N° d'ordre : 11/2018- D/ELC

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediene

Faculté d'Electronique et Informatique

THÈSE de Doctorat en science

Présentée pour l'obtention du grade de Docteur

EN : ELECTRONIQUE

Spécialité : Télécommunication

Par : **FEDILA Meriem**

Intitulé :

Amélioration de la Robustesse des Systèmes de Reconnaissance Automatique du Locuteur

Soutenue publiquement le 20/12/2018, devant le jury d'examen composé de :

Mme	**FALEK Leïla**	**Professeur à l'USTHB**	**Présidente**
M	**AMROUCHE Abderrahmane**	**Professeur à l'USTHB**	**Directeur de Thèse**
M	**HAMZA Abdelkrim**	**Maître de conférences/ A l'USTHB**	**Examinateur**
Mme	**HAMAMI Latifa**	**Professeur à l'ENP**	**Examinatrice**
Mme	**BENBLIDIA Nadjia**	**Professeur à l'USD Blida1**	**Examinatrice**
M	**BENSELAMA Zoubir**	**Professeur à l'USD Blida1**	**Examinateur**

Dédicace

Je dédie cette thèse…

A mes très chers parents Bachir et Nacera

Vous m'avez comblé avec votre tendresse et affection tout au long de mon parcours. Vous n'avez cessé de me soutenir et de m'encourager durant toutes les années de mes études, vous avez toujours été présents à mes côtés. En ce jour mémorable, pour moi ainsi que pour vous, recevez ce travail en signe de ma vive reconnaissance et ma profonde estime.
Puisse Dieu, le tout puissant, vous préserver et vous accorder santé, longue vie et bonheur.

A mon très cher mari Abdellah

Tes sacrifices, ton soutien moral et matériel, ta gentillesse sans égal, ton profond attachement, ta présence, ton amour et ta tendresse m'ont permis de réussir mes études. Merci d'être toujours à mes côtés. Que ce travail soit témoignage de ma reconnaissance et de mon amour sincère et fidèle.

A ma chère petite fille Tasnim

Je t'aime mon ange. Que dieu, le tout puissant, te protège mon bébé.

A ma très chère sœur Aicha, mon cher petit frère Abdellah ainsi mon frère Mouad et son épouse Zineb

Ce travail est un témoignage de mon attachement, mon amour et mon affection que je porte pour vous. Que Dieu vous protège et vous prête bonne santé et longue vie.

A mes neveux Abderahmene, Abderahim et le prince Amir

Je vous dédie ce travail avec tous mes vœux de bonheur, de santé et de réussite.

A mes beaux-parents DJEBALI Abdelhafid et Najia

Je vous dédie ce travail avec tous mes vœux de bonheur, de santé et longue vie.

A ma chère grand-mère Dahbia

Puisse Dieu, le tout puissant, vous préserver et vous accorder santé, bonheur et longue vie.

A mes amies et collègues

En témoignage des souvenirs qui nous uni, je vous dédie ce travail et je vous souhaite une vie pleine de santé et de bonheur.

A tous les membres de ma famille, petits et grands.

Veuillez trouver dans ce modeste travail l'expression de mon affection.

Remerciement

Je tiens à remercier tout d'abord *ALLAH* le tout-puissant, de m'avoir donné autant de courage, de patience et de volonté afin d'accomplir ce travail. *EL-hamdou li Allah.*

J'adresse mes plus vifs remerciements à mon Directeur de thèse, Monsieur *Abderrahmane AMROUCHE*, Professeur à la Faculté d'Electronique et d'Informatique, Université des Sciences et de la Technologie Houari Boumediene USTHB, d'avoir dirigé ce travail avec compétence et professionnalisme. Je le remercie pour tous ses conseils précieux et remarques pertinentes. Ce travail est l'occasion de vous témoigner ma profonde gratitude.

Je tiens à remercier également Madame *FALEK Leïla,* Professeur à la Faculté d'Electronique et d'Informatique, USTHB, pour m'avoir fait l'honneur de présider ce jury de thèse.

Mes remerciements s'adressent aussi à Monsieur *HAMZA Abdelkrim* Maître de conférences à la Faculté d'Electronique et d'Informatique, USTHB, Madame *HAMAMI Latifa*, Professeur à l'Ecole Nationale Polytechnique (ENP), Madame *BENBLIDIA Nadjia*, Professeur à l'Université Saâd Dahleb de Blida1 (USDB1) et Monsieur *BENSELAMA Zoubir*, Professeur à l'USDB1, qui ont bien voulu accepter d'évaluer le présent travail. Je les remercie pour le temps qu'ils y consacrent et je souhaiterai qu'ils y trouvent entière satisfaction.

Je voudrais également remercier Monsieur *BENGHERABI Messaoud* qui m'a donné l'opportunité de travailler au sein de son équipe ‘‘Traitement de signal pour les systèmes Biométriques, la Sécurité Multimédia et la Criminalistique’’ (BIOSMC) au Centre de recherche de Développement des Technologies Avancées (CDTA), Je lui suis reconnaissante pour ses qualités scientifiques, ses précieux conseils ainsi de m'avoir permis d'améliorer mes capacités scientifiques.

Mes remerciements vont également à *ma famille* pour leur soutien et leur affection.

Enfin je ne pourrais pas terminer ces remerciements sans une pensée à l'ensemble de *mes Enseignants* des cycles primaire, moyen, secondaire et supérieur qui sont à l'origine de mon savoir.

Que Dieu vous bénisse TOUS.

ملخص

إن الهدف من هذا البحث هو جعل انظمة التعرف الالي على المتحدث أقل حساسية بالمؤثرات التي تطرأ على المحيط الخارجي أين لا تكون ظروف التسجيل تحت السيطرة وهذا باقتراحنا معاملات جديدة حيث قمنا باستغلال العلاقة بين طيف فرق الطور و طيف الشدة و أيضا مرشح غاماتون . قمنا بدراسة المعاملات المحصل عليها GPSCC و ذلك باستعمال بنك المعلومات الصوتي TIMIT و النموذج الأكثر إستعمالا في هذا الميدان GMM. النتائج المحصل عليها تبين أن المعاملات الجديدة تحسن من أداء نظام التعرف وذلك في ظروف محيطية صعبة ليس فقط بوجود تشويشات مختلفة ولكن أيضاً في ظروف تطبيقات المحمول حيت سجلنا تخفيض في الخطأ المطلق يقدر ب 12% مقارنةً مع استعمال المعاملات المرجعية MFCC . من جهةٍ أخرى قمنا باعتماد طريقة جديدة للكشف عن نشاط الصوت VAD تعتمد على إستعمال تقنية SVM. نتائج الطريقة الجديدة LSVM-VAD المتحصل عليها باستعمال بنك المعلومات الصوتي MOBIO ونمذجة I-vector تتقارن بشكل إيجابي مع أفضل الانظمة المعروضة في مسابقة ICB2013 SRE .

