[53] Hermansky H., RASTA-PLP Speech Analysis Technique, In *IEEE ICASSP*, pp. I.121-I.124, 1992.

[54] Hermansky H., Morgan N., RASTA Processing of Speech, In *IEEE Trans. On Speech and Audio Processing*, pp. 578-589, Vol. 2, No. 4, 1994.

[55] Higgins A. and Bahler L., Text-independent speaker verification by discriminator counting, In *IEEE ICASSP*, pp. 405-408, 1991.

[56] Homayounpour M.M. and Choller G., A comparison of some relevant parametric representations for speaker verification, In *Workshop on Automatic Speaker Recognition and Verification*, pp. 185-188, April, 1994, Martigny, Switzerland.

[57] Jankowski C.R, Quatieri T.F. and Reynolds D.A., Measuring fine structure in speech : Application to speaker identification, In *IEEE ICASSP*, pp. 325-328, 1995.

[58] Johnson S. E., Who spoke when?- Automatic segmentaion and clustering for determining speaker turns, In *Proc. Eurospeech*, September 1999, Budapest, Hungary.

[59] Jonathan Phillips P., Martin Alvin, Wilson C. L. and Przybocki Mark, An introduction to evaluating biometric systems, In *IEEE Computer Society*, February, 2000.

[60] Kharroubi Jamal, Petrovska-Delacrétaz Dijana and Chollet Gérard, Text independent speaker verification using support vector machines, In *A speaker Odyssey, The Speaker Recognition Workshop*, pp. 51-54, June 2001, Crete, Greece.

[61] Kajarekar Sachin S. and Hermansky Hynek, Speaker Verification Based on Broad Phonetic Categories, In *A speaker Odyssey, The Speaker Recognition Workshop*, pp. 201-205, June 2001, Crete, Greece.

[62] Kyung Y. and Lee H., Text Speaker Recognition using micro-prosody, In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 1998.

[63] Ludovic F., Sénac C., Vallès-Parlangeau N. and André-Obrecht R., Indexation de la bande sonore :les composantes Parole/Musique, In *in Journée de jeunes doctorants*, 2000.

[64] Maes Stéphane H., LPC Poles Traker for Music/Speech/Noise Segmentation and Music Cancellation, In *Eurospeech*, 1997.

[65] Magrin-Chagnolleau I., Bonastre J.F. and Bimbot F., Effect of Utterance Duration and Phonetic Content on Speaker Identification using Second Order Statistical Methods, In *EUROSPEECH*, pp. 337-340, 1995.

[66] Malayath N. , Hermansky H., Kain A. and Carlson R., Speaker-Independent Feature Extraction by Oriented Principal Component Analysis, In *EUROSPEECH*, 1997.

[67] Maragos Petros, Kaiser James F. and Quatieri Thomas F., Energy separation in signal modulations with application to speech analysis, In *IEEE Trans. on Signal Processing*, pp. 3024-3051, Vol. 41, No. 10, October 1993.

[68] Markel D., Oshika B. T. and Gray A. H., Long-Term Feature Averaging for Speaker Recognition, In *IEEE Trans. on ASSP*, pp. 330-337, Vol. 25, No. 4, 1977.

[69] Markel J. D. and Davis S. B, Text-independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base, In *IEEE Trans. on ASSP*, pp. 74-82, Vol. 27, No. 1, 1979.

[70] Matsui T. and Furui S., Text-Independent Speaker Recognition Using Vocal Tract and Pitch Informations, In *ICSLP*, pp. 137-140, 1990.

[71] Matsui T. and Furui S., A Text-Independent Speaker Recognition Method Robust Against Utterance Variations, In *IEEE ICASSP*, pp. 377-380, 1991.

[72] Matsui T. and Furui S., A comparison of text-independent speaker recognition methods using VQ-distorsion and discret continuous HMMs, In *Proc. ICASSP 92*, pp. 157-160, 1992, San-Francisco, USA.

[73] Matsui T. and Furui S., Speaker Adaptation of Tied-Mixture-Based Phoneme Models for Text-Prompted Speaker Recognition, In *IEEE ICASSP*, pp. I-125.128, 1994.

[74] Meignier Sylvain, Bonastre Jean-François and Igounet Stéphane, E-HMM appoach for learning and adapting sound models for speaker indexing, In *A speaker Odyssey, The Speaker Recognition Workshop*, pp. 175-180, June 2001, Crete, Greece.

[75] Murthy H. A., Beaufays F., Heck L. P. and Weintraub M., Robust Text-Independent Speaker Identification over Telephone Channels, In *IEEE Transactions on Speech and Audio Processing*, pp. 554-568, Vol. 7, No. 5, september 1999.

[76] Ong S., Moody M.P and Sridharan S., Confidence analysis for speaker identification : The effectivenes of various features, In *Workshop on Automatic Speaker Recognition and Verification*, pp. 91-94, April, 1994, Martigny, Switzerland.

[77] Openshaw J., Sun Z. and Mason J., A comparison of composite features under degraded speech in speaker recognition, In *IEEE ICASSP*, pp. 371-374, 1993.

[78] Potamianos Alexandros and Maragos Petros, Time-frequency distributions for automatic speech recognition, In *IEEE Trans. On Speech Audio Processing*, Vol. 9, No. 3, pp. 196-200, March 2001.

[79] Patterson R.D., Auditory filter shapes derived with noise stimuli, In *JASA*, pp. 640-654, Vol. 59, No. 3, 1976.

[80] Phillips P. Jonathan, Martin Alvin, Wilson C. L. and Przybocki Mark, An introduction to evaluating biometric systems, In *IEEE Computer Society*, February 2000.

[81] Plumpe M.D, Quatieri T.F., and Reynolds D.A, Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification, In *IEEE Transactions on Speech and Audio*, 1999.

[82] Popper A. N. and Fay R., The Mammalian Auditory Pathway : Neurophysiology, In *Springer–Verlag*, 1992.

[83] Poritz A.B., Linear Predictive Hidden Markov Models and the Speech Signal, In *IEEE ICASSP*, pp. 1291-1294, 1982.

[84] Potamianos Alexandros and Maragos Petros, Speech formant frequency and bandwidth tracking using multiband energy demodulation, In *J. Acoust. Soc. Am. (JASA)*, Vol. 99 , No. 6 , pp. 3795-3805, June 1996.

[85] Pruzansky S., Pattern matching procedures for automatic talker recognition, In *The Journal of the Acoustical Society of America*, pp. 354, Vol. 35, No. 3, 1963.

[86] Qiu L., Yang H. and Koh S.N., Fundamental frequency determination based on instantaneous frequency estimation, *Signal Processing*, pp. 233-241, Vol. 44, June 1995.

[87] Rang D. Zilca, Using second order statistics for text independent speaker verification, In *A speaker Odyssey, The Speaker Recognition Workshop*, June 2001, Crete, Greece.

[88] Reynolds Douglas A., Experimental evaluation of features for robust speaker identification, In *IEEE Trans. on ASSP*, pp. 639-643, Vol. 4, No. 2, 1994.

[89] Reynolds Douglas A., A gaussian mixture modeling approach to text independent speaker identification, *Thesis, Georgia Institute of Technology*, August 1992.

[90] Reynolds Douglas A. and Rose Richard C., Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, In *IEEE Trans. on SAP*, pp. 72-83, Vol. 3, No. 1, 1995.

[91] Reynolds Douglas A., The effects of handset variability on speaker recognition performance : Experiments on the Switchboard corpus, In *IEEE ICASSP*, 1996.

[92] Rosen S., Structure temporelle de la parole, aspects acoustiques, auditifs, linguistiques, In *Congress on Auditory Processing of Complex Sounds*, pp. 4-5, Vol. , December 1991, London.

[93] Rosenberg A. and Soong F., Evaluation of a Vector Quantization talker recognition system in text independent and text dependent modes, In *Computer Speech and Language* , Vol. 2 , No. 3/4 , 1988.

[94] Rouat J., Nonlinear operators for speech analysis, chapter. 36, publisher :J. Wiley and Sons, Visual representations of speech signals, editor. M. Cooke and S. Beet and M. Crawford, pp. 335 340, 1993.

[95] Rudazi L., and Zahorian S. A., Text-Independent Talker Identification with Neural Network, In *IEEE ICASSP*, pp. 389-392, 1991.

[96] Sambur M.R., Selection of Acoustic Features for Speaker Identification, In *IEEE Trans. on ASSP*, pp. 176-182, Vol. 23, No. 2, 1975.

[97] Sambur M.R., Speaker Recognition Using Orthogonal Linear Prediction, In *IEEE Trans. on ASSP*, pp. 283-289, Vol. 24, 1976.

[98] Samouelian A., Robert-Ribes J. and Plumpe M., Speech, Silence, Music and Noise Classification of TV Broadcast Material, In *ICSLP'98*, 1998, Sydney (AU).

[99] Saunders John, Real-time discrimination of broadcast Speech/Music, In *IEEE ICASSP*, pp. 993-996, 1996.

[100] Scheirer E. and Stanley M., Construction and evaluation of a robust multifeature speech/music discriminator, In *IEEEICASSP'97* , Vol. II, pp. 1331-1334, 1997, Munich.

[101] Sharath Pankanti, Ruud M. Boll and Anil Jain, Biometrics : The future of identifiation, In *IEEE Computer Society*, pp. 46-49, February 2000.

[102] O'Shaughnessy Douglas, Speech Communications : Human and Machine, IEEE Press, 2000.

[103] Solomonoff A., Mielke A., Schmidt M. and Herbert G., Clustering Speakers by their Voices, In *IEEE ICASSP* , pp. 757-760, 1998, Seattle.

[104] Kemal Sönmez, Larry Heck, Mitchel Weintraub and Elisabeth Shriberg, A lognormal tied mixture model of pitch for prosody-based speaker recognition, *In Proc. of Eurospeech*, pp. 1391-1394, 1997.

[105] Kemal Sönmez, Elisabeth Shriberg, Laryy Heck and Mitchel Weintraub, Modeling dynamic prosodic variation for speaker verification, In *Proc. of International Conference on Spoken Language Processing*, pp. 3189-3192, 1998.

[106] Soong F. K. S., Rosenberg A. E., Rabiner, L. R. and Juang B. H., A Vector Quantization Approach to Speaker Recognition, In *IEEE ICASSP*, pp. 387-390, 1985.

[107] Spina M. S. and Zue V.W., Automatic Transcription of General Audio Data :Preliminary Analyses, In *Proc. ICSLP-96*, pp. 594-597, 1996.

[108] Tancerel L., Ragot S., and Lefebvre R., Speech/Music Discrimination for universal audio coding, In *20th Biennal Symposium on Communications*, pp. 28-31, 2000, Queen's University, Kingston, Canada.

[109] Teager H.M., and Teager S.M., Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract, In *Speech Production and Speech Modelling, edited*

by W. J. Hardcastle and A. Marchal, NATO Advanced Study Insitute Series D, Bonas, pp. 241-261, France, July 1989.

[110] Teager H.M., Some Observations on Oral Air Flow During Phonation, In *IEEE Tran. Acoust., Speech, Signal Processing, ASSP-28, No. 5*, pp. 599-601, October 1980.

[111] Teager H.M., Some Useful Properties of Teager's Energy Operators , In *IEEE ICASSP*, Vol. 3 , pp. 149-152, New York, 1993.

[112] Teager H.M. and Teager S.M. A phenomenological Model for Vowel Production in the Vocal Tract, In *R.G. Dabiloff (ed.) Speech Sciences : Recent Advances*, pp. 73-109, College-Hill Press, San Diego, CA, October 1983.

[113] Thevenaz Ph. and Hugli H, Usefulness of the LPC-residue in text-independent speaker verification, In *Speech Communication*, pp. 145-158, Vol. 17(1-2), August 1995.

[114] Tishby N. Z., Information theoretic factorization of speaker and language in Hidden Markov Models with application to speaker recognition, In *IEEE ICASSP*, pp. 87-90, 1988.

[115] Tishby N.Z., On the application of mixture AR HMMs to text-independent speaker recognition, in *IEEE Transactions on Signal Processing*, Vol. 39, pp. 563-570, March 1991.

[116] Wenndt S. and Shamsunder S., Bispectrum features for robust speaker identification, In *Proc. ICASSP 97*, pp. 1095-1098, 1997, Munich, Germany.

[117] Williams G. and Ellis D. P.W., Speech/Music Discrimination Based on posterior probability Features, in *Eurospeech'99*, Budapest, 1999.

[118] Wolf J., Efficient acoustic parameters for speaker recognition, In *The Journal of the Acoustical Society of America*, pp. 2044-2056, No. 51(6), 1972.

[119] Xu L., Oglesby J. and Mason J., The optimization of perceptually-based features for speaker identification, In *IEEE ICASSP*, pp. 520-523, 1989.

[120] Xu L. and Mason J., Optimization of perceptually-based spectral transforms in speaker identification, In *EUROSPEECH*, pp. 439-442, 1991.

[121] Yanguas L.R., Quartieri T.F and Goodman F., Implications of glottal source for speaker and dialect identification, In *IEEE ICASSP*, 1999.

[122] Yong Chun Liu, Un détecteur perceptif de la hauteur tonale pour la parole téléphonique, *Mémoire présenté à UQAC, Québec, Canada*, 1992.

[123] Zhang Tong and Jay Kuo C.-C, Content-based Classification and Retrieval of Audio, In *SPIE's 43rd Annual Meeting - Conference on Advanced Signal Processing Algorithms, Architectures, and Implementations VIII, SPIE*, pp. 432-443, Vol. 3461, San Diego, July 1998.