الكلمات المفتاحية: نظام التعرف الالي على المتحدث، مرشح غاماتون، طيف فرق الطور، المنتج الطيفي ، VAD، GMM-UBM، I-vector ، G722.2

Résumé

Le travail de cette thèse propose d'améliorer la robustesse des systèmes de Reconnaissance Automatique du Locuteur pour les applications en environnement réel, où les conditions ne sont pas contrôlées. Deux contributions ont été développées : i) La première concerne la proposition de nouveaux paramètres, exploitant la relation symbiotique entre la phase à court terme, le spectre d'amplitude et les bancs de filtres Gammatones. Ces nouveaux paramètres, appelés GPSCC (Gammatone Product Spectrum Cepstral Coefficients), ont été testés sur la base de données TIMIT en utilisant les Modèles de Mélange de Gaussiennes GMM. Ces paramètres présentent une grande robustesse à des conditions de bruit sévères, et de même pour un environnement mobile, avec une réduction d'erreur absolue de 12% par rapport à l'utilisation des paramètres de référence MFCC. ii) D'autre part, nous avons implémenté une nouvelle technique de détection d'activité vocale VAD basée sur l'utilisation des Support Vecteur Machine SVM. La méthode proposée, nommée LSVM-VAD présente de bonnes performances comparativement aux récents systèmes présentés dans le concours ICB2013 SRE (Speaker Recognition Evaluation) sur la base de données MOBIO en utilisant la technique I-vecteur.

Mots clés : Vérification Automatique du Locuteur, Filtres Gammatones, Retard de groupe, Produit Spectral, VAD, GMM-UBM, I-vecteur, G722.2

Abstract

In order to improve the robustness of Automatic Speaker Recognition systems for real-world applications, we have proposed new features exploiting the symbiotic relationship between the short-term phase, the spectrum amplitude and Gammatone filter banks. The proposed features, namely PSGCC (Gammatone Product Spectrum Cepstral Coefficients), were tested on the TIMIT database using Gaussian Mixture Models GMM. The PSGCC features are very robust in adverse conditions and mobile environment, with an absolute error reduction of 12% compared to the use of the much known MFCC parameters. The second contribution concerns a new VAD speech activity detection technique based on the use of SVM vector support. The proposed LSVM-VAD method compares favorably with the best systems presented in the ICB2013 SRE (Speaker Recognition Evaluation) competition on the MOBIO database using I-vector.

Keywords: Automatic speaker verification, Gammatone filter-bank, Group-delay, spectral product, VAD, GMM-UBM, I-vector, G722.2.

Table des matières

Introduction générale

Chapitre 1. La Reconnaissance Automatique du Locuteur

Chapitre 2. La modélisation des vecteurs acoustiques

Chapitre 3. Nouveaux paramètres acoustiques pour la reconnaissance robuste

Chapitre 4. Expérimentation I: Evaluation du système de vérification du locuteur par les GPSCC et les MGCC

Chapitre 5. Expérimentation II: Nouvelle technique de Détection d'Activité Vocale

Conclusions et perspectives

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABREVIATIONS

Les abréviations dans cette thèse sont citées en anglais.

AMR-WB	Adaptatif Multi-Rate WideBand
CELP	Code Excited Linear Prediction
CMN	Cepstral Mean Normalization
DCF	Detection Cost Function
DCT	Discrete Cosine Transform
DFT	Discret Fourier Transform
DET	Detection Error Tradeoff
DTW	Dynamic Time Warping
EER	Equal Error Rate
EM	Expectation Maximization
ERB	Equivalent Rectangular Bandwidth
FA	Factor Analysis
FAR	False Acceptance Rate
FRR	False Rejection Rate
GMM	Gaussian Mixture Model
GPLDA	Gaussian Probabilistic Linear Discrimination Analysis
GPSCC	Gammatone Product Spectral Cepstral Coefficient
HMM	Hidden Markov Model
IAL/ASI	Automatic Speaker Identification
I-vector	Identity -vector
JFA	Joint Factor Analysis
LBG	Linde–Buzo–Gray
LDA	Linear Discrimination
LSVM-VAD	Linear Support Vector Machine Voice Activity Detection
MAP	Maximum A Posteriori
MFCC	Mel Frequency Cepstral Coefficient
MFPSCC	Mel Frequency Product Spectral Cepstral Coefficient
MGCC	Multitaper Gammatone Cepstral Coefficient

MSR	Microsoft Research
NIST	National Institute of Standards and Technology
PLDA	Probabilistic Linear Discriminant Analysis
QV	Quantification Vectorielle
ROC	Receiver Operating Characteristic
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
SNR	Signal Noise Rate
SVM	Support Vector Machine
S-Norm	Symetric Normalization
TIMIT	Texas Instruments-Massachusetts Institute of Technology
UBM	Universal Background Model
VAL/ASV	Automatic Speaker Verification
VAD	Voice Activity Detection
VQVAD	Vector quantization Voice Activity Detection

Introduction générale

Introduction

Plus qu'un moyen naturel de communication entre les individus, la parole transporte également des informations physiologiques et comportementales sur l'identité du locuteur ayant prononcé le message. Les humains se servent de ces informations pour identifier les personnes qu'ils connaissent, et c'est la raison pour laquelle les grandes entreprises se tournent aujourd'hui vers l'authentification vocale et, en particulier, la Reconnaissance Automatique du Locuteur (RAL).

La reconnaissance par la voix fait toujours l'objet de travaux de recherches entrepris par de nombreuses équipes de recherches dans le monde. Elle présente certains avantages par rapport aux autres technologies biométriques, tels que les empreintes digitales, l'iris, la rétine et l'ADN. Elle peut être utilisée pour authentifier l'accès à distance à des bases de données ou à des réseaux informatiques où l'utilisateur n'est pas physiquement présent. La reconnaissance du locuteur est utilisée par des banques pour fournir une authentification vocale rapide dans les services bancaires téléphoniques et les transactions par carte de crédit. Ainsi seulement un microphone est nécessaire, pour l'acquisition d'un signal de parole, pour le contrôle d'accès physique afin de sécuriser des zones telles que les aéroports et les bureaux, alors que d'autres technologies biométriques nécessitent des équipements spéciaux comme les scanners d'empreintes digitales et les scanners rétiniens.

À vrai dire, la reconnaissance du locuteur est l'une des sous-disciplines du traitement de la parole. Elle se divise à son tour en plusieurs tâches, soit la vérification, l'identification et le regroupement en locuteurs.

Historiques sur la Reconnaissance Automatique du locuteur

Depuis plus de 40 ans, la reconnaissance du locuteur est devenue un domaine actif de traitement de la communication parlée. La reconnaissance du locuteur par les humains a été largement étudiée. La motivation de ces études a été d'apprendre comment l'homme reconnait les locuteurs. Dans les années 1960, des chercheurs ont introduit le spectrogramme en tant que moyen d'identification personnelle et en effet, il a stimulé la recherche sur la reconnaissance du locuteur par ordinateur dans les années 1970 et devient la reconnaissance automatique du

locuteur. A cette époque, les techniques de prédiction linéaire, la transformée de Fourier et les techniques d'analyse cepstrale ont été appliquées pour générer des paramètres du locuteur, pour des systèmes de reconnaissance du locuteur à petite échelle (< 20 locuteurs).

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

La mise à disposition de bases de données parole plus importantes (par exemple, corpus YOHO), dans les années 1990, a boosté les études sur des modèles plus compliqués par exemple, les Modèles de Markov Caches (HMM), le Modèle de Mélange de Gaussiennes (GMM), les Réseaux de Neurones et les Machines à Vecteurs de Support (SVM), etc. Le Modèle de Mélange de Gaussiennes a été considéré comme la technique de modélisation dominante pour les systèmes de reconnaissance du locuteur à cause de son efficacité à caractériser la distribution de la densité des données de la parole. En ce qui concerne l'extraction des caractéristiques, les coefficients cepstraux incorporant le modèle auditif, connus sous le nom de Coefficients Cepstraux à Fréquence Mel (MFCC) et leurs coefficients dynamiques ont été considéré comme les paramètres dominants. A partir des années 2000, des techniques de normalisation des scores ont également été étudiées pour la reconnaissance du locuteur robuste. Un système avec les paramètres MFCC, une modélisation GMM et le modèle du monde universel (UBM) pour la normalisation des scores, a été considéré comme celui qui obtient les meilleurs résultats et est accepté comme un système de base pour comparer les nouvelles technologies.

Motivation et objectif de cette thèse

L'état de l'art des systèmes de reconnaissance du locuteur fonctionnent très bien dans des conditions contrôlées des laboratoires. Les résultats expérimentaux montrent que la reconnaissance automatique du locuteur dans un environnement idéal atteint un niveau de performance aussi bon que celui de la reconnaissance par les êtres humains. Toutefois, comme une technique orientée vers les applications du monde réel, où les conditions ne sont pas contrôlées, les performances des systèmes actuels de reconnaissance du locuteur sont loin d'être robustes et fiables. En effet 1) Les variations inter-locuteurs (qui démontre les différences du

signal de parole en fonction du locuteur) 2) les variations intra-locuteurs (conditionnée par l'état physique et émotionnel du locuteur), et 3) les variations des conditions d'enregistrement du signal de parole (le transcodage de la parole, type de microphone, bande passante du canal de transmission, environnement bruité) peuvent dégrader considérablement les performances d'un système de reconnaissance automatique du locuteur.

Notre principal objectif dans cette thèse est donc d'étudier les moyens qui peuvent rendre le système de reconnaissance insensible aux changements des conditions environnementales en précédant le moteur de reconnaissance par des modules de traitement du signal capables d'améliorer le signal de parole à l'entrée du système de reconnaissance. Plus particulièrement, nous entamons deux facteurs limitant la robustesse des systèmes de vérification de locuteurs. Le premier facteur à traiter est celui de bruit d'environnement. Quant au deuxième facteur, il est lié au transcodage de la parole dans les applications mobiles.

Dans ce cadre, et dans la première partie de thèse, nous proposons des nouveaux paramètres acoustiques fondés sur le système auditif humain en utilisant les filtres Gammatones et exploitant le module et la phase du spectre de fréquence.

Dans la deuxième partie de la thèse, nous nous sommes focalisés sur les techniques de prétraitement du signal de parole à savoir la Détection d'Activité Vocale (Voice Activity Detection, VAD) utilisée pour détecter la présence de la parole dans un signal audio. Ainsi, dans le cadre de cette thèse, nous avons également implémenté une nouvelle technique de détection basée sur l'utilisation des Machines à Vecteurs du Support (SVM).

Méthodologies

Dans ce travail de thèse, nous considérons les systèmes de l'état de l'art, à savoir les systèmes à base des MFCC et le modèle de mélange de gaussiennes (GMM), comme des systèmes de référence dans toutes nos séries d'expériences de la première partie, en utilisant la base de données TIMIT.

Quant aux recherches de la deuxième partie, la nouvelle technique que nous proposons a été implémentée dans le contexte de l'espace des I-vecteurs en utilisant la base de données MOBIO.

Structure de la thèse

La thèse est organisée en cinq chapitres de la façon suivante :

- Le premier chapitre présente l'architecture du système RAL et celle de ses deux tâches principales à savoir la vérification et l'identification. Une description détaillée des différents blocs, et celle des techniques utilisées est également présentée dans ce chapitre.
- Dans le deuxième chapitre, nous nous focalisons davantage sur deux méthodes de la classification à savoir les modèles de Mélanges de Gaussiennes (GMM) et le I-vecteur. Ces méthodes sont adoptées dans l'implémentation des systèmes de l'état de l'art actuel dans le domaine de la vérification du locuteur.
- Dans le troisième chapitre, nous présentons quelques paramètres utilisés en reconnaissance automatique du locuteur. Nous introduisons par la suite des nouveaux paramètres acoustiques fondés sur le système auditif humain en utilisant les filtres Gammatones et exploitant le module et la phase du spectre de fréquence.
- Le quatrième chapitre donne les premiers résultats obtenus suite aux expérimentations réalisées sur la base de données TIMIT. Des expérimentations pour évaluer la pertinence des paramètres proposés aux dégradations causées par la présence de bruit d'environnement d'une part et le transcodage de la parole dans les applications mobiles d'autre part.
- La détection d'activité vocale (VAD) utilisée pour détecter la présence de la parole dans un signal audio représente notre centre d'intérêt dans le dernier chapitre de cette thèse. Dans ce chapitre, nous retrouvons le détail de la nouvelle technique de détection basée sur l'utilisation des Machines à Vecteurs du Support (SVM).
- En conclusion, nous récapitulons les résultats et les principales contributions de cette thèse avant d'exposer les perspectives envisagées.

Chapitre 1

La Reconnaissance Automatique du Locuteur

Sommaire

Résumé

La Reconnaissance Automatique du Locuteur RAL est le processus qui identifie automatiquement celui qui parle en se basant sur des informations individuelles incluses dans le signal de parole. Un système de reconnaissance de locuteur procède en quatre étapes : le prétraitement du signal de parole, la paramétrisation, la modélisation du locuteur et une dernière étape de décision. Dans ce chapitre, nous introduisons le principe de la Reconnaissance Automatique du Locuteur et nous présentons les différentes étapes du système.