SPEECH, MUSIC AND SONGS DISCRIMINATION IN THE CONTEXT OF HANDSETS VARIABILITY

Hassan Ezzaidi and Jean Rouat

ERMETIS, DSA, Université du Québec à Chicoutimi,
555 boul. de l'Université, Chicoutimi, Québec, Canada G7H 2B1.
http://wwwdsa.uqac.uquebec.ca/ermetis

1. ABSTRACT

The problem of speech, music and music with songs discrimination in telephony with handsets variability is addressed in this paper. Two systems are proposed. The first system uses three Gaussian Mixture Models (GMM) for speech, music and songs respectively. Each GMM comprises 8 Gaussians trained on very short sessions. Twenty six speakers (13 females, 13 males) have been randomly chosen from the SPIDRE corpus. The music were obtained from a large set of data and comprises various styles. For 138 minutes of testing time, a speech discrimination score of 97.9% is obtained when no channel normalization is used. These performance are obtained for a relatively short analysis frame (32ms sliding window, buffering of 100 ms). When using channel normalization, an important score reduction (on the order of 10 to 20%) is observed. The second system has been designed for applications requiring shorter processing times along with shorter training sessions. It is based on an empirical transformation of the ΔMFCC that enhances the dynamical evolution of tonality. It yields in average an acceptable discrimination rate of 90% (speech-/music) and 84% (speech, music and songs with music).

2. INTRODUCTION

As the digital storage and processing of audio signals are increasing and become popular, speech and music discrimination systems are crucial to extend the functionality of various information and communication systems. In Multimedia applications, such systems can be useful to achieve automatic classification, indexation, archiving and retrieving of information from large multimedia databases [5]. Speech and music discrimination can also play a significant role in speaker or speech recognition systems, by rejecting non speech segments. The new generation of low bit rate coders and compression technologies need an estimation of the signal nature in order to achieve a better compression. Therefore, a fast and efficient speech/music discriminator is crucial for that type of coders as the performance is strongly related to the accuracy of the speech/music discriminator [4].

To retrieve and achieve a good discrimination, the dynamical time evolution of the vocal tract, during the speech production (vowels, consonants, coarticulation, etc.), is one salient property that has been strongly exploited in many ways in the literature. In brief, the vocal signal is characterized by a formantic and dynamical structure contrary to the musical signal that is rather characterized by an harmonic and regular structure over longer durations. Generally, the proposed techniques are simple in order to be applicable to a vast form of music style and to be independent of speaker and speech for most cases. Saunders [6] proposes four features derived principally from the zero crossing rate and the energy contour. A set of 13 features related to the amplitude, fine spectrum and signal frequency also are studied by Scheirer and Stanley [7]. Carey et *al.* [2] examine the discrimination achieved by several different features (filterbank energy, cepstre, pitch and zero-crossing) using common training and test sets and the same classifier. Samouelian and *al.* [5] perform the automatic labeling and classification of TV broadcast material into speech, music, silence and noise segments. El-Maleh and *al.* [3] suggest the use of line spectral frequencies (LSFs) and zero-crossing-based features for frame-level narrowband speech/music discrimination. Recently, Ajmera and *al.* [1] use a posterior probability based entropy and dynamism features that are integrated over time through a 2 state HMM. Generally, the authors use features that are estimated over a long time range (0.5 to 5 seconds). One exception to this, is the work by El-Maleh et *al.* [3], where duration tests as short as 20 ms are used. Their scores with such a short window length are still interesting and average to 82.5% for the speech segments and 79.2% in the music case.

In the present work we are interested in a system that performs indexation and retrieving of telephone speech information from a corpus that comprises conversations in alternance with music or songs. The indexation system performs first the speech/music discrimination and then, the speaker identification and speech recognition tasks. The design of the discrimination system is therefore constrained to the use of the MFCC vectors like the speaker/speech recognition systems. Furthermore, performance of the discrimination system should be independent on the recording conditions and more specifically of the telephone handset variability. We propose two discrimination systems. The first system uses a parametric multi-Gaussian modeling that comprises 8 mixtures of Gaussians per model. The second system is based on the comparison of distances between ΔMFCC derived features and a preset threshold.

3. DATABASE AND EVALUATION CRITERION

3.1. Database

The data were downloaded from various broadcasting Internet sites and from the university telephone switchboard. The audio quality ranges from narrow-bandwidth of AM radios (3 kHz) to high fidelity music (16 kHz bandwidth). The music files where then played trough a loudspeaker. A telephone handset was placed in front of the loudspeaker to transmit the music through the tele-

This work has been partially supported by NSERC and CSE.

phone network (digital in the university, analogue elsewhere). This operation limits the bandwidth to that of the telephone lines and introduces to a certain extend the channel effect and handset variability. The recordings were carried out during several sessions and comprises various types of music and songs.

Table 1. Music database with style description and durations respectively for the training (GMM system) and the testing sessions.

Styles for Music only	Training time (min)	Testing time (min)
Classique	1 77	11 6166
Standard	0 5	5 4354
Jazz	0 5	5 5962
Rock	0 06	0 1517
Metal	0 5	2 4975
Western	0 9	5 6269
Country	0 54	5 4231
Total for Music	4 75	46 5454
Styles for Music with Songs	Training time (min)	Testing time (min)
Blues	0 5	1 7920
Country	1 5	5 5712
Rap	1 5	2 4879
Rock	1 0	2 4985
Films	0 62	4 6887
Various	1 16	6 7456
Reggae	0 13	1 2899
Total for Music with Songs	6 61	29 762
Total for Speech	6 00	74 2169
Total Duration	17 1	150 46.25

Table 1 gives the different music style and time durations used during training and testing.

A subset of the speech SPIDRE–Swichboard Corpus is used to extract randomly speech data. Each speaker has 4 conversations originating from 3 different telephone handsets. 74 minutes of speech that comprises all conversations taken from ten males and ten females have been used for testing. In addition, we randomly choose one conversation of 3 women and 3 men from which we extracted 1s when training was necessary. The total speech duration in the training session is therefore of 6 minutes. The training speakers are not presented during the test. All the database is sampled at 8 Khz.

3.2. Discrimination score

The discrimination score $S = \frac{C}{T}$ is defined as the ratio of the number C of correctly classified frames over the number T of tested frames.The same score S is used for all the reported experiments.

4. ML SPEECH MUSIC DISCRIMINATION

In this section we design a Gaussian Mixture Models speech/music discriminator to study the influence of channel normalization – commonly used to improve speech recognition or identification rates in the context of telephone handsets variability – and the benefit of using a model for music with songs.

4.1. Perspectives and models

In a first set of experiments, one model for speech and another for music (with or without songs) are used. In that situation, – in the context of music with singer(s) – depending on their relative dominance the system will decide whether it is speech or music. A second set of experiments is designed where 3 models are used (speech, music only, music with voice or singer(s)).

4.2. Training session

For each style of music (Classic, Western, Pop, Rock, Jazz, Soft music, Instrumental and Rap), only a few seconds have been used to train the music GMM. Also, for each style of songs, a few seconds of signal, are extracted from the following categories: Pop, Rock, Blues and Rap to train the GMM songs. The speech model is trained on 6 minutes taken from 6 conversations of three men and three women and originating from different telephone handsets. Each conversation corresponds to one minute duration. None of the training data are used for the testing.

4.3. Testing session

A sliding window of 32ms length and 10ms shift is placed on the signal to estimate every 10ms the log–likelihood of the parameter observation for each model. Results are presented for averages over 100ms (10 frames) or 200ms (20 frames) of the log–likelihood. Finally, the model with the maximum averaged log–likelihood is retained.

4.4. Results

4.4.1. Influence of the channel normalization

Table 2. Parametric models and discrimination scores. 2 models: one GMM for Speech, one for Music **and** Songs; 3 models: one GMM for Speech, one for Music, one for Songs; CN: Channel Normalization. Sg in the music styles indicates presence of songs or singer with the music.

	2 models, no CN		2 models, CN		3 models, no CN	
Styles	100ms	200ms	100ms	200ms	100ms	200ms
classique	93.35	94.52	86.03	88.51	97.79	98.50
Standard	96.84	97.16	90.03	91.48	98.96	98.98
Jazz	91.66	94.13	85.38	88.85	98.53	99.45
Rock	91.31	93.13	84.73	85.76	98.01	98.88
Metal	98.37	99.33	93.98	97.22	99.28	99.62
Western	87.65	88.92	69.19	67.51	95.52	96.74
Country	80.76	80.89	75.18	75.40	95.46	97.68
Music	91.42	92.58	83.50	84.96	97.65	98.55
BluesSg	54.80	52.33	55.75	51.34	72.06	71.36
CountrySg	78.19	79.33	61.29	57.72	94.37	96.11
RapSg	71.80	72.04	42.54	32.62	95.05	98.11
RockSg	61.84	55.18	37.06	25.88	85.04	91.26
Films	80.07	80.17	74.23	74.02	88.3111	88.37
VariousSg	87.20	88.23	74.24	74.38	95.0686	96.19
ReggaeSg	63.08	57.89	36.82	25.86	92.27	94.01
MusicSg	71.00	69.31	54.56	48.83	88.88	90.77
Speech	92.46	96.62	89.96	96.52	94.28	97.93
Total	86.84	88.78	79.49	81.71	93.77	96.29

Channel Normalization (CN) – subtraction of the mean feature vector from each of the feature vectors – is usually used to reduce the influence of the channel variability over the speaker or speech recognition systems. We study the influence of CN by comparing the discrimination rates for classification achieved with or without channel normalization. In each case, two GMM models are used (one for Speech, another for Music **and** Songs). Averaged scores are given respectively for 100ms and 200ms test time durations (Table 2 (columns 2 to 5). The first column is the name category.

By comparing the columns, we notice that the scores discrimination of speech remains almost similar (92.46% without normalization, 89.96% with normalization for 10 frames averaging; 96.62% without normalization, 96.52% with normalization for 20 frames averaging). However, the discrimination is much better on music and songs when no normalization is used. Particulary, the discrimination for music is reduced from 92.58% to 84.96% when

normalization is used (20 frames averaging). In fact, the normalization, unavoidably removes some features that are supposed to characterize principally the regularity of music tonality. Therefore the normalization is inappropriate for music signal. When the music is mixed with singers the difference is greater (from 69.31% to 48.83% with 20 frames averaging). This situation is probably related to the fact that the music regularity is perturbed and speech singer is improved when the normalization is taken into account. Here we can retain that the normalization is inappropriate for parameters deriver from music and song signals.

4.4.2. Integration of a model for Songs

In order to increase the discriminability, we introduce a third GMM model that is dedicated to music mixed with songs. An experiment is performed with three GMM models. The first is trained with speech, the second with music only and the third with songs mixed with music. The same data and durations than in previous subsections have been presented to the system.

It is known that during all training sessions, exhaustive data should be presented to the system in order to better adapt the model parameters (various speaker conversations and music styles are required). For this purpose, the computational time is expensive and can even exceeds the hardware limits. Also, the number of mixtures can increase considerably to accurately take into account all the situations and styles. Another alternative would be in modelling separately each music style with a different model. However the execution time would be considerably longer. We think that such training technique is intended to characterize the intra-speaker and intra-music styles variabilities. We do not need such accurate models but prefer models that loosely estimate the distributions of speech, music and music with songs, as we want general characteristics for each classes while achieving a good discrimination. So, we propose here to limit the number of GMMs to 8 and to use short duration data for training. Even if in previous works, up to 64 number of mixture were used for the same problem, we show that good discrimination is obtained by using 8 mixtures of Gaussians per model of classes.

Discrimination scores are reported in table 2 (columns 6 and 7). No normalization has been performed. An averaged discrimination score of 93.77% (10 frames averaging) or of 96.29% (20 frames) is obtained in comparison to 86.83% and 88.78% for the same conditions but with 2 GMMs (one for speech, one for music and songs). Therefore, the use of 2 models to differentiate music and music with songs (instead of one model for both) is a good strategy.

5. A FASTER SPEECH-MUSIC DISCRIMINATOR: TRACKING THE SHORT-TERM VARIABILITY BASED ON ΔMFCC

5.1. Feature extraction

We extract features that exploit the short-time variability and irregular duration of speech tonalities. In a preliminary study we observed that ΔMFCC yield a better discrimination than MFCC. Therefore, in the following, ΔMFCC are used as basic features. For each 32 ms frame, 12 delta coefficients are obtained. The 32 ms frame is shifted every 10 ms. The dynamic tonality indirectly is captured by different measurements. We first compute the Euclidian distance between $\Delta MFCC$ parameter vectors from 3 adjacent segments. Let us write $\Delta MFCC_i(n)$ the $i\ th$ dimension of vector $\Delta MFCC(n)$, where $\Delta MFCC(n)$ is the vector of 12 delta Mel Cepstrum Coefficients computed at frame n. We define $d_1(n)$ and $d_2(n)$ as:

$$d_1(n) = \sum_{i=1}^{12} (\Delta MFCC_i(n) - \Delta MFCC_i(n-1))^2 \quad (1)$$

$$d_2(n) = \sum_{i=1}^{12} (\Delta MFCC_i(n) - \Delta MFCC_i(n-2))^2 \quad (2)$$

It is observed that the time evolution of the two distances between segments is almost stable with weak fluctuations for all the music segments, while significant fluctuations are observed for speech segments. Then, the standard deviation of these distances over a duration of 100 ms is calculated.

$$\sigma_1(n) = \frac{\sqrt{\sum_{i=n-L}^{n} (d_1(n) - \overline{d_1})^2}}{L} \quad (3)$$

$$\sigma_2(n) = \frac{\sqrt{\sum_{i=n-L}^{n} (d_2(n) - \overline{d_2})^2}}{L} \quad (4)$$

In preliminary experiments, we found that speech is always characterized by strong transitions and strong amplitude of $\sigma_i(n)$, which is not the case for the music. This motivates us to define a parameter $P_\sigma(n)$ that takes into consideration this observation.

$$P_\sigma(n) = (\sigma_1(n) - \sigma_2(n))^2 + max(\sigma_1(n), \sigma_2(n)) \quad (5)$$

Based on this parameter, we proceed now in the classification experiments between *Speech*, *Music* and *Music with Songs*.