1. Introduction

La parole est un signal aléatoire complexe qui contient plusieurs types d'informations. Le signal de parole véhicule un message linguistique qui sert à la communication entre les individus. Mais transporte également des informations sur l'identité de l'individu ayant prononcé le message. Les humains se servent de ces informations pour identifier les personnes qu'ils connaissent, en particulier lorsqu'ils ne peuvent pas voir leur interlocuteur, au téléphone par exemple. La reconnaissance automatique du locuteur RAL s'intéresse précisément à ces caractéristiques particulières et exploite les techniques du traitement automatique de la parole et de la reconnaissance des formes pour accomplir le but ultime de la biométrie, à savoir, l'authentification des personnes.

Outre que la reconnaissance du locuteur, la discipline du traitement automatique de la parole recouvre également une vaste panoplie d'activités (voir Figure 1.1), tels que : la synthèse et la reconnaissance de la parole, le codage, la compression, la reconnaissance de la langue et du dialecte, la reconnaissance des émotions, etc.

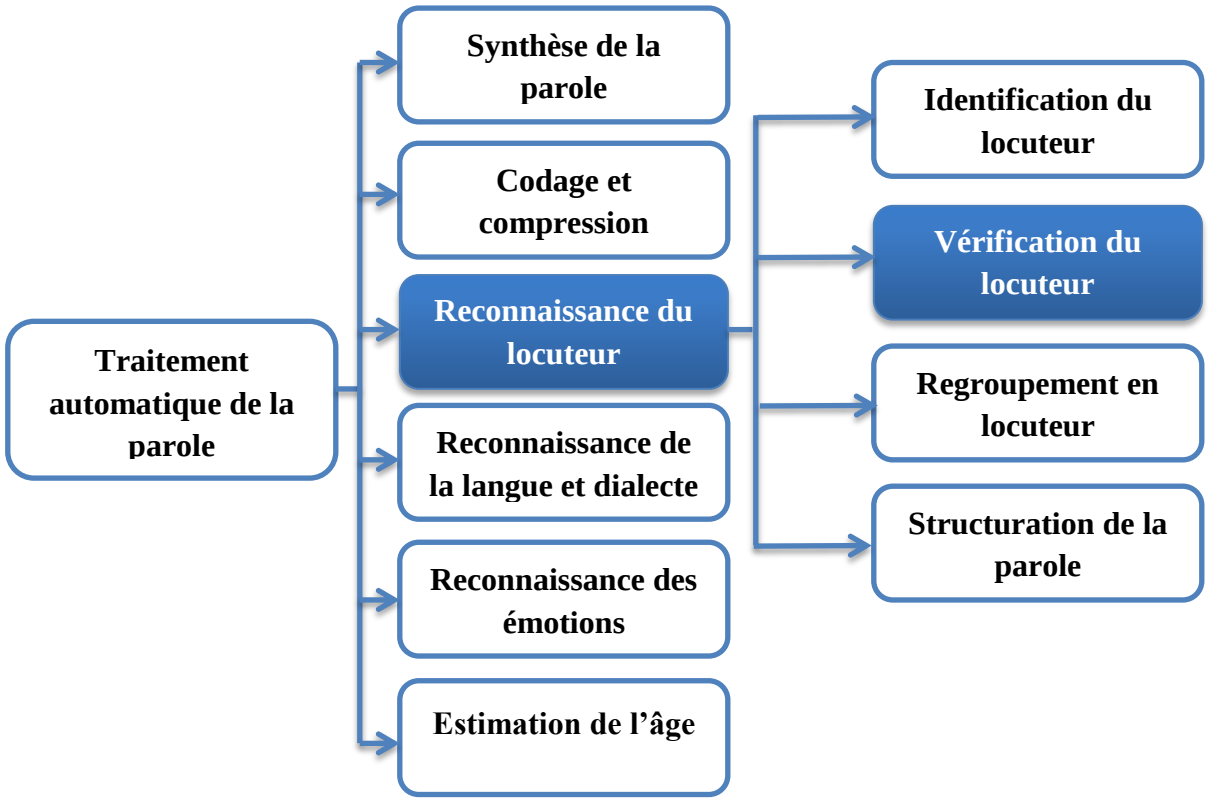

Figure 1.1 Arbre du traitement automatique de la parole avec une focalisation (en couleur) sur les sujets traités dans cette thèse.

2. Types de la variabilité de la parole

2.1. Variabilité intra-locuteur

La variabilité intra-locuteur identifie les différences dans le signal produit par une même personne. Cette variation résulte de l'état physique et moral du locuteur. Une maladie des voies respiratoires peut ainsi dégrader la qualité du signal de parole de manière à ce que celui-ci devienne totalement incompréhensible. L'humeur ou l'émotion du locuteur peut également influencer son rythme d'élocution ou son intonation.

2.2. Variabilité inter-locuteurs

La variabilité inter-locuteurs résulte en réalité des différences morphologiques et physiologiques des locuteurs. La parole est principalement produite grâce aux cordes vocales qui génèrent un son à une fréquence de base, le fondamental (Sönmez, Heck, Weintraub, & Shriberg, 1997). Cette fréquence sera différente d'un individu à l'autre et plus généralement d'un genre à l'autre. De plus, la transformation du son par l'intermédiaire du conduit vocal permet de générer des sons différents qui sont regroupés selon les classes. Or le conduit vocal est de forme et de longueur variables selon les individus et, plus généralement, selon le genre et l'âge. La variabilité inter-locuteurs trouve également son origine dans les différences de prononciation qui existent au sein d'une même langue et qui constituent les accents régionaux.

2.3. Variabilité due à l'environnement

Le bruit d'environnement peut également entraîner des déformations du signal de parole en obligeant le locuteur à accentuer son effort vocal.

2.4. Variabilité inter-sessions

La variabilité inter-session résulte des distorsions provoquées dans le signal de parole quand les canaux d'enregistrement et de transmission changent entre les sessions d'apprentissage et de test. De nombreux travaux expérimentaux ont montré que des variations de matériel entre les phases d'apprentissage et de test sont à l'origine de graves dégradations des performances (Van Vuuren, 1996). Par exemple, l'acquisition d'un signal de parole sur le réseau GSM introduit les dégradations suivantes sur le signal de parole :

- l'ajout du bruit de l'environnement,
- le sous-échantillonnage à 8 kHz du signal,

- le filtrage sur la bande de fréquence [300-3400] Hz,
- le codage-décodage (transcodage) à bas débit de la parole,
- l'ajout du bruit de quantification des paramètres émis,
- la transmission sur un lien sans-fil avec pertes.