5.2. Decision boundaries

In this section, we study the possibility of discriminating speech from any kind of music or songs by finding the boundary between the classes. As the parameter $P_\sigma(n)$ is scalar, the Bayesian classification scheme is reduced to the problem of finding the hresholds that yield the optimal results. As we suppose the apriori equiprobability of the classes, minimizing the Bayesian cost function is equivalent in maximizing the recognition rate when a lossless decision matrix is used.

In next subsection, we show that speech/music classification can be succesfully achieved with only one threshold instead of two.

5.3. Results

Three case studies are presented and illustrated (Figure 1).
1. *Speech/Music* **discrimination:** (Fig. 1 (a))
The optimal threshold is located at the curves intersection with approximately a value of 0.001 for the threshold and a discrimination score of 90%. However, if one wants to introduce the notion of loss associated to decisions by favoring the recognition of one class relatively to another, the optimal threshold should be moved to the right or to the left direction A better discrimination of speech is observed when the threshold is lower. On the opposite, when the threshold is increased, the scenario is completely reversed (good discrimination for music but not for speech).
2. *Speech/Music with Songs* **discrimination:** (Fig. 1 (b))
The optimal threshold is moved considerably to the right, makes the average discriminant score fall to 80%. This means that the

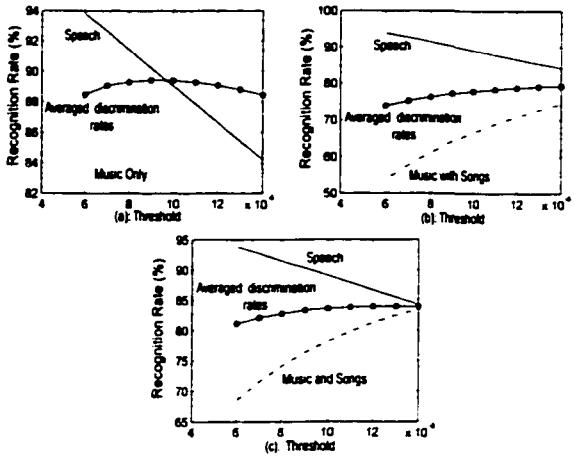

Fig. 1. System performance with threshold strategy: **(a)**: classification of *Speech/Music Only*. **(b)**: classification of *Speech/Songs*. **(c)**: classification of *Speech/Music with Songs*. Continuous curves in (a),(b) and (c) are the recognition rate for *Speech*. Dot line curves are the recognition rates (a) for *Music Only*, (b) for *Songs* and (c) for both *Music* **and** *Songs*.-*- is the total averaged discrimination rate.

overlap with speech is more significant for the class *Music with Song* than for *Music Only*.

3. **Speech/merged** *Music* **class with the** *Song* **class discrimination:** (Fig. 1 (c))

The optimal threshold has a value of 0.0014 and a score of 84%.

For the reason that 3 classes (*Speech, Music, Music with Songs*) are considered, the classification should be performed with at least two boundaries (i.e. thresholds). But we observe that the *Music with Songs* class can be merged with the *Music Only* class to create one class. By doing so, the discrimination can still be performed with satisfactory results even if this architecture is not associated with the optimal solution of the problem. An optimal solution would require a greater processing time.

5.4. Discussion

In comparison to most common works that usually use test durations on the order of 1000 ms, we obtain a comparable score with a shorter test duration (100 ms). This represents a factor of reduction of 9/10. To increase the discrimination rate, we can estimate the standard deviations of the proposed distances over durations longer than 100 ms.

On the other hand, for *Music with Songs* and when no specific model is used for *Music* or *Songs*, the performance decreases according to the coupling between the music and the singer signal. We retain that the singer signal which does not dominate the music signal (energy) is recognized as music. On the other hand, if the singer signal dominates and affects the regularity of music signal, it is recognized as speech. Combining an energy tracker and a silence detector could be a good strategy to retrieve the singer segments. But, the processing requires more computations and a greater time analysis.

6. CONCLUSION

We investigated two systems to discriminate *Speech*, *Music* and *Music with Songs* in the context of telephony with handsets variability. The first system uses three Gaussian Mixture Models (GMM) for speech, music and songs respectively. The reference system uses two GMMs, the first for *Speech* and the second for *Music Only* along with *Music and Songs*. Each GMM comprises 8 Gaussians trained on very short sessions and tested on very large sessions.

We have shown that, when channel normalization is used to improve speech recognition or identification rates in the context of handsets variability, then the speech/music discrimination score reduces by 10% for *Music* and by 20% for *Music with Songs*. Indeed, the normalization of feature vectors removed the regularity structure of musical signal.

Over a 100 ms of test duration, the scores obtained for the reference system with normalization and the proposed system respectively are 89.96% and 94.28% in the case of speech, 83.50% and 97.65% in the case of music and 70.67% and 91.91%in the case of song music. The score for test durations of 100 ms and 200 ms are almost similar, so we think that if we reduce the duration to 50 ms the score should not change too much.

The second system is based on an empirical transformation of the ΔMFCC that enhances the dynamical evolution of tonality. Even if it is optimized for integration in real time applications, it yields an acceptable discrimination rate of 84% with a 100 ms test duration. The results are comparable to other scores reported in the literature that use on the average 1 second. Therefore a reduction of 9/10 is observed.

Finally, in the context of handsets variability, when the discrimination system is used in conjunction with Speech or Speaker Recognition, the normalization of the MFCC should be only performed by the recognition systems and not by the speech/music discriminator

7. REFERENCES

[1] Ajmera J., McCowan I. , and Bourlard H., "Robust HMM-Based Speech/Music Segmentation," in *ICASSP'02*, 2002.

[2] Carey M. J., Parris E. S., and Lloyd-Thomas H., "A comparison of features for speech, music discrimination," in *ICASSP'99*, 1999.

[3] El-Maleh K., Klein M., Petrucci G., and Kabal P., "Speech/music discrimination for multimedia applications," in *ICASSP'00*, 2000.

[4] Tancerel L., Ragot S., and Lefebvre R., "Speech/music discrimination for universal audio coding," in *20th Biennal Symposium on Communications*, 2000, pp. 28–31.

[5] Samouelian A., Robert-Ribes J., and Plumpe M., "Speech, silence, music ans noise classification of tv broadcast material," in *ICSLP'98*, 1998.

[6] Saunders John, "Real-time discrimination of broadcast speech/music," in *ICASSP'96*, 1996, pp. 993–996.

[7] Scheirer E. and Stanley M., "Construction and evaluation of a robust multifeature speech/music discriminator," in *ICASSP'97*, 1997, vol. II, pp. 1331–1334.

TOWARDS COMBINING PITCH AND MFCC FOR SPEAKER IDENTIFICATION SYSTEMS

Hassan Ezzaidi, Jean Rouat and Douglas O'Shaughnessy[+]

Ermetis, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada, G7H 2B1
[+]INRS-Télécommunications, Université du Québec
900 de la Gauchetiere west, Box 644, Montreal, Québec, Canada, H5A 1C6

ABSTRACT

Usually, speaker recognition systems do not take into account the dependence between the vocal source and the vocal tract. A feasibility study that retains this dependence is presented here. A model of joint probability functions of the pitch and the feature vectors is proposed. Three strategies are designed and compared for all female speakers taken from the SPIDRE corpus. The first operates on all voiced and unvoiced speech segments (baseline strategy). The second strategy considers only the voiced speech segments and the last includes the pitch information along with the standard MFCC. We use two pattern recognizers: LVQ–SLP and GMM. In all cases, we observe an increase of the identification rates and more specifically when using a time duration of 500ms (6% higher).

1. INTRODUCTION

The vibration frequency of the vocal folds is known to be an important feature to characterize speech and has been found effective for automatic speech and speaker recognition [1] [6]. An important characteristic of pitch is its robustness to noise and channel distortions. Many parametrizations of pitch such as pitch value, averaged pitch, pitch contour, pitch jitter and location [1] [4] have been proposed for speaker verification or identification. Speaker recognition systems exclusively based on pitch do well when the number of speakers [1] is small. However, performance decreases significantly when the number of speakers increases, but pitch information can be reliably used to distinguish the sex of speakers [5]. In spite of the weak contribution of pitch to contemporary speaker identification, it remains true that the mechanisms involved in speech production are complex, and imply dependence of articulators and vocal folds, which can be useful for speaker verification or identification.

The most popular way used to model pitch is by a Gaussian density or a mixture of Gaussians. Statistical independence of the glottis and the vocal tract is assumed by these models. In this paper, we propose to take into account the correlation between the glottis and the vocal tract. We study the influence of this dependence in the context of a text–independent Speaker Identification System (SIS). We use a joint probability function to take into account the correlation between source and vocal tract. The proposed approach consists of generating models of the feature vectors for each pitch range.

hezzaidi@uqac.uquebec.ca, jrouat@uqac.uquebec.ca, dougo@inrs-telecom.uquebec.ca

This work was funded by NSERC, Communications Security Establishment and the FUQAC. Many thanks to Karl Boutin for his support.

The next section describes the motivation for this research. The baseline and the proposed systems are described in sections 5 and 7. Sections 8 and 9 present the results, discussion and conclusion.

2. MOTIVATION

Most systems use parameters that encode vocal tract features, but contributions of the glottis are largely ignored. Even if MFCC are theoretically known to deconvolve the source and the vocal tract, in practice, cepstrum coefficients are affected by high pitched voices (women and infants). One can illustrate the role of pitch when dependence of the source and the vocal tract are maintained. Figure 1 exhibits four spectrograms and pitch histograms, each column corresponds to a different male speaker, obtained from the YOHO database. All speakers pronounced the same digit utterance 'twenty six'. The pitch range is divided into 56 linearly bins of 10 Hz width. The spectrograms show a significant similarity of formant distributions between speakers. The spatial distribution of formants depends on the interspeaker variability as described in [2]. However, the pitch histograms are different and vary from one speaker to another for the same context. If one compares the histograms by taking into account their frequency amplitude and width, it is observed that speaker 2 from the second column and speaker 4 from the fourth column do have a similar pitch distribution. On the other hand, speakers 1 and 3 are characterized by

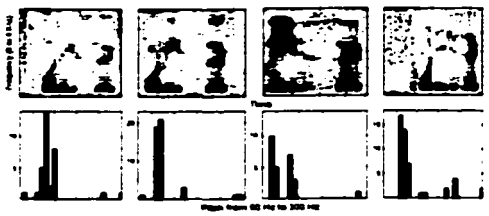

Figure 1: For each of four male speakers: Pitch histograms and spectrograms for the same English digit utterance '26'.

dissimilar pitch histograms. Consequently, if one takes into consideration the pitch information, the interspeaker variability can be restricted to speakers with similar pitch distributions, and the other speakers will be considered as belonging to other clusters. Speakers with similar pitch will be recognized based on the spectral characteristics.

In summary, pitch and vocal tract features can be jointly exploited in order to establish probability models of feature vectors

assuming the a priori knowledge of the pitch distribution.

3. PROPOSED MODEL

3.1. Theoretical framework

We suppose that pitch and vocal tract features are two random process respectively denoted as $X(t)$ and $Y(t)$. Let's write $\widehat{X(n)}$, the estimated discretized pitch frequency at time $n\Delta t$ and $\widehat{Y(n)}$ the estimated discretized vocal tract feature vector at time $n\Delta t$. $\widehat{Y(n)}$ is an l-dimensional vector. In practice $\widehat{Y(n)}$ is an LPC or $MFCC$ vector estimated from a centered signal window at time $n\Delta t$. For each process $\widehat{X(n)}$ and $\widehat{Y(n)}$, we assume time independence of their respective realizations. As a consequence, $\widehat{X(n+1)}$ is independent of the realization of $\widehat{X(n)}$. The same restriction applies to $\widehat{Y(n)}$. In the following, we drop the time and consider the simultaneous realization of $\widehat{X(n)}$ and $\widehat{Y(n)}$ as being time independent. The crosscorrelation between $\widehat{X(n)}$ and $\widehat{Y(n)}$ is still preserved.

Let us write $\{x_1, x_2, .., x_n\}$, the increasing sequence of realizations of $\widehat{X}$, with $x_i \in [60\ Hz, 660\ Hz]$. We suppose that the set of realizations of $\widehat{Y}$ is finite (by using vector quantization for example) and equal to $\{\overrightarrow{y_1}, \overrightarrow{y_2}, .., \overrightarrow{y_m}\}$, with $\overrightarrow{y_i} \in R^l$. Let f to be the joint probability of $\widehat{X}$ and $\widehat{Y}$.

$$f(x_i, \overrightarrow{y_j}) = P(\widehat{X} = x_i, \widehat{Y} = \overrightarrow{y_j})\ with \tag{1}$$

$$0 \leq f(x_i, \overrightarrow{y_j}) \leq 1\ and \sum_{i=1}^{i=n} \sum_{j=1}^{j=m} f(x_i, \overrightarrow{y_j}) = 1. \tag{2}$$

The respective marginal probability functions are:

$$f(x_i) = \sum_{j=1}^{m} f(x_i, \overrightarrow{y_j})\ and\ f(y_j) = \sum_{i=1}^{n} f(x_i, \overrightarrow{y_j}). \tag{3}$$

Each speaker s is supposed to be defined by its probability function,

$$f_s(x_i, \overrightarrow{y_j}) = P_s(\widehat{X} = x_i, \widehat{Y} = \overrightarrow{y_j}). \tag{4}$$

We observe that

$$f_s(x_i, \overrightarrow{y_j}) \quad = \quad f_s(\overrightarrow{y_j}/x_i) f_s(x_i) \tag{5}$$

$f_s(x_i)$ is the *a priori* probability of a pitch frequency to be equal to x_i and $f_s(\overrightarrow{y_j}/x_i)$ is the *a posteriori* probability of observing a feature vector to be equal to $\overrightarrow{y_j}$ given the knowledge of the pitch frequency x_i. The estimation of the a priori probability of the pitch frequency is relatively straightforward while the estimation of $f_s(\overrightarrow{y_j}/x_i)$ can be long and tedious.