3. Domaine d'application de la RAL

Les applications de la RAL concernent les problèmes de confidentialité et d'authentification. Nous distinguerons :

1. les applications "sur site" comme l'utilisation des serrures vocales pour contrôle d'accès.
2. Les applications liées aux télécommunications: ces applications concernent l'identification du locuteur à travers le réseau téléphonique pour accéder à un service de transactions bancaires par exemple afin de fournir une authentification vocale rapide ou encore les transactions par carte de crédit.
3. les applications judiciaires: recherche de suspects, orientations d'enquêtes, preuves lors d'un jugement.

La difficulté de la tâche de reconnaissance n'est pas la même d'une application à l'autre. Dans le cas des applications 'sur site', l'environnement de prononciation de la phrase ou du mot de passe est plus facilement contrôlé que dans le cas des applications via le réseau téléphonique (distorsions dues au canal, différences entre les combinés téléphoniques, bande passante limitée).

4. Tâches de la RAL

En reconnaissance du locuteur, on fait la différence entre l'identification et la vérification du locuteur, selon que le problème est de déterminer qui, parmi un nombre fini et préétabli de locuteurs, a produit le signal analysé, ou qu'il s'agit de vérifier que la voix analysée correspond bien à la personne qui est censée de la produire.

4.1. Identification du Locuteur

L'Identification Automatique du Locuteur IAL est le processus qui consiste à déterminer, parmi une population de locuteurs connus, celui susceptible de produire un enregistrement vocal. D'un point de vue schématique (Figure1.2), une séquence de parole, donnée en entrée du système d'IAL, est comparée à un ensemble de références caractéristiques des locuteurs. L'identité du

locuteur dont la référence est la plus proche de la séquence de parole est donnée en sortie du système d'IAL comme identité reconnue.

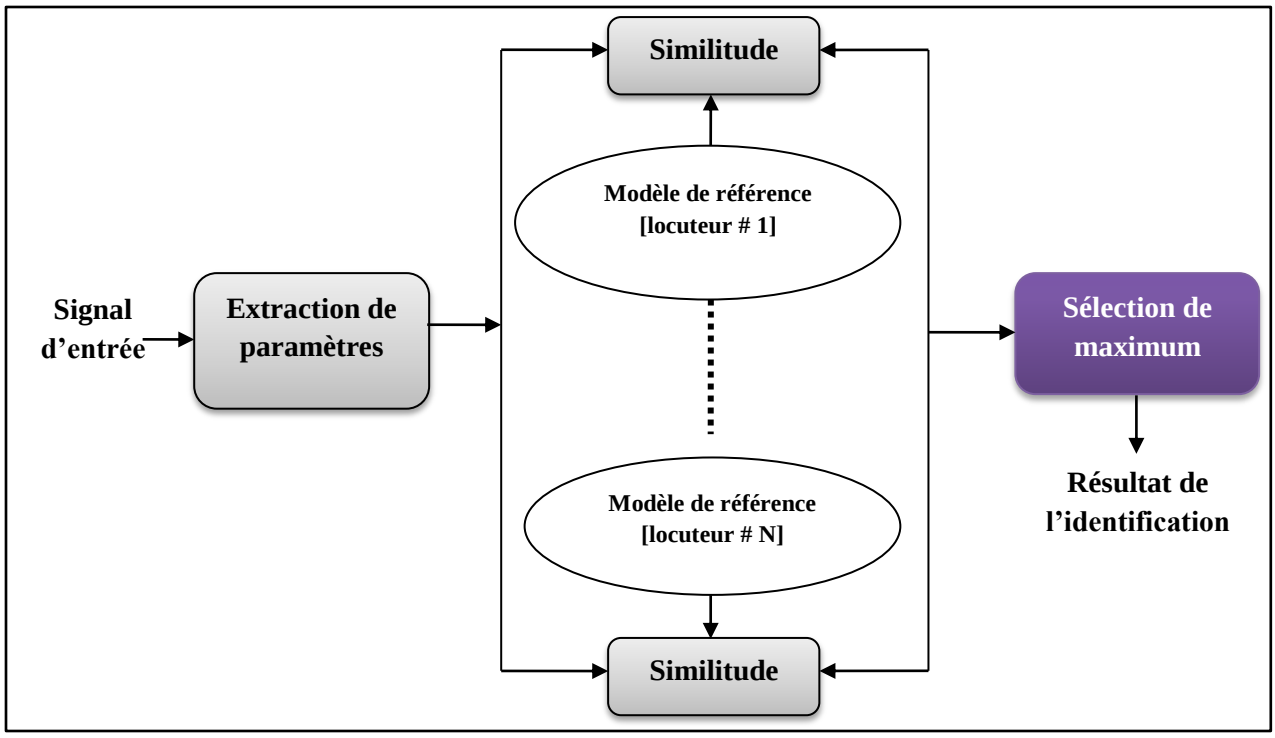

Figure 1.2 Schéma d'un système d'identification du locuteur.

Deux modes d'identification sont possibles:

4.1.1. Identification à ensemble fermé

Dans un ensemble fermé, on suppose que la voix traitée est produite par l'un des locuteurs de l'ensemble de référence. Un tel système ne doit alors être utilisé que dans un environnement au sein duquel tous les individus sont connus.

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

$$Y = argmax\, f(X|S) \qquad (1.1)$$

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

4.1.2. Identification à ensemble ouvert

Dans un ensemble ouvert, on ne sait pas si le propriétaire de la voix traitée appartient à l'ensemble de référence ou non. Dans l'identification à ensemble ouvert, on ne peut pas dire que le locuteur qui a la distance la plus petite est la source de la voix traitée. On doit avoir un seuil

prédéfini d'une façon que le locuteur qui a une distance inférieur au seuil sera considéré comme étant la source de la voix traitée. Pour ce faire, les données$S$sont comparées à chaque référence$X$ connue par le système. Chaque comparaison fournit un score $f(X|S)$. Le score le plus élevé est alors comparé à un seuil $\mathcal{O}$, fixe au préalable. Si le score est supérieur à ce seuil, le système décide qu'il s'agit de la personne correspondante à la référence sélectionnée. Si le score est inférieur à ce seuil, le système décide qu'il ne s'agit pas d'une personne connue.

L'identité la plus probable est obtenue, comme dans le cas de l'identification en milieu ferme, et la réponse du système est donnée par :

$$f(Y|S)\begin{cases} > \mathcal{O} & identité\ est\ celle\ du\ client\ X \\ < \mathcal{O} & identité\ non\ reconnue \end{cases} \qquad (1.2)$$

Où $\mathcal{O}$ est le seuil de décision fixe au préalable.

4.2. La Vérification du Locuteur

Il s'agit de vérifier que la voix analysée correspond bien à la personne qui est censée de la produire(Bimbot et al., 2004). La Vérification Automatique du Locuteur VAL est le processus qui prend la décision d'accepter ou de rejeter l'identité d'un locuteur susceptible d'être la source d'un enregistrement vocal. Le schéma de principe d'un système VAL est donné par la Figure 1.3.

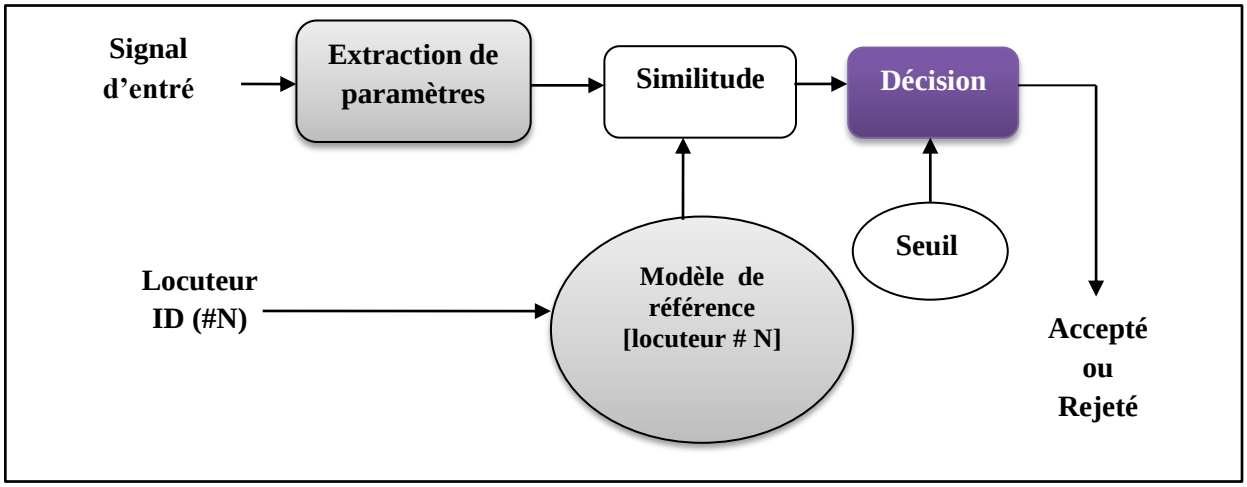

Figure 1.3 Schéma d'un système de vérification du locuteur.

4.3. Une comparaison entre l'identification et la vérification automatique du locuteur

Il est clair que les deux systèmes de reconnaissance automatique du locuteur, à savoir, l'identification et la vérification, ont une tâche commune qui est la comparaison d'une voix connue avec une autre non connue. Mais ces derniers présentent de grandes différences. La distinction la plus importante concerne les propriétés de l'ensemble de référence des locuteurs, elle est faite selon que, premièrement, l'ensemble est *fermé* ou *ouvert (voir section 4.1.1 et 4.1.2)*, deuxièmement, l'ensemble est *connu* ou *inconnu* (Rose & Clermont, 2001).

4.3.1. Ensemble connu et ensemble inconnu

La distinction entre l'ensemble fermé et l'ensemble ouvert n'a aucun sens dans un système de vérification du locuteur. Dans ce dernier, on suppose que les identités à proclamer appartiennent à l'ensemble de référence des locuteurs. Cependant, il y a une autre propriété de l'ensemble de référence qui est : *connu* ou *inconnu*. Dans la vérification traditionnelle l'ensemble de référence est connu, il peut être les employés d'une entreprise, les clients d'une banque ou en général l'identité de toutes personnes qui doivent être vérifiées de temps en temps.

4.3.2. Nombre de comparaisons

Le nombre de comparaisons faites lors d'une tâche d'identification ou de vérification diffère selon cette tâche. La vérification du locuteur nécessite une seule comparaison entre l'identité proclamée et la voix traitée. Dans l'identification du locuteur, la comparaison apparaît itérativement entre la voix traitée et chaque locuteur de l'ensemble de référence.

5. Architecture générale d'un Système de Reconnaissance Automatique du Locuteur

Le processus de reconnaissance de personnes se décompose en deux phases (voir Figure 1.4). La première étape consiste à obtenir une représentation de l'utilisateur. Elle est appelée *enrôlement ou apprentissage*. Lors de cette phase, le système construit sa représentation de l'individu, qui sera utilisée par la suite pour l'authentification. Cette représentation de l'utilisateur doit être unique et permanente. La deuxième étape est l'étape de test, elle consiste à mesurer la ressemblance entre les données fournies d'un locuteur souhaitant être authentifie et le modèle existant correspondant à l'identité annoncée.

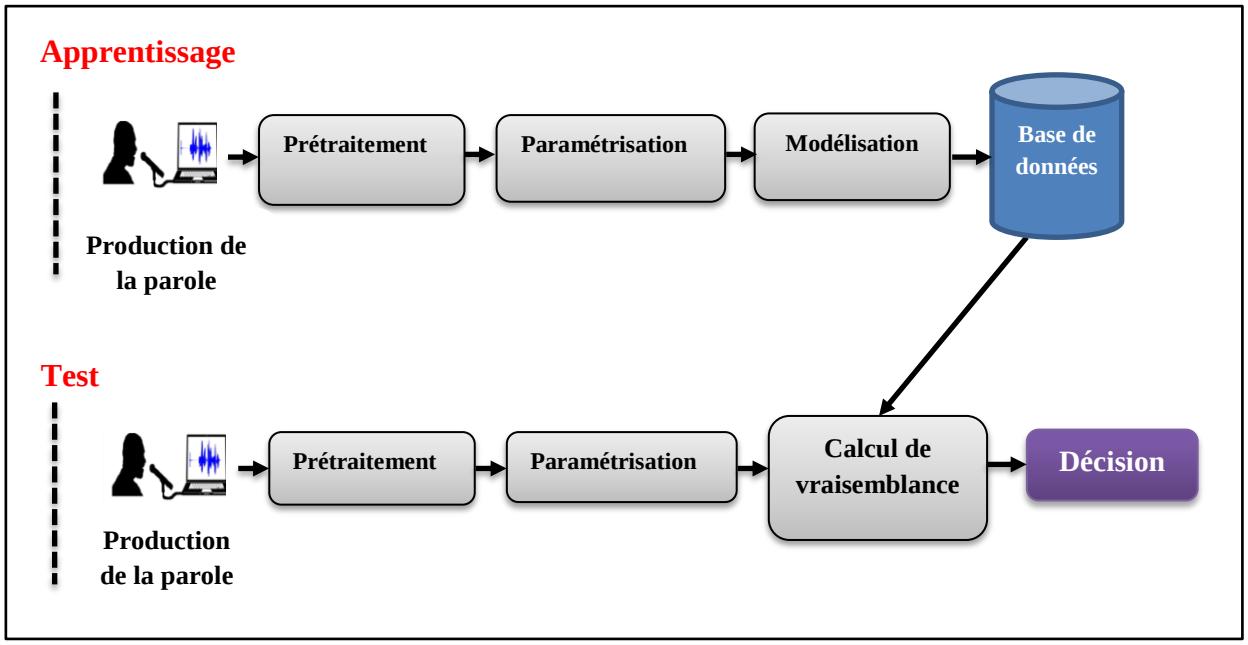

Figure 1.4 Structure générale d'un système de RAL.