3.2. Feature vector distributions based on pitch knowledge

In the present work we focus on the estimation and integration of the posteriori probability, $f_s(\overrightarrow{y_j}/x_i)$, in speaker recognition systems. The consideration of the factor $f_s(x_i)$ from equation 5 is left as a future work.

We propose to subdivide the space $(x, \overrightarrow{y})$ into subspaces H_k where $f_s(\overrightarrow{y_j}/x_i)$ is supposed to be locally independent of the pitch value. Let us define I_k, $k = 1, \ldots, N$ as sub–intervals of the pitch set $\{x_1, x_2, .., x_n\}$. We recall that $x_1 = 66\ Hz$ and $x_n = 660\ Hz$. N is the number of intervals with $I_1 \cup \ldots \cup I_N = \{x_1, x_2, ... x_n\}$. Each subspace H_k is associated to a pitch interval I_k. For each H_k, we suppose that the probability function is stationary and independent of the pitch inside the interval I_k, that is,

$$f_s(\overrightarrow{y_j}/x_i) \simeq f_s(\overrightarrow{y_j}/I_k). \tag{6}$$

Theoretically, the number of models $f_s(\overrightarrow{y_j}/x_i)$ would be equal

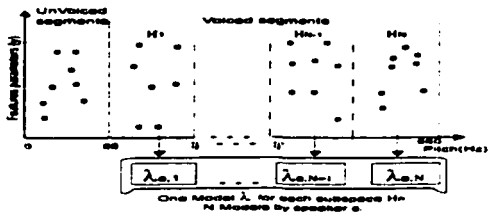

Figure 2: Proposed approach for generating sub–models.

to n. By subdividing the space into N subspaces, we reduce that number to N. Figure 2 illustrates the notion of subspaces and models of probability functions $f_s(\overrightarrow{y_j}/I_k)$. The interval length of I_k is based on the shape of the pitch histogram (section 7.3).

4. SPEECH ANALYSIS

Mel Cepstrum Coefficients derived from a bank of filters (MFCC) are used as features to characterize the identity of speakers. We use coefficients c_1 to c_{12}. The speech is first preemphasized (0.97); then, a sliding Hamming window with a length of 32 ms and a shift of 10 ms is positioned on the signal. Cepstral mean normalization and liftering are also performed. Delta and delta-delta MFCC are not used, as the comparison between the systems would be biased. In fact, adjacent segments can have different pitch values belonging to different sub–intervals I_k.

5. PATTERN RECOGNITION

5.1. Framework

Two pattern recognizers are used for the experimental task: one parametric and one non-parametric.

5.1.1. Parametric model

We use a Gaussian Mixture Model (GMM) [7] with 32 ($M = 32$) weighted sums of Gaussians. Each GMM is defined for a specific speaker s and pitch interval I_k. Let us define $p(\overrightarrow{y}/\lambda_{s,k})$, the Gaussian mixture density associated to the probability function $f_s(\overrightarrow{y_j}/I_k)$ for speaker s, as

$$p(\overrightarrow{y}/\lambda_{s,k}) = \sum_{i=1}^{M} w_{i,k} b_{i,k}(\overrightarrow{y}) \tag{7}$$

with

$b_{i,k}(\overrightarrow{y}) = \frac{1}{(2\pi)^{l/2}|\Sigma_{i,k}|^{1/2}} exp\{-\frac{1}{2}(\overrightarrow{y} - \overrightarrow{\mu_{i,k}})' \Sigma_{i,k}^{-1} (\overrightarrow{y} - \overrightarrow{\mu_{i,k}})\}$.

M is the GMM order, $\overrightarrow{y}$ is the l-dimensional vector estimating the vocal tract contribution (MFCC vector), $b_{i,k}$ is the i-th Gaussian density with mean $\mu_{i,k}$ and covariance matrix $\Sigma_{i,k}$ and $w_{i,k}$ are the mixture weights. $b_{i,k}$, $\mu_{i,k}$, $\Sigma_{i,k}$ and $w_{i,k}$ are defined for pitch

interval I_k and for speaker s. Each speaker is characterized by N models $\lambda_{s,k}$ corresponding to N pitch intervals I_k.

5.1.2. Non-parametric model

We use the hybrid LVQ-SLP network as proposed by He *et al.* [3]. Each speaker s, with a pitch belonging to I_k, is characterized by a codebook $C_{s,k}$. The codebook size is the same for all speakers. We performed experiments with codebook sizes of 512 for each speaker.

5.2. Recognition

5.2.1. Parametric model

We define T as being the test length over which the recognition is performed. A frame-by-frame estimation of log-likelihood for each speaker s and pitch interval I_k is first performed. Each frame (32 ms length) is shifted by 10 ms. Then, the maximum log-likehood for each speaker is estimated over T. When the test sentence is longer than T, the average of the score over the number of segments with a length of T is computed according to equation 8.

$$S_T^s = \frac{nb.\ of\ seg.\ correctly\ tested\ for\ T\ duration}{total\ nb.\ of\ seg.\ tested\ for\ T\ duration} \quad (8)$$

The final identification score (equation 9) is obtained by averaging over the number of speakers N_s:

$$Score = \frac{\sum_{i=1}^{N_s} S_T^s}{Ns} \quad (9)$$

5.2.2. Non-parametric model

For each frame, the feature vector is classified by using the Nearest Neighbor criteria. A speaker is recognized if, for the entire test conversation, it is selected more frequently than the other speakers.

6. SPEECH DATABASE

A subset of the SPIDRE-Swichboard Corpus is used and comprises eighteen (18) female speakers of the database. Each speaker has 4 conversations originating from 3 different handsets. The training data contains 3 conversations, with 2 conversations coming from the same handset. The last conversation, using the third handset (different from the others), is presented as the test data. This combination is referred to as the *mismatched condition*. The *matched condition* refers to situations where training and testing data are recorded from the same handset.

7. STRATEGIES

7.1. The baseline strategy

The baseline strategy uses both the voiced and unvoiced segments. The suppression of silence was carried out based on the energy evolution and comparison with fixed thresholds.

7.2. Recognition based on voiced speech segments

We include a module that estimates the pitch and selects the voiced segments. We use a pitch tracker and a voiced-unvoiced detection system [8] in conjunction with the SID system analysis module. In this case, silence and unvoiced segments are automatically rejected. During training and for each pitch period, we centered a 32 ms duration window and extracted the MFCC coefficients.

7.3. Recognition based on the estimated a posteriori probabilites

For the third strategy, four pitch intervals $I_1,\ldots,I_4$ are created according to the pitch frequency histogram. More than 90% of the pitch frequencies belong to the interval [150Hz,220Hz]. We distributed the pitch frequencies over 4 intervals I_1=[150,180], I_2=[170,200], I_3=[190,220] and I_4=[66,150]$\cup$[220,660]. The choice of four intervals is a trade off between fine pitch intervals and sufficient training size of the models. During training and for each interval I_k, the MFCC vectors are used to generate model parameters for each speaker. Therefore each speaker is characterized by 4 models. With the aim of overcoming the pitch estimation errors, we choose an overlap of 10 Hz between the intervals. Thus, the MFCC vectors from speech whose fundamental frequency belongs to two adjacent intervals (I_k, I_{k+1}), will be used to train two models, respectively, associated to subspaces H_k and H_{k+1}. Then, during the testing session, the evaluation is carried out over these two subspaces and we keep the best score.

In the case of LVQ-SLP, the coodbook generation is made according to two procedures. One attributes the same codebook size to each subspace, and the other distributes the number of prototypes per codebook according to the number of events in each subspace.

In the case of the GMM models, one model λ_s is generated for the baseline system, one model is also used for recognition on voiced speech and four models $\lambda_{s,k}$ are generated for the recognition taking into account the a posteriori probabilities of voiced speech according to the pitch.

8. RESULTS AND DISCUSSION

8.1. Evaluation with a LVQ-SLP model

LVQ-SLP results for 18 women of the SPIDRE database are reported in tables 1 and 2.

Table 1: LVQ-SLP: Identification rate increases for 18 female speakers with fixed codebook sizes. Baseline system: 55%.

	Voiced (512)	H_1 (128)	H_2 (128)	H_3 (128)	H_4 (128)
Matched	6	14	10	1	0
Mismatched	3	4	4	5	2

When the unvoiced segments are not taken into account (Voiced column), the identification rate increases to 61% (6% more in table 1) for matched handsets and to 58% (3%) for mismatched handsets. When pitch is taken into account (columns H_k in tables 1 and 2), the increase is almost the double in H_1 and H_2 and weaker in H_3 and H_4. H_1 and H_2 are the subspaces with the greatest number of events and H_3 and H_4 with the smallest number of events. When the number of prototypes per codebook is

Table 2: LVQ–SLP: Identification rate increases for 18 female speakers with codebook sizes proportional to the number of events in each subspace. Baseline system: 55%.

	H_1	H_2	H_3	H_4
Matched	14	3	0	-1
Mismatched	4	5	6	1

proportional to the number of events, performance falls in H_2 and H_4 and remains constant in H_1.

When recognition rates are weighted according to the number of events per subspace, we obtain an averaged increase of 8% in matched conditions, and 4% in mismatched situations. It is observed that the identification results are sensitive to several factors: 1) codebook sizes, 2) training techniques, 3) scores combination, and 4) pitch estimation. The best increase in performance is observed for subspaces with the greatest number of events.

8.2. Evaluation with a GMM model

Table 3 reports the identification results observed with the three strategies: 1) Baseline (voiced and unvoiced segments), 2) Voiced (only voiced segments) and 3) Voiced segments with partition of space into H_1 to H_4. The first column gives the value of T, that is, the duration of maximum log–likehood estimation. The baseline

Table 3: GMM: Mismatched identification rates for 18 female speakers.

Time(seconds)	Baseline(%)	Voiced(%)	Voiced & pitch(%)
0.1	36.8	37.7	40.5
0.5	63.4	66.7	69.4
1	75.4	79.9	80.8
2	84.2	87.9	88.0
3	88.0	90.6	90.5
4	90.0	93.9	93.3
5	91.4	95.4	94.7
6	92.7	95.3	95.2

strategy yields the lowest identification rates. When voiced segments are used, the best increase is 4.5% and the weakest is 1%. When a preliminary subdivision based on pitch is performed, the greatest increase is 6% and the weakest is 2.5%.

When the test duration is greater than 2 seconds, strategies based on voiced segments yield similar results. When T is less than 2 seconds, the best results are observed with the strategy that takes into account the posterior probability (subdivision into subspaces). The increase is on the order of 2%.

8.3. Discussion

Identification rates of LVQ–SLP and GMM are not strictly comparable, as the recognition criteria is different. The comparison of Tables 1 and 2 suggests that the a priori probability $f_s(x_i)$ should be taken into account. In Table 3, a T of 1 second is equivalent to 100 MFCC vectors and is independent of the strategy. The weaker performance of the baseline system might be partially due to the smaller number of voiced frames in a fixed T. In several cases, the pitch is not well estimated and affects the performance. If these errors are corrected, we can possibly achieve a better training and evaluation. The voiced pitch strategy requires more calculation and is more sensitive to errors especially during training.

9. CONCLUSION

A new approach that preserves the dependence between the vocal source and the vocal tract has been proposed. Experiments that integrate the a posteriori probability of observing a MFCC vector given the knowledge of the pitch frequency have been reported. They are compared with a baseline system operating on all voiced and unvoiced speech segments and with a second system that operates on voiced speech segments only. Closed set Speaker Identification experiments were performed on a subset of the SPIDRE corpus that comprises highly confusable female speakers. Systems based on voiced segments yield the best scores.

When the dependence of the source and vocal tract is taken into account, the best results are observed for durations T lower than 3 seconds (up to 4.5% for $T = 500$ ms). For $T \geq 3$ seconds scores are 1% higher, in favour of the system based on voiced segments only.

Despite the small improvement in performance, it appears to us that the approach is promising. In fact, many restrictive hypotheses have been made to set up the experiments. The pitch tracker has been supposed to be reliable; sufficient training data for subspaces decomposition, local independence of MFCC in relation to pitch in a subspace (equation 6), and time independence of pitch and MFCC have been assumed.

We therefore suggest, as future work, to increase the size of the corpus for a better statistical convergence, to optimize the number and width of the pitch intervals (I_k) and to introduce weighting by the a priori probability distribution of the pitch ($f_s(x_i)$) in accordance with equation 5. We also suggest restricting the application to a text-dependent system, for which the variability of the parameters is usually smaller.

10. REFERENCES

[1] B. S. Atal. Automatic recognition of speakers from their voices. In *Proc. IEEE 1976*, volume 64, pages 460–475, 1976.

[2] G. R. Doddington. Speaker recognition-identification people by their voices. In *Proc. IEEE Vol.73, No 11*, number 11, pages 1651–1664, 1985.

[3] J. He, L. Liu, and G. Palm. Speaker identification using hybrid lvq-slp networks. In *Proc. IEEE ICNN Vol.4*, pages 2051–2055, 1995.

[4] C.R Jankowski Jr., T.F. Quatieri, and D.A. Reynolds. Measuring fine structure in speech: Application to speaker idendification. In *IEEE-ICASSP*, pages 325–328, 1995.

[5] J.Rouat, H. Ezzaidi, and M. Lapointe. Nouv. alg. d'extract. en vue de caract. le loc. Technical report, March 1999. Contrat W2213-9-2234/SL, 67 pages.

[6] Douglas O'Shaughnessy and Hesham Tolba. Towards a robust/fast continuous speech recognition system using a voiced-unvoiced decision. pages 413–416, ICASSP, 1999.

[7] Douglas A. Reynolds and Richard C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. 3(1):72–83, 1995.

[8] J. Rouat, Y.C. Liu, and D. Morissette. A pitch determination and voiced/unvoiced decision algorithm for noisy speech. *Speech Communication*, 21:191–207, 1997.

COMPARISON OF MFCC AND PITCH SYNCHRONOUS AM, FM PARAMETERS FOR SPEAKER IDENTIFICATION

Hassan EZZAIDI and Jean ROUAT

ERMETIS, DSA, Université du Québec à Chicoutimi
Chicoutimi, Québec, Canada, G7H 2B1
hezzaidi@uqac.uquebec.ca, jrouat@uqac.uquebec.ca

ABSTRACT

We study robust pitch synchronous parameters that are derived from envelope and instantaneous frequencies estimated via a bank of cochlear filters. Closed set Speaker Identification experiments are performed on the SPIDRE corpus with matched and mismatched handsets conditions. The recognizer is based on a hybrid Linear Vector Quantization and Single Layer Perceptron (LVQ-SLP). Experiments are reported with different codebook sizes. In mismatched condition, the Mel Frequency Cepstral Coefficients (MFCC) yield slightly better rating (68%) than Envelope (58%) and Instantaneous Frequency (65%) parameters when used independently. When the MFCC based recognizer is used in conjunction with the envelope based recognizer, the recognition rate increases to 80%. We also report identification rates based on two classes: women and men. In another experiment, listeners were asked to discriminate speakers on a subset of ten females. We discuss their performance. We also discuss the potential of the approach and of judicious combination of the parameters to improve Speaker Identification Systems.

1. INTRODUCTION

Although many experiments on clean speech report high identification rates [5], results on noisy speech [8] [5] are usually too poor for practical identification tasks [11]. The *cleaning* of speech before analysis and recognition can improve significantly the performance but it is not always feasible. Telephone handsets introduce nonlinear distortions of the speech signal. Since the handset properties are not a priori known, it is not possible to filter out the handset distortions [8]. In this context, robust parameter extraction and new identification strategies have to be elaborated.
Today, most of the speech systems assume that the signal is stationary under the analysis window. This yields a representation that is an estimate of the time averaged parameter values. Therefore, the fine structure of speech (as short as 2–3ms) is partially hidden by the analysis and can not be exploited. One of the exceptions is the work of Jankowski *et al.* who use cepstral parameters derived from the Teager Energy operator and the DESA-1 energy separation algorithm [5].
We first use a cochlear filter bank to generate two spatio-temporal representations of speech: Envelopes and Instantaneous Frequencies (IF). Then, new feature vectors are derived from the two spatio-temporal representations. They are synchronous with the glottis and are used for speaker identification (SID). The proposed features characterize the very fine structure of speech on intervals as short as 4-5 ms. No assumption of stationary speech has to be made.
In the next section, we describe the AM and IF structure of speech. Section 3 presents the structure of the speaker identification system. The experiments with results are reported in section 4 and 5. Finally, section 6 and section 7 give a discussion and conclusion.

This work was supported by NSERC, CSE and fondation de l'université du Québec à Chicoutimi. Thanks to Karl Boutin and Luc Gagnon from CSE, Mathieu Lapointe, P. Dumouchel and P. Ouellet.

2. AM AND IF: A SPEECH STRUCTURE

2.1. Modulation in the Auditory System

The modulation information is one of the main cues extracted by the auditory system. Various work is made regarding the physiology of AM and FM processing [7] [3].
The work by Delgutte *et al.* [1] [2] and other authors suggest that the auditory system should be able to track simultaneously formants and pitch by relying on phase-coupling of auditory fibers and on patterns of modulation for fibers influenced by a summation of stimulus harmonics. This partially motivates the proposed analysis.

2.2. The Envelope and IF Fine Structures

Examples of short-term envelope and IF fine structures are given below. A bank of 24 cochlear filters centered on 330Hz to 4700Hz is used [6]. The output of each filter is a bandpass signal with a narrow-band spectrum centered around f_{c_i} where f_{c_i} is the central frequency (CF) of channel i. The output signal $s_i(t)$ from channel i can be considered to be modulated in amplitude and phase with a carrier frequency of f_{c_i}.

$$s_i(t) = A_i(t)\cos[2\pi f_{c_i} t + \phi_i(t)] \qquad (1)$$

$A_i(t)$ is the modulated amplitude (envelope) and $\phi_i(t)$ is the modulated phase. The instantaneous frequency $F_i(t)$ is defined as

$$F_i(t) = f_{c_i} + \frac{1}{2\pi}\frac{d\phi_i(t)}{dt}$$

If $\widehat{s_i(t)}$ is the Hilbert transform of $s_i(t)$ and $s_i(t)'$ the time derivative of $s_i(t)$ we can write

$$A_i(t) = \sqrt{s_i(t)^2 + \widehat{s_i(t)}^2} \qquad (2)$$

$$F_i(t) = \frac{s_i(t)\widehat{s_i(t)}' - s_i(t)'\widehat{s_i(t)}}{s_i(t)^2 + \widehat{s_i(t)}^2} \qquad (3)$$

A representation of the fine envelope and IF structures versus time is illustrated on figure 1.

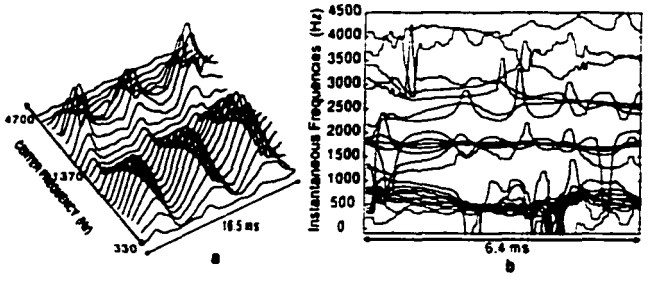

Figure 1: **a: Envelopes $A_i(t)$ ($i = 1,...,24$) for a woman, the y axis is expressed in Hertz according to the ERB scale. b: Instantaneous Frequencies $F_i(t)$ ($i = 1,...,24$) for a man (after median filter).**

3. THE SPEAKER IDENTIFICATION SYSTEM

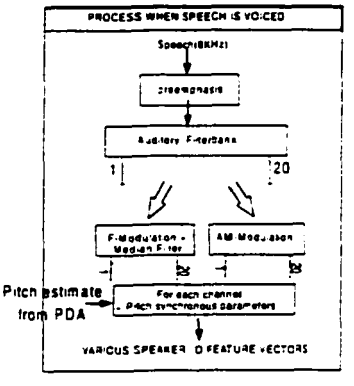

Figure 2: The Speaker Identification System.

We use a pitch tracker and voiced/unvoiced detection system (PDA) [10] in conjunction with the analysis module (figure 2). We use the twenty first filters of the bank described in section 2.2 (center frequencies from 330Hz to 3030Hz) and generate the envelopes and instantaneous frequencies.

3.1. The proposed envelope and IF parameters

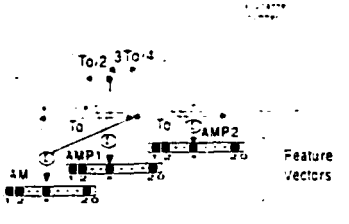

Figure 3: Example of extraction of parameters the x^{th} dimension of AM, AMP1 and AMP2 from $A_{11}(t)$, $x = 11$.

The Pitch Determination Algorithm is first used to estimate the fundamental period T_0. The exact value of T_0 is further refined by peak-picking the $A_i(t)$. For each interval $[0,T_0]$, $A_i(t)$ and $F_i(t)$ are computed according to equations 2 and 3. The time origin of each interval $[0,T_0]$ is the position of the glottal peak. Then, we create six feature vectors that we call IF, DIF, AM, DAMP, AMP1 and AMP2. Figure 3 is an example of these feature vectors measured on three time intervals ($[0,T_0]$,$[T_0/2,3T_0/4]$, $[3T_0/4,T_0]$) and coming from the same glottal interval.

The feature vectors are created for each glottal interval (and for each channel). They are averaged values of $A_i(t)$ (or $F_i(t)$) estimated from different segments. For each interval, the parameter value coming from channel i is used as the ith. dimension of the feature vector. Therefore, the feature vectors we use are of 20 dimensions (except DIF: 19 dimensions) and one feature vector is obtained for each glottal period.

AM (or IF) is the averaged value of $A_i(t)$ (or $F_i(t)$) over $[0,T_0]$. AMP1 is the averaged value of $A_i(t)$ in $[T_0/2, 3T_0/4]$. AMP2 is the averaged value of $A_i(t)$ in $[3T_0/4, T_0]$. DAM is the time difference between adjacent AM. DIF is the averaged difference of $F_i(t)$ through adjacent channels.

AM and IF contains the information from one period. AMP1 is supposed to estimate the **power** of secondary pulses of envelopes. On the other hand AMP2 is supposed to characterize the envelope's **ascending slope** of the source excitation. Estimates of the power rather than the location is supposed to be less sensitive to noise and distortion introduced by the telephone channel. DAM and DIF were proposed to reduce the handset and context dependence of the parameters. They capture the dynamical information. However, the invariant information is also removed.

It is interesting to note that Jankowski *et al.* [5] have already studied the **location** of secondary pulses estimated through the Teager Energy operator. It is also important to take into consideration that the Teager Energy operator yields an estimate of the product of $A_i^2(t)$ by $F_i^2(t)$. For the time being, the use of the Hilbert transform alleviates the problem of the distortion terms that we observe in the use of the Teager Energy operator [9].

3.2. The Recognition Module

We use a nonparametric pattern recognizer to compare the parameters. In fact, a preliminary statistical study shows that the envelope and IF parameters would need specific parametric pattern recognizers in order to optimize the performance respectively to each parameter (such as GMM for MFCC). This optimization could be long and tedious. For now, we use the LVQ-SLP as proposed by J. He and al. [4]. Each speaker is characterized by one codebook. The codebook size is the same for all speakers. We performed experiments with codebook sizes of 32, 64, 128 and 256.

4. EXPERIMENTS AND RESULTS

4.1. The Speech Database

A SPIDRE subset of the Swichboard Corpus (23 males and 17 females) is used (40 speakers).Each speaker has 4 conversations originating from 3 different handsets. The training data contains 3 conversations with 2 conversations coming from the same handset. The last conversation, using the third handset (different from the others), is presented as the test data. That combination is referred as *mismatched* condition.

4.2. The Reference System

The MFCC are used as reference. The speech is first preemphazised (0.97), then, a sliding Hamming window with a length of 30 ms and a shift of 10 ms is positioned on the signal. Twelve cepstral coefficients, twelve delta cepstral coefficients (computed according to the regression weighting), one log power and one delta log power are then extracted by using a liftering of 24 . Cepstral mean normalization is also performed. The final dimension of the MFCC vectors is a of 26.

4.3. Results

Recognition criterion
A speaker is recognized if, for the entire test conversation, the speaker is selected more frequently than the other speakers.

4.3.1. Mismatched handsets

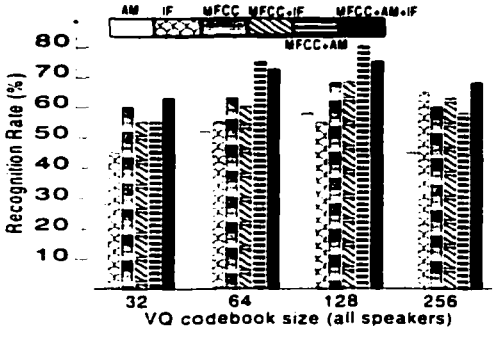

Figure 4: Recognition rates for the AM, IF and MFCC parameters (all 40 speakers).

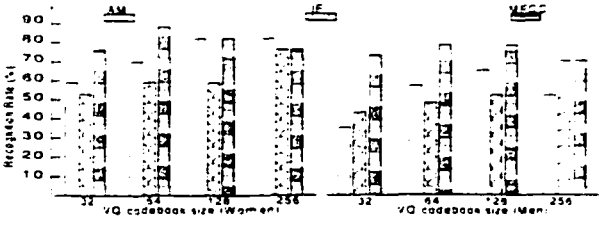

Figure 5: Recognition rates for the AM, IF and MFCC parameters: Same speakers as in Fig. 4. The experiment has been performed on men and women subsets.

With the IF features, the speaker identification recognition rate increases as a function of the codebook size. However, AM and MFCC features performance peaks at 64 or 128 codebook size. The best performance with IF is 70% for men, 76% for women and 65% for all speakers. The AM features performance is 65% for men, 82% for women and 58% for all speakers. As reference, the MFCC features registered a score of 78% for men, 88% for women and 68% for all speakers.

Three *global* Speaker Identification Systems (SIDS) are simulated. They are made of *elementary* SIDS trained respectively on AM, IF and MFCC features. We define a global score matrix of a SIDS as being the sum of the local score matrices from the elementary SIDS. The global score matrix is then used to decide which speaker has been recognized (majority vote). For example, a MFCC+AM *global* SIDS is simulated by adding the score matrix of the MFCC–SIDS with that of the AM–SIDS. Results are also reported in Figure 4. The MFCC+AM speaker identification system yields a recognition rate (80%) that is higher than the MFCC speaker identification system (68%).