Les étapes d'enrôlement et de test sont décrites dans la suite, sous la forme d'une succession de modules fonctionnels.

5.1. Production de la parole

La production de la parole est un mécanisme très complexe qui repose sur l'interaction entre le système neurologique et physiologique. Après que l'idée et la volonté de parler soient survenues, le cerveau envoie des signaux au système physiologique qui se compose de plusieurs organes et muscles (Figure 1.5). L'énergie fournit par un air expulsé des poumons met en mouvement les cordes vocales - contenues dans le larynx - qui par leur rapprochement et leur éloignement interrompent et libèrent cet air pour produire la voix qui est une succession de bouffées chargées d'énergie acoustique; grâce aux organes situés au-dessus du larynx, le pharynx, les fosses nasales et la cavité buccale qui comporte tous les éléments nécessaires à l'articulation: voile du palais, palais dur, mâchoire, dents, langue et lèvres, la parole est produite (Mami, 2003).

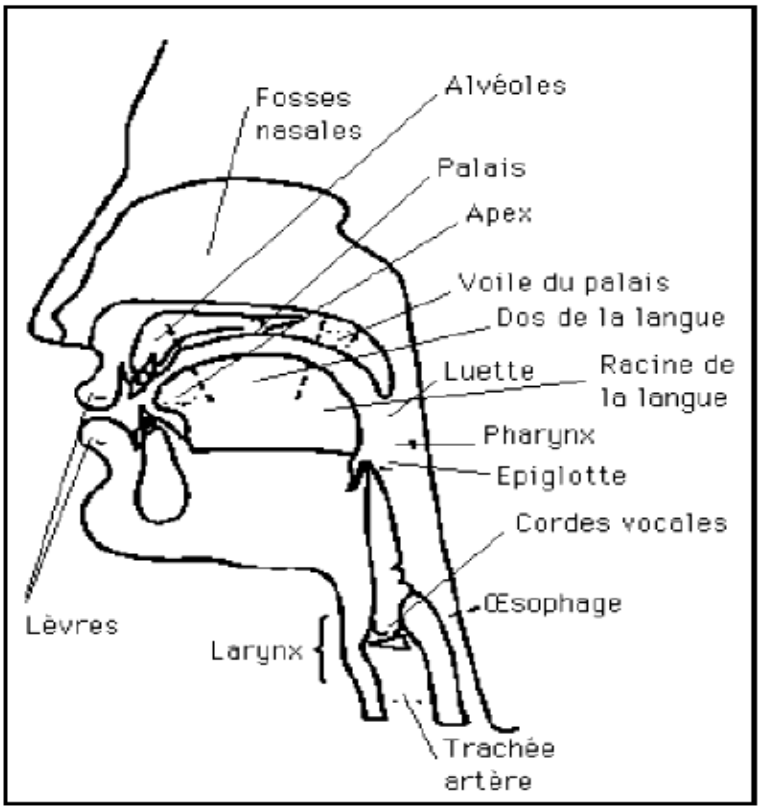

Figure 1.5 Organes de production de la parole (Pinto & Sato, 2016).

Le son résultant peut être classé comme voisé ou non voisé selon que l'air émis a fait vibré les cordes vocales ou non. Dans le cas des sons voisés, la fréquence de vibration des cordes vocales, dite fréquence fondamentale ou pitch s'étend généralement de 70 à 400 hertz. L'évolution de la fréquence fondamentale détermine la mélodie de la parole. Son étendue dépend des locuteurs, de leurs habitudes mais aussi de leurs âges (selon la taille du larynx qui va du large chez les adultes au moins large chez les enfants) et du type du locuteur (Pépiot, 2014) (environs 120 Hz pour les hommes et 200 Hz pour les femmes). Outre la fréquence fondamentale, la voix possède d'autres caractéristiques: l'intensité qui se mesure par l'énergie du signal (parler fort ou doucement), et le timbre qui définit la personnalité vocale et sa physiologie cachée (Wilhelm, 2012).

5.2. Prétraitement du signal de parole

5.2.1. La préaccentuation

Le spectre d'un signal de parole a une décroissance globale de l'énergie. Pour compenser cette décroissance, on effectue une préaccentuation en utilisant un filtre passe-haut. Le filtre le plus utilisé est le filtre à réponse impulsionnelle finie décrit ci-dessous:

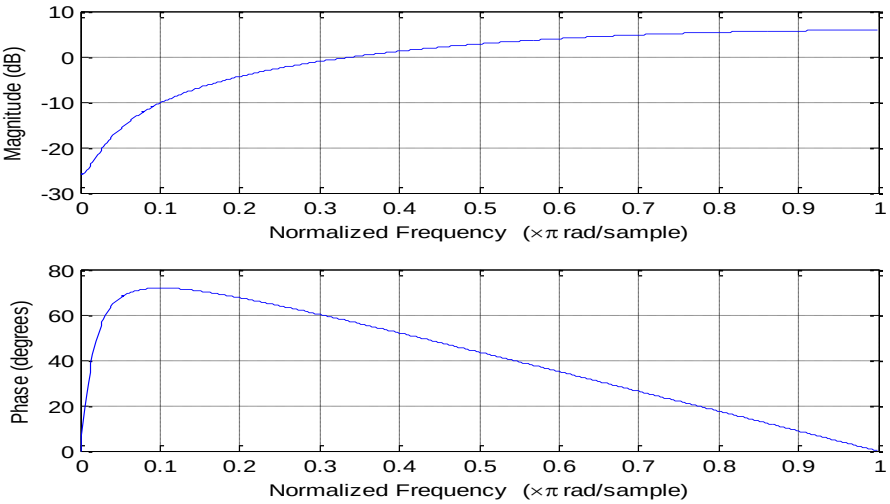

Figure 1.6 Le filtre de la préaccentuation.

$$H(z) = 1 - \alpha z^{-1}, \quad avec\ 0.9 \leq \alpha \leq 1 \qquad (1.3)$$

La réponse de ce filtre peut être vue dans la Figure 1.6.

5.2.2. Segmentation parole/silence ou la détection d'activité vocale

Les humains sont doués pour distinguer entre la parole et le silence, ce qui est aussi essentiel dans la reconnaissance du locuteur de retirer les trames de silence ou de bruits qui ne reflètent pas les caractéristiques du locuteur et qui diminuent les performances du système de RAL (Besacier & Bonastre, 1998). L'utilisation de la détection d'activité vocale (VAD), qui localise les segments de parole dans un énoncé, est une tâche très importante. Les mesures les plus utilisées pour trouver et éliminer le silence sont : l'énergie du signal (Sönmez et al., 1997), la puissance du signal et le taux de passage par zéro. La détection d'activité vocale semble relativement triviale, mais elle présente quelques difficultés dans la pratique. Récemment, plusieurs chercheurs se sont intéressés à la recherche de nouveaux algorithmes VAD. Nous fournirons plus de détails sur ce sujet dans le chapitre 5 de cette thèse.