4.3.2. Mismatched and matched handsets

The impact of mismatched and matched conditions is evaluated by comparing the performance of parameters derived from the fine structure. In matched condition, the same telephone handset is used for training and testing. One conversation is used for training and the second one for testing . Five of the forty speakers had a conversation that was corrupted and that could not be used for comparison purposes. The recognition rate is degraded

Table 1: Matched handsets, 21 male and 14 female speakers. Codebook size of 256.

	IF	DIF	AM	DAMP	AMP1	AMP2
Men	81%	90%	75%	62%	57%	76%
Women	65%	65%	72%	58%	79%	72%
All Speakers	74%	74%	63%	58%	58%	69%

Table 2: Mismatched handsets, 21 male and 14 female speakers. Codebook size of 256.

	IF	DIF	AM	DAM	AMP1	AMP2
Men	71%	67 %	57%	33%	52%	71%
Women	79%	64%	92%	50%	79%	65%
All Speakers	69%	58%	52%	32%	32%	38%

when there is a mismatch between the handsets (table 2). However, for women, FM, DIF and AM rates are lower with matched (table 1) than with mismatched handsets. It is likely that the number of training samples is not sufficient for the matched condition since only one conversation is used during training in comparison to three for the mismatched condition. The IF, DIF and AM yield better recognition rate than the other proposed features. On the other hand, the AMP2 parameter is better for men speaker, while the AMP1 parameter produces good results for women speaker. A possible explanation is that the relaxation mode of the vocal tract is better measured for men as the glottal period is longer than for women. IF is less affected by the handset mismatch than AM.

5. LISTENING EXPERIMENTS

Pairs of sentences were randomly chosen and played through Senn–heiser HD250 linearII headphones. Three naïve listeners were asked to tell if the speaker was the same for both sentences. The sentences were taken from conversations involving ten female speakers from the SPIDRE database. The listeners could not use sex as discrimination criteria. The listeners are French speaking and could not understand American English. After presentation of two sentences, the listeners had to answer YES (same speaker) or NO (different speaker). The random generator has been biased in order to present 40% of the pairs with the same speaker.

Analysis of the tests clearly show that the handset has a predominant influence on the perception of listeners. In many situations

with the same speaker and two handsets for the two sentences, listeners identified the two conversations as coming from different talkers. After hundred presentations of random pairs the respective performance was of 73%, 68.5% and 65% for the three listeners. No distinction between mismatched and matched handsets has been made in the evaluation. Even if the task was easy in comparison to the identification of forty speakers, the relatively low performance of the listeners gives an idea of the complexity of the SPIDRE database.

6. DISCUSSION

Most of the time, MFCC yields better recognition rate than the recognition based on AM and IF. But performance is still comparable. The crucial point is that the AM, IF or MFCC based recognizers are complementary. They do not confuse the same speakers. The best performance is observed by combining the MFCC recognizer with the AM recognizer.
For women, the IF features yield better performance than AM. In fact, the voiced speech of women comprises fewer harmonics. Therefore the cochlear filters have a better harmonic resolution, yielding a better demodulation for women [12]. Furthermore, The highest the fundamental frequency is, the greatest will be the number of training vectors.
When the sex of a speaker is already known it is possible to obtain better performance. The improvement is sufficiently high to motivate the design and test of a SID system based on a sex preclassification module.

7. CONCLUSION

A cochlear filter bank was first used to generate two spatio-temporal representations of speech: Envelopes and Instantaneous Frequencies. Then, new feature vectors were derived from the two spatio-temporal representations. They are synchronous with the glottis and were used for speaker identification (SID). In comparison with other works, no assumption of stationary speech had to be made and no smoothing or stationary analysis of the features has to be performed. The recognition rates are reasonable for the SPIDRE database that is known to be very difficult. The proposed parameters are very good candidates to improve contemporary SID in difficult environment.
The essence of the proposed parameters and the MFCC is different. The cepstral coefficients use a fixed window length of 30 ms over which an average of 24 points Mel spectrum is estimated (24 filters) before logarithmic and cosine transform. Furthermore, the delta coefficient integrates information on larger time scale (until 80 ms, with the regression formulation). The proposed coefficients are synchronized with the glottis and are estimated over segments that can be shorter than 5 ms (female speakers). They capture the fine structure of speech which is not exploitable when working with the MFCC.

The recognizers we used are simple and do not reflect all the potential of the feature vectors. First, we did not use the temporal structure of the proposed representation, as each feature vectors (obtained for each glottal interval) has been considered to be independent of each other via the LVQ–SLP. Second, we did not combine the feature vectors. In fact, a preliminary experiment on eleven speakers combining AM and DAM in a 40 dimensional vector yielded a significant increase of recognition rate.

As future work, we suggest to exploit the temporal evolution of the proposed parameters, to combine envelops, IF parameters and MFCC and to design more sophisticated pattern recognizers in order to increase the reported performance.

The feature parameters are available at
http://wwdsa.uqac.uquebec.ca/~j2rouat/speakerid.html

8. REFERENCES

[1] B. Delgutte. Representation of speech-like sounds in the discharge patterns of auditory nerve fibers. *JASA*, 68:843–857, 1980.

[2] B. Delgutte and N. Y. Kiang. Speech coding in the auditory nerve:v. vowels in background noise. *JASA*, 75:908–918, 1984.

[3] C. Daniel Geisler. *From Sound to Synapse: physiology of the mammalian ear*. Oxford, 1998.

[4] Jialong He, Li Liu, and Günther Palm. Speaker identification using hybrid lvq-slp networks. In *Proc. IEEE ICNN'95*, volume 4, pages 2051–2055, 1995.

[5] C.R Jankowski Jr., T.F. Quatieri, and D.A. Reynolds. Measuring fine structure in speech: Application to speaker identification. In *IEEE-ICASSP*, pages 325–328, 1995.

[6] R.D. Patterson. Auditory filter shapes derived with noise stimuli. *JASA*, 59(3):640–654, 1976.

[7] A. N. Popper and R. Fay, editors. *The Mammalian Auditory Pathway: Neurophysiology*. Springer–Verlag, 1992.

[8] D.A. Reynolds. The effects of handset variability on speaker recognition performance: Experiments on the switchboard corpus. In *IEEE-ICASSP*, 1996.

[9] J. Rouat. Nonlinear operators for speech analysis. In M. Cooke, S. Beet, and M. Crawford, editors, *Visual representations of speech signals*, pages 335–340. J. Wiley, 1993.

[10] J. Rouat, Y.C. Liu, and D. Morissette. A pitch determination and voiced/unvoiced decision algorithm for noisy speech. *Speech Communication*, 21:191–207, 1997.

[11] H. Murthy *et al.* Robust text-independent speaker identification over telephone channels. *IEEE SAP*, (5):554–568, 1999.

[12] D. Wei and A. C. Bovik. On the instantaneous frequencies of multicomponent am–fm. *IEEE SP Letters*, 5(4):84–86, 1998.

SPEAKER IDENTIFICATION BY COMPUTER AND HUMAN EVALUATED ON THE SPIDRE CORPUS

Hassan EZZAIDI and Jean ROUAT
ERMETIS. DSA. Université du Québec à Chicoutimi. Chicoutimi. Québec. Canada. G7H 2B1.

1. INTRODUCTION

Although many experiments on clean speech report high identification rates for computer systems, results on noisy telephone speech with different handsets are usually too poor for practical identification tasks (noise, limited bandwidth, effect of the channel, telephone handsets variability)[1].

What would be the identification rate of humans in the same conditions? A reference is necessary in order to evaluate the performance of computer systems. The comparison between computer and human has been already made. For a review one can refer for example to the work by Doddington [2]. As the performance of human has been chown to be dependent of the speech nature, we propose to examine the effect of telephone handset variability for text-independent speaker identification of telephone speech. We report human and computer speaker identification with the SPIDRE database.

Section 2 describes the experimental conditions while section 3 and 4 are the results and discussion. Section 5 is the conclusion.

2. EXPERIMENTAL CONDITIONS

2.1 SPIDRE database

Closed set Speaker Identification experiments were performed on a SPIDRE subset of the Swichboard corpus with *matched* and *mismatched* telephone handset conditions. We refer to the *matched* conditions, when (for a same speaker) the training and testing sessions were collected from the same telephone handset. In the *mismatched* conditions, different handsets were used for training and testing sessions.

Based on the pitch frequency (F_0) distribution, we first ran a speaker identification experiment on the 45 speakers of the SPIDRE corpus. We then extracted the most confusable speakers to create a subset of 10 women speakers (Female speakers with similar F_0 distribution). Each speaker has 4 conversations originating from 3 different handsets. The sampling rate is 8 Khz.

2.2 Listening conditions

Sixty pairs of sentences were randomly chosen and played through a Sennheiser HD250 linearII headphone. Ten naïve listeners (one woman and nine men) were asked to tell if the speaker was the same for both sentences. The listeners could not use sex as discrimination criteria. For each sentence, five seconds of speech were played. The listeners are French speaking and most of them could not understand spoken American English. For each pair, the listener had to make four choices:1. certainly the same speaker; 2. probably the same speaker; 3. probably different speakers; 4. certainly different speakers.

2.3 Computer experiments

Speech analysis

The speech is first preemphazised (0.97), then, a sliding Hamming window with a length of 32 ms and a shift of 10 ms is positioned on the signal. Twelve cepstral Mel coefficients, twelve delta Mel cepstral coefficients (computed according to the regression weighting), one log power and one delta log power are then extracted by using a liftering of 22 . Cepstral mean normalization is also performed. The final dimension of the MFCC vectors is 26.

Identification

We use two clustering technics. The first recognizer is based on a nonparametric pattern recognizer. For now, we use the LVQ-SLP as proposed by J. He and al. [3]. Each speaker is characterized by one codebook. The codebook size is the same for all speakers. We performed experiments with codebook sizes of 128, 256 and 512. The second recognizer is based on a parametric estimation of the probability distributions of the MFCC. A Gaussian Mixture Model (GMM) is associated to each speaker (one model for each speaker). For a given speaker, the GMM is supposed to model the statistical distribution of the MFCC. We used models with 32 Gaussian mixtures. We also assumed a diagonal variance matrix for each mixture component and parameters were estimated via the E.M algorithm.

Training and testing

The impact of mismatched and matched conditions is evaluated. In matched condition, the same telephone handset is used for training and testing. One conversation is used for training and the second one for testing. With mismatched handsets, training is performed on 3 conversations (pronounced through 2 handsets) and testing is made on the 4^{th} Conversation coming from the 3^{rd} handset.

Recognition criterion

The tested conversations were divided into fixed block lengths of 10 ms. With the GMM, the log likelihood of each block is computed, whereas, the nearest neighbour algorithm is used for the LVQ-SLP. A speaker is recognized if, for the entire test conversation (all blocks) it has the minimal distance (LVQ-SLP) or the maximum-likelihood (GMM).

3. EXPERIMENTS AND RESULTS

3.1 Listening tests

Table 1 reports rates for intra-speaker (columns 2 and 3) and inter-speaker (column 4) identification. Listeners are reported in column 1. In the matched conditions, one finds an averaged rate of 81%. The variance is significant and is mainly due to listeners L1 and L6. For the mismatched conditions, the recognition rate falls of 11%, with a weaker variance. In the case of the inter-speaker identification it is not possible to verify if the same telephone handset can yield confusions between speakers (speakers declared to be same speaker instead of declaring different) as the database labeling does not include the description of the handset characteristics. It is observed that the identification rates are coherent with those of

columns 2 and 3.

Analysis of the tests clearly shows that the handset has a predominant influence on the perception of listeners. In many situations with the same speaker and two handsets for the two sentences, listeners identified the two conversations as coming from different talkers.

Listeners	Identical speaker in test		Different speakers in test
	In Matched condition	In Mismatched condition	
L1	60 %	68 %	67 %
L2	90 %	74 %	77 %
L3	90 %	72 %	75 %
L4	70 %	72 %	72 %
L5	90 %	72 %	75 %
L6	50 %	64 %	65 %
L7	90 %	71 %	92 %
L8	83 %	62 %	75 %
L9	100 %	72 %	62 %
L10	90 %	72 %	75 %
Mean	81 %	70 %	73.5 %
σ	16	4	8.2

Table 1 : Averaged scores for 10 listeners . Last column, refers to conversation pairs involving two different speakers using unknown handsets.

3.2 Computer tests

Handsets condition	*Codebook size (LVQ-SLP)*			*32 GMM*
	512	*256*	*128*	
Matched	90 %	90 %	90 %	-
Mismatched	60 %	60 %	60 %	90 %

Table 2: Computer speaker recognition rates. *Matched* (One conversation in training, another in testing, identical handset): *Mismatched* (Three conversations in training, another in testing, different handset).

The LVQ-SLP recognizer yields an identification rate of 90% when talkers use the same telephone handset. With different handsets in training and testing (mismatched) the scores drop to 60%. It is interesting to note that the LVQ-SLP rates are independent of the codebook sizes.

The GMM recognizer outperforms the LVQ-SLP recognizer when mismatched handsets are used (90% in comparison to 60%). With the same handset we would expect better results.

4 DISCUSSION

It is observed that the confusion between speakers is mainly due to the strong telephone handset influence. Thus, the speaker acoustical characteristics are found to be largely degraded by the telephone handsets.

Except for L1 and L6, all listeners presented the same faculty to distinguish between the 10 women. In matched conditions, if one does not consider L1 and L6 in the statistics, the average identification rate can increase around 90%.

Interestingly, the difference in performance when changing from matched to mismatched condition is smaller for listeners than for computers.

Although the task presented to listeners and computers is not comparable – the listeners task is easier with a comparison of two speech segments and the computer has to carry out the classification between 10 speakers presented simultanously - it is observed that the mismatched conditions degrade the performance for human and computer.

It is possible to infer that the success of the computer is related to the efficiency of the models and to the quality of the parameters (reduction of the channel and handset effects) for the subset of 10 women.