5.3. Paramétrisation

Les caractéristiques vocales d'un individu peuvent ne pas être faciles à distinguer, mais elles sont uniques ; prenons comme exemple le cas de vrais jumeaux qui ont des différences dans leurs voix, mais les études montrent qu'ils ont une forme similaire des conduits vocaux et des

propriétés acoustiques, et c'est difficile de les distinguer d'un point de vue perceptuel. Ainsi, si la reconnaissance est effectuée par les humains (un expert par exemple) ou par des machines, certains aspects mesurables et prédéfinis de la parole doivent être pris en considération pour effectuer des comparaisons significatives entre les voix (Lippmann, 1997). En général, L'extraction des paramètres acoustiques est une étape très importante dans un système de reconnaissance automatique du locuteur. Le type de paramètres utilisés a un grand effet sur la performance du système. Dans le cas idéal, les paramètres acoustiques utilisés dans un système de reconnaissance automatique du locuteur doivent satisfaire les contraintes suivantes (Nolan, 1980):

- ✓ Difficiles à imiter par des imposteurs.
- ✓ Relativement faciles à extraire et à mesurer.
- ✓ Très fréquents dans le signal vocal.
- ✓ Robustes aux différents bruits et variations inter-sessions, ayant une grande variabilité inter-locuteurs et une faible variabilité intra-locuteur pour les considérer idéaux.

Ces paramètres peuvent être :

- Paramètres spectraux à court terme extraits sur des petites fenêtres et qui reflètent les caractéristiques du conduit vocal.
- Paramètres prosodiques extraits sur des grands segments (de dizaines à des milliers de millisecondes), ce sont des paramètres représentatifs de la source vocale.
- Paramètres de haut niveau (caractérisant la sémantique, l'accent, la prononciation, etc.) des locuteurs.

5.3.1. Paramètres à court terme

Le signal de parole présente des changements continus, par conséquent, le signal doit être décomposé en trames courtes d'une durée de 10-30ms (Jourani, 2012). Dans cet intervalle, le signal est considéré quasi-stationnaire et un vecteur de caractéristiques acoustiques est extrait de chaque trame d'où l'utilisation des paramètres à court terme dans la reconnaissance automatique du locuteur. Les principaux méthodes de l'analyse à court terme utilisés en RAL sont :

- Les analyses par bancs de filtres: il s'agit d'une analyse du système auditif qui consiste à calculer l'énergie du signal vocal dans différentes bandes de fréquences.
- Les analyses à base de transformée de Fourier: les coefficients issus de la transformée de Fourier peuvent être utilisés pour une analyse en bancs de filtres ainsi que pour les calculs

de coefficients cepstraux. Parmi les plus connus se trouvent les Coefficients Cepstraux à Fréquence Linéaire LFCC (Linear Frequency Cepstrum Coefficients) et les Coefficients Cepstraux à Fréquence Mel MFCC (Mel Frequency Cepstral Coefficients) (Davis & Mermelstein, 1980).

- Les analyses par prédiction linéaire(Bonastre et al., 2003): les coefficients issus de l'analyse LPC du modèle autorégressif peuvent être utilisés comme paramètres caractéristiques. Parmi les plus connus se trouvent les Coefficients Cepstraux à Prédiction Linéaire LPCC (Linear Predictive Cepstrum Coefficient) et les coefficients LSF (Line Spectral Frequency).

5.3.2. Paramètre prosodique

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

5.3.3. Paramètres de haut niveau

En plus des différences physiologiques, chaque personne a sa propre habitude linguistique et manière de parler, qui diffèrent en fonction de son milieu socioculturel. L'idiolecte qui est le système linguistique propre à une personne se manifeste par des choix particuliers dans la grammaire, le vocabulaire et les variations dans la prononciation et l'intonation.

5.3.4. Normalisation des caractéristiques

Différentes techniques de normalisation des paramètres acoustiques ont été proposées en RAL dans le but d'augmenter la robustesse des systèmes face aux variations des conditions d'enregistrement et de transmission. Cette partie exposera les principales d'entre elles.

5.3.4.1. Normalisation de moyenne et de variance cepstrales

La normalisation de moyenne et de variance cepstrales CMVN (Cepstral Mean and Variance Normalization) consiste à retirer la moyenne (la composante continue) de la distribution de chacun des paramètres cepstraux, et à ramener la variance à une variance unitaire en divisant par l'écart type (Viikki, Bye, & Laurila, 1998). On parle de CMN (Cepstral Mean Normalization) (Furui, 1981), quand seule la moyenne est normalisée. Le passage du domaine temporel au

domaine cepstral, transforme les bruits convolutifs en bruits additifs. Les bruits convolutifs variant lentement dans le temps seront donc représentés par une composante additive presque constante. Par conséquent, la suppression de la composante continue permet de réduire l'effet de ces bruits.

5.3.4.2. Le filtrage RASTA

Le filtrage RASTA correspond à un filtrage passe-bande dans le domaine cepstral. Ce filtrage est utilisé pour la réduction des bruits convolutifs engendrés par le canal de communication, et qui sont généralement à variation beaucoup plus lente que la parole (Hermansky & Morgan, 1994).

5.3.4.3. La Gaussianisation des paramètres (Feature warping)

C'est une technique qui modifie les paramètres acoustiques afin que leurs distributions suivent une autre distribution donnée, typiquement une distribution gaussienne de moyenne nulle et de variance unité (Pelecanos & Sridharan, 2001).

Ces trois premières techniques sont des techniques non supervisées, qui n'utilisent aucune connaissance sur le canal.

5.3.4.4. Le mappage des paramètres (Feature mapping)

C'est une technique de normalisation supervisée qui permet la réduction des effets de la variabilité du canal entre les conditions d'apprentissage et de test. Cette technique repose sur la projection des paramètres acoustiques liés à une condition d'enregistrement donnée dans un nouvel espace de caractéristiques indépendant du canal (Reynolds, 2003).

5.4. Modélisation des vecteurs acoustiques

Dans un système de reconnaissance automatique du locuteur, les paramètres acoustiques sont utilisés pour estimer un modèle, qui doit satisfaire les contraintes suivantes (Blouet, 2002):

- Il doit nécessiter le plus faible espace de stockage possible.
- Il doit avoir une méthode d'estimation la moins complexe possible.
- Il doit permettre une décision rapide lors de la phase de test.
- Il doit être le plus robuste possible aux variations