5 CONCLUSION

Even if the task was easy in comparison to the identification of forty speakers, the relatively low performance of the listeners gives an idea of the complexity of the SPIDRE corpus.

The selection of ten females has been based on pitch frequency distribution. They have a similar distribution of pitch. We already found that based on the pitch, the task was tedious for computers when using exclusively cues derived from the pitch distribution [4]. Manipulation of the pitch frequency confuses listeners when identifying speakers [5]. This suggests that pitch frequency is in fact a fundamental cue which can not be fully exploited by listeners on our test set because of the pitch distribution similarity. Furthermore, Itoh and Saito [6] found that the spectrum envelope is more important in speaker identification than excitation. The MFCC recognizers rely mainly on the spectrum envelope and formants and are probably more accurate than listeners to identify speakers based on spectral characteristics only.

For a subset of 10 female speakers with high confusable pitch distribution the recognizer based on the MFCC and GMM outperforms the listeners.

ACKNOWLEDGEMENT

Mohammed Bahoura wrote the listening tests and performed the listening evaluations. Many thanks are due to the 10 listeners.

REFERENCES

[1] C.R Jankowski Jr., T.F. Quatieri, and D.A Reynolds. Measuring fine structure in speech: Application to speaker In *IEEE-ICASSP*, pages 325-328, 1995.

[2] G. Doddington: Speaker recognition-identifying people by their voices In *Proc. IEEE* Vol 73, pp1651-1664.

[3] Jialong He, Li Liu, and Günther Palm. Speaker identification using hybrid LVQ-SLP networks. In Proc. *IEEE ICNN'95*, volume 4, pages 2051—2055, 1995.

[4] J. Rouat, H. Ezzaidi et M. Lapointe. Nouveaux algorithmes d'extraction en vue de caracteriser le locuteur. *Technical report*, ERMETIS, Universite du Québec a Chicoutimi, Mars 1999. Contrat W2213-9-2234 SL, rapport final, 67 pages.

[5] S. V. Bemis and S. W. Nunn. Acoustic features and human perception of speaker identity In Proc. AVIOS, pages 85-96, 1998.

[6] Itoh K. and Saito S. Effects of Acoustical Feature Parameters on Perceptual Speaker Identity *In Review of the Electrical Communications Laboratories*, vol. 35, N0.1.

Combining pitch and MFCC for speaker recognition systems

Hassan Ezzaidi, Jean Rouat and Douglas O'Shaughnessy[+]

ERMETIS, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada, G7H 2B1
[+]INRS-Télécommunications, Université du Québec
900 de la Gauchetiere west, Box 644, Montreal, Québec, Canada, H5A 1C6

hezzaidi@uqac.uquebec.ca, jrouat@uqac.uquebec.ca, louds@inrs-telecom.uquebec.ca

Abstract

Usually, speaker recognition systems do not take into account the short-term dependence between the vocal source and the vocal tract. A feasibility study that retains this dependence is presented here. A model of joint probability functions of the pitch and the feature vectors is proposed. Three strategies are designed and compared for all female speakers taken from the SPIDRE corpus. The first operates on all voiced and unvoiced speech segments (baseline strategy). The second strategy considers only the voiced speech segments and the last includes the short-term pitch information along with the standard MFCC. We use two pattern recognizers: LVQ-SLP and GMM. In all cases, we observe an increase in the identification rates and more specifically when using a time duration of 500 ms (6% higher).

1. Introduction

The vibration frequency of the vocal folds is known to be an important feature to characterize speech and has been found effective for automatic speech and speaker recognition [1] [2]. An important characteristic of pitch is its robustness to noise and channel distortions. Many parametrizations of pitch such as pitch value, averaged pitch, pitch contour, pitch jitter and location, pitch histograms and prosody [1] [3] [4][5] have been proposed for speaker verification or identification.

1.1. Long-term and short-term pitch

Depending on the time scale, the pitch information can be used differently:

1. Prosody, pitch histograms and pitch evolution are characteristics that are commonly obtained via long term pitch parametrization. These characteristics are known to be complementary with vocal tract parametrization (such as MFCC or LPC) [6]. One can cite for example the work by Sönmez *et al.* [4][5] where it is shown that prosodic variation can be successfully combined with cepstrum coefficients to improve speaker verification systems. Pitch histograms have also been suggested to improve speech verification [4] and identification [7] systems. Speaker recognition systems exclusively based on pitch do well when the number of speakers [1] [6] [7] is small. However, performance decreases significantly when the number of speakers increases, but pitch information can be reliably used to distinguish the sex of speakers [8]. When the number of speakers is small (on the order of 10), pitch can be reliably used to discriminate male speakers. When the number of speakers is greater, pitch has to be combined with other parameters. Furthermore, pitch and MFCC have been shown to be complementary features that can be combined to improve speaker recognition systems.

2. On a short-term time scale, the usefulness of pitch seems to be somehow controversial. We therefore propose a general framework for the study of short-term pitch contributions in order to gain a better understanding of these contributions to speaker recognition.

1.2. Source and vocal tract coupling

In spite of the weak contribution of pitch to contemporary speaker identification research, it remains true that the mechanisms involved in speech production are complex, and imply dependence of articulators and vocal folds, which can be useful for speaker verification or identification.

Most of the models reported in the literature assume the short-term independence of the glottis and the vocal tract. We propose to take into account the correlation between the glottis and the vocal tract. We study the influence of this dependence in the context of a text-independent Speaker Identification System (SIS). We use a joint probability function to take into account the correlation between source and vocal tract. The proposed approach consists of generating models of the feature vec-

This work was funded by NSERC, Communications Security Establishment and the FUQAC. Many thanks to Kurt Boutin for his support and to the anonymous reviewers for constructive comments.

tors for each pitch range.

The next section describes the motivation for this research. The baseline and the proposed systems are described in sections 5 and 7. Sections 8 and 9 present the results, discussion and conclusion.

2. Motivation

Most systems use long-term or short-term parameters that should encode vocal tract features, but contributions of the glottis to these features are largely ignored. Even if MFCC (Mel Frequency Cepstrum Coefficients) are theoretically known to deconvolve the source and the vocal tract; in practice, cepstrum coefficients are affected by high pitched voices (women and infants).

One can illustrate the role of pitch when dependence of the source and the vocal tract are maintained. Figure 1 exhibits four spectrograms and pitch histograms; each column corresponds to a different male speaker, obtained from the YOHO database. All speakers pronounced the same digit utterance "twenty six". The pitch range is divided into 56 equal bins of 10 Hz width. The spectrograms show a significant similarity of formant distributions between speakers. The spatial distribution of formants depends on the interspeaker variability as described in [9]. However, the pitch histograms are different and vary from one speaker to another for the same context. If one compares the histograms by taking into account their frequency amplitude and width, it is observed that speaker 2 from the second column and speaker 4 from the fourth column do have a similar pitch distribution. On the

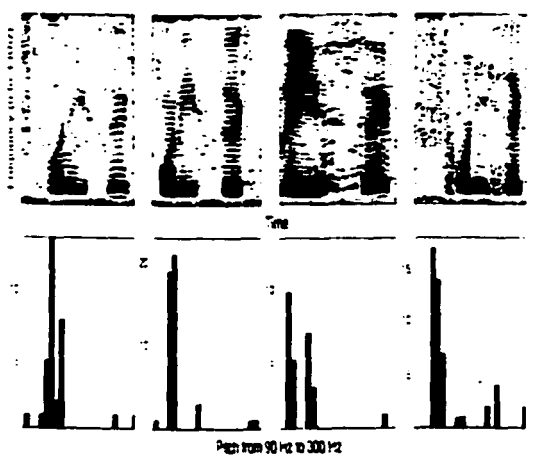

Figure 1: For each of four male speakers: Spectrograms and pitch histograms for the same English digit utterance "26".

other hand, speakers 1 and 3 are characterized by dissimilar pitch histograms. Consequently, if one takes into consideration the pitch information, the interspeaker variability can be restricted to speakers with similar pitch distributions, and the other speakers will be considered as belonging to other clusters. Speakers with similar pitch will be recognized based on their spectral characteristics.

In summary, short-term pitch and vocal tract features can be jointly exploited in order to establish probability models of feature vectors assuming the a priori knowledge of the pitch distribution.

3. Proposed Model

3.1. Theoretical framework

We suppose that pitch and vocal tract features are two random processes respectively denoted as $X(n)$ and $Y(n)$. Let's write $\widehat{X(n)}$ the estimated discretized pitch frequency at time $n\Delta t$ and $\widehat{Y(n)}$ the estimated discretized vocal tract feature vector at time $n\Delta t$. $\widehat{Y(n)}$ is an r-dimensional vector. In practice $\widehat{Y(n)}$ is an LPC_r or $MFCC_r$ vector estimated from a centered signal window at time $n\Delta t$. For each process $\widehat{X(n)}$ and $\widehat{Y(n)}$, we assume time independence of their respective realizations. As a consequence, $\widehat{X(n-1)}$ is supposed to be independent of the realization of $\widehat{X(n)}$. The same restriction applies to $\widehat{Y(n)}$. In the following, we drop the time and consider the simultaneous realization of $\widehat{X(n)}$ and $\widehat{Y(n)}$ as being time independent. The crosscorrelation between $\widehat{X(n)}$ and $\widehat{Y(n)}$ is still preserved.

Let us write $\{x_1, x_2, \ldots, x_n\}$ the increasing sequence of realizations of $\hat{X}$, with $x_1 \in [63\ Hz, 300\ Hz]$. For simplification purposes, we assume that the set of realizations of $\hat{Y}$ is finite (by using vector quantization and codebooks, for example) and equal to $\{\overrightarrow{y_1}, \overrightarrow{y_2}, \ldots, \overrightarrow{y_m}\}$, with $\overrightarrow{y_i} \in R^r$. Let f be the joint probability of $\hat{X}$ and $\hat{Y}$.

$$f(x_i, \overrightarrow{y_j}) = P(\hat{X} = x_i, \hat{Y} = \overrightarrow{y_j}) \text{ with} \quad (1)$$

$$0 \leq f(x_i, \overrightarrow{y_j}) \leq 1 \text{ and } \sum_{i=1}^{n}\sum_{j=1}^{m} f(x_i, \overrightarrow{y_j}) = 1 \quad (2)$$

The respective marginal probability functions are:

$$f(x_i) = \sum_{j=1}^{m} f(x_i, \overrightarrow{y_j}) \text{ and } f(y_j) = \sum_{i=1}^{n} f(x_i, \overrightarrow{y_j}). \quad (3)$$

Each speaker s is supposed to be defined by its probability function f_s that takes into account the coupling between X and Y:

$$f_s(x_i, \overrightarrow{y_j}) = P_s(\hat{X} = x_i, \hat{Y} = \overrightarrow{y_j}). \quad (4)$$

We observe that

$$f_s(x_i, \overrightarrow{y_j}) = f_s(\overrightarrow{y_j}/x_i) f_s(x_i). \quad (5)$$

$f_s(x_i)$ is the *a priori* probability of a pitch frequency to be equal to x_i and $f_s(\overrightarrow{y_j}/x_i)$ is the *a posteriori* probability of observing a feature vector to be equal to $\overrightarrow{y_j}$ given the

knowledge of the pitch frequency x_i. The estimation of the a priori probability of the pitch frequency is relatively straightforward while the estimation of $f_s(\overrightarrow{y_j}/x_i)$ can be long and tedious.

3.2. Feature vector distributions based on pitch knowledge

In the present work we focus on the estimation and integration of the posteriori probability, $f_s(\overrightarrow{y_j}/x_i)$, in speaker recognition systems. The consideration of the factor $f_s(x_i)$ from equation 5 is left as a future work.

We propose to subdivide the space $(x, \overrightarrow{y})$ into subspaces H_k where $f_s(\overrightarrow{y_j}/x_i)$ is supposed to be locally independent of the pitch value. Let us define I_k, $k = 1, \ldots, N$ as sub–intervals of the pitch set $\{x_1, x_2, \ldots, x_n\}$. We recall that $x_1 = 60$ Hz and $x_n = 500$ Hz. N is the number of intervals with $I_1 \cup \ldots \cup I_N = \{x_1, x_2, \ldots, x_n\}$. Each subspace H_k is associated to a pitch interval I_k. For each H_k, we suppose that the probability function $f_s(\overrightarrow{y_j}/x_i)$ is stationary and independent of the pitch inside the interval I_k (in other words, inside an interval I_k the pitch frequency is supposed to be the same)[1], that is:

$$f_s(\overrightarrow{y_j}/x_i) = P(\overrightarrow{Y} = \overrightarrow{y_j}/I_k, \text{ Speaker} = s \text{ with } x_i \in I_k) \quad (6)$$

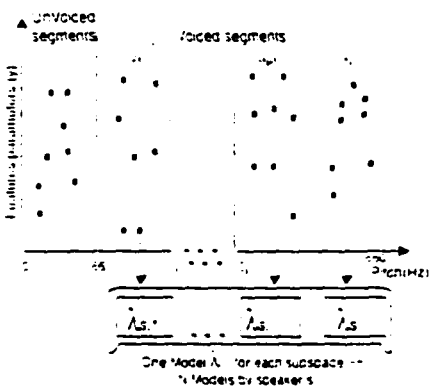

Figure 2: Proposed approach for generating sub–models.

Theoretically, the number of models $f_s(\overrightarrow{y_j}/I_k) = \lambda_{s,k}$ would be equal to n. By subdividing the space into N subspaces, we reduce that number to N. Figure 2 illustrates the notion of subspaces and models of probability functions $f_s(\overrightarrow{y_j}/I_k)$. The interval length of I_k is based on the shape of the pitch histogram (section 7.3).

4. Speech Analysis

Mel Cepstrum Coefficients derived from a bank of filters (MFCC) are used as features to characterize the identity of speakers. We use coefficients c_1 to c_{12}. The speech is first preemphasized (0.97); then, a sliding Hamming window with a length of 32 ms and a shift of 10 ms is positioned on the signal. Cepstral mean normalization and liftering are also performed. Delta and delta-delta MFCC are not used, as the comparison between the systems would be biased. In fact, adjacent segments can have different pitch values belonging to different sub–intervals I_k.

5. Pattern recognition

5.1. Framework

Two pattern recognizers are used for the experimental task: one parametric and one non-parametric.

5.1.1. Parametric model

We use a Gaussian Mixture Model (GMM) [10] with a weighted sum of 32 ($M = 32$) Gaussians. Each GMM is defined for a specific speaker s and pitch interval I_k. Let us define $p(\overrightarrow{y}, \lambda_{s,k})$, the Gaussian mixture density associated with the probability function $f_s(\overrightarrow{y}, I_k)$ for speaker s, as

$$p(\overrightarrow{y}, \lambda_{s,k}) = \sum_{i=1}^{M} w_{i,s,k} b_{i,s,k}(\overrightarrow{y}) \quad (7)$$

with

$$b_{i,s,k}(\overrightarrow{y}) = \frac{1}{(2\pi)^{d/2} |\Sigma_{i,s,k}|^{1/2}} e^{-\frac{1}{2}(\overrightarrow{y} - \overrightarrow{\mu_{i,s,k}})' \Sigma_{i,s,k}^{-1} (\overrightarrow{y} - \overrightarrow{\mu_{i,s,k}})}$$

M is the GMM order, $\overrightarrow{y}$ is the d-dimensional vector estimating the vocal tract contribution (MFCC vector), $b_{i,s,k}$ is the i-th Gaussian density with mean $\overrightarrow{\mu_{i,s,k}}$ and covariance matrix $\Sigma_{i,s,k}$, and $w_{i,s,k}$ are the mixture weights. $\overrightarrow{\mu_{i,s,k}}$, $\Sigma_{i,s,k}$ and $w_{i,s,k}$ are defined for pitch interval I_k and for speaker s. Each speaker is characterized by N models $\lambda_{s,k}$ corresponding to N pitch intervals I_k.

5.1.2. Non–parametric model

We use the hybrid LVQ-SLP network as proposed by He *et al.* [11]. Each speaker s, with a pitch belonging to I_k, is characterized by a codebook $C_{s,k}$. The codebook size is the same for all speakers. We performed experiments with codebook sizes of 512 for each speaker.

5.2. Recognition

5.2.1. Parametric model

We define T as being the test length over which the recognition is performed. A frame-by-frame estimation of log–likelihood for each speaker s and pitch interval I_k is first performed. Each frame (32 ms length) is shifted by 10 ms. Then, the maximum log–likehood for each speaker is estimated over T. When the test sentence is longer than T,

[1] this restriction can be suppressed by assuming $N = n$

the average of the score over the number of segments with a length of T is computed according to equation 8.

$$S_T = \frac{\text{nb. of seg. correctly tested for } T \text{ duration}}{\text{total nb. of seg. tested for } T \text{ duration}} \quad (8)$$

The final identification score (equation 9) is obtained by averaging over the number of speakers N_s:

$$S = \frac{\sum_{T} S_T}{N_s} \quad (9)$$

5.2.2. Non-parametric model

For each frame, the feature vector is classified by using the Nearest Neighbor criteria. A speaker is recognized if, for the entire test conversation, it is selected more frequently than the other speakers.

6. Speech database

A subset of the SPIDRE–Switchboard Corpus is used and comprises the eighteen (18) female speakers of the database. Each speaker has 4 conversations originating from 3 different handsets. The training data comprises 3 conversations, with 2 conversations coming from the same handset. The last conversation, using the third handset (different from the others), is presented as the test data. This combination is referred to as the *mismatched condition*. The *matched condition* refers to situations where training and testing data are recorded from the same handset.

7. Strategies

7.1. The baseline strategy

The baseline strategy uses both the voiced and unvoiced segments. The suppression of silence was carried out based on the energy evolution and comparison with fixed thresholds.

7.2. Recognition based on voiced speech segments

We include a module that estimates the pitch and selects the voiced segments. We use a pitch tracker and a voiced-unvoiced detection system [12] in conjunction with the SID system analysis module. In this case, silence and unvoiced segments are automatically rejected. During training and for each pitch period, we centered a 32 ms duration window and extracted the MFCC coefficients.

7.3. Recognition based on the estimated a posteriori probabilites

For the third strategy, four pitch intervals I_1.....I_4 are created according to the pitch frequency histogram. More than 90% of the pitch frequencies belong to the interval [150Hz,220Hz]. We distributed the pitch frequencies over 4 intervals I_1=[150,180], I_2=[170,200], I_3=[190,220] and I_4=[63,150]∪[220,600]. The choice of four intervals is a trade off between fine pitch intervals and sufficient training size of the models. During training and for each interval I_k, the MFCC vectors are used to generate model parameters for each speaker. Therefore each speaker is characterized by 4 models. With the aim of overcoming the pitch estimation errors, we choose an overlap of 10 Hz between the intervals. Thus, the MFCC vectors from speech whose fundamental frequency belongs to two adjacent intervals (I_k, I_{k+1}), will be used to train two models, respectively, associated to subspaces H_k and H_{k+1}. Then, during the testing session, the evaluation is carried out over these two subspaces and we keep the best score.

In the case of LVQ–SLP, the codebook generation is made according to two procedures. One attributes the same codebook size to each subspace, and the other distributes the number of prototypes per codebook according to the number of events in each subspace.

In the case of the GMM models, one model λ_1 is generated for the baseline system, one model is also used for recognition on voiced speech and four models $\lambda_{1,k}$ are generated for the recognition taking into account the a posteriori probabilities of voiced speech according to the pitch.

8. Results and discussion

8.1. Evaluation with a LVQ–SLP model

LVQ–SLP results for 18 women of the SPIDRE database are reported in tables 1 and 2.

Table 1: LVQ–SLP: Identification rate increases for 18 female speakers with fixed codebook sizes. Baseline system: 55%.

	Voiced (512)	H_1 (128)	H_2 (128)	H_3 (128)	H_4 (128)
Matched	6	14	10	1	0
Mismatched	3	4	4	5	2

Table 2: LVQ–SLP: Identification rate increases for 18 female speakers with codebook sizes proportional to the number of events in each subspace. Baseline system: 55%.

	H_1	H_2	H_3	H_4
Matched	14	3	0	-1
Mismatched	4	5	6	1

When the unvoiced segments are not taken into account (Voiced column), the identification rate increases to 61% (6% more in table 1) for matched handsets and to 58% (3%) for mismatched handsets. When pitch is taken into account (columns H_k in tables 1 and 2), the increase is almost double in H_1 and H_2 and weaker in H_3 and H_4.

H_1 and H_2 are the subspaces with the greatest number of events and H_3 and H_4 with the smallest number of events. When the number of prototypes per codebook is proportional to the number of events, performance falls in H_2 and H_3 and remains constant in H_1.

When recognition rates are weighted according to the number of events per subspace, we obtain an averaged increase of 8% in matched conditions, and 4% in mismatched situations. It is observed that the identification results are sensitive to several factors: 1) codebook sizes, 2) training techniques, 3) score combination, and 4) pitch estimation. The best increase in performance is observed for subspaces with the greatest number of events.

8.2. Evaluation with a GMM model

Table 3 reports the identification results observed with the three strategies: 1) Baseline (voiced and unvoiced segments), 2) Voiced (only voiced segments) and 3) Voiced segments with partition of space into H_1 to H_4. The first column gives the value of T, that is, the duration of maximum log-likehood estimation.

Table 3: GMM: Mismatched identification rates for 18 female speakers.

Time (seconds)	Baseline (%)	Voiced (%)	Voiced & pitch (%)
0.1	56.3	37.5	40.5
0.5	83.4	86.7	89.4
1	85.4	90.3	90.5
2	94.2	97.4	98.0
3	98.2	99.5	99.5
4	99.2	99.4	99.3
5	99.4	99.5	99.7
6	99.7	99.8	99.8

The baseline strategy yields the lowest identification rates. When voiced segments are used, the best increase is 4.5% and the weakest is 1%. When a preliminary subdivision based on pitch is performed, the greatest increase is 6% and the weakest is 2.5%.

When the test duration is greater than 2 seconds, strategies based on voiced segments yield similar results. When T is less than 2 seconds, the best results are observed with the strategy that takes into account the posterior probability (subdivision into subspaces). The increase is on the order of 2%.

8.3. Discussion

Identification rates of LVQ-SLP and GMM are not strictly comparable, as the recognition criteria are different. The comparison of Tables 1 and 2 suggests that the a priori probability $f_{\Omega}(\omega)$ should be taken into account. In Table 3, a T of 1 second is equivalent to 100 MFCC vectors and is independent of the strategy. The weaker performance of the baseline system might be partially due to the smaller number of voiced frames in a fixed T.

When the dependence of the source and vocal tract is taken into account, the best results are observed for durations T lower than 3 seconds (up to 4.5% for $T = 500$ ms). For $T \geq 3$ seconds scores are 1% higher, in favour of the system based on voiced segments only.

It has been observed that the size of the corpus and data training can have a strong influence on performance. In fact, the best recognition rates are observed for subspaces H_4 with the highest number of occurences.

9. Conclusion

A new approach that preserves the dependence between the vocal source and the vocal tract has been proposed. Experiments that integrate the a posteriori probability of observing a MFCC vector given the knowledge of the pitch frequency have been reported. They are compared with a baseline system operating on all voiced and unvoiced speech segments and with a second system that operates on voiced speech segments only. Closed set Speaker Identification experiments were performed on a subset of 18 female speakers taken from the SPIDRE corpus that comprises highly confusable speakers. Results have been reported and speaker identification comparisons have been performed by using the short-time pitch in order to take into account its contribution to the cepstrum coefficients. In fact, it is known that the MFCC do not really deconvolve the source from the vocal tract for high pitch voices.

There is a significant increase in performance when the recognition uses a short window length T as a decision duration. When the system can rely on a longer decision window, the improvement is not that significant. It is suspected that a greatest time integration can compensate for local decision errors. We recall that a local likelihood is obtained for every 10 ms and T is the overall interval over which the log-likelihood are combined.

Many restrictive hypotheses have been made to set up the experiments. The pitch tracker has been supposed to be reliable; sufficient training data for subspace decomposition; local independence of MFCC in relation to pitch in a subspace (equation 6); and time independence of pitch and MFCC have been assumed. Most of these hypotheses are not really realistic. For example, in several cases, the pitch is not well estimated (double and half pitch) and affects the performance. If these errors are taken into account, we can possibly achieve better training and evaluation.

We also suggest, as future work, to increase the size of the corpus for a better statistical convergence, to optimize the number and width of the pitch intervals $\Delta\omega_i$ and to introduce weighting by the a priori probability distribution of the pitch ($f_{\Omega}(\omega)$) in accordance with equation 5. We also suggest restricting the application to a text-dependent or verification system, for which the variability of the parameters is usually smaller.

10. References

[1] B. S. Atal. "Automatic recognition of speakers from their voices." in *Proc. IEEE*. 1976. vol. 64. pp. 460–475.

[2] Douglas O'Shaughnessy and Hesham Tolba. "Towards a robust/fast continuous speech recognition system using a voiced-unvoiced decision." pp. 413–416. ICASSP. 1999.

[3] C.R Jankowski Jr.. T.F. Quatieri. and D.A. Reynolds. "Measuring fine structure in speech: Application to speaker identification." in *IEEE-ICASSP*. 1995. pp. 325–328.

[4] Kemal Sönmez. Larry Heck. Mitchel Weintraub. and Elisabeth Shriberg. "A lognormal tied mixture model of pitch for prosody-based speaker recognition." in *Proc. of EUROSPEECH*. september 1997. pp. 1391-1394.

[5] Kemal Sönmez. Elisabeth Shriberg. Larry Heck. and Mitchel Weintraub. "Modeling dynamic prosodic variation for speaker verification." in *Proc. of the International Conference on Spoken Language Processing*. 1998. pp. 3189-3192.

[6] Douglas O'Shaughnessy. *Speech Communications: Human and Machine*. IEEE Press. 2000.

[7] Hassan Ezzaidi. Tuong Vinh Ho. Mathieu Lapointe. and Jean Rouat. "Nouveaux algorithmes d'extraction liés aux caractéristiques de la parole destinés à l'identification du locuteur." Tech. Rep.. ERMETIS. Universit du Qubec Chicoutimi. July 1998. Contrat. W2213-7-2005/001/SV. rapport de progrs n04. 104 pages.

[8] J.Rouat. H. Ezzaidi. and M. Lapointe. "Nouveaux algorithmes d'extraction en vue de caractériser le locuteur.." Tech. Rep.. March 1999. Contrat W2213-9-2234/SL. 67 pages.

[9] G. R. Doddington. "Speaker recognition-identifying people by their voices." in *Proc. IEEE Vol.73, No 11*. 1985, number 11. pp. 1651–1664.

[10] Douglas A. Reynolds and Richard C. Rose. "Robust text-independent speaker identification using gaussian mixture speaker models." vol. 3. no. 1. pp. 72–83. 1995.

[11] J. He. L. Liu. and G. Palm. "Speaker identification using hybrid lvq-slp networks." in *Proc. IEEE ICNN Vol.4*. 1995. pp. 2051–2055.

[12] J. Rouat. Y.C. Liu. and D. Morissette. "A pitch determination and voiced/unvoiced decision algorithm for noisy speech." *Speech Communication*. vol. 21. pp. 191–207. 1997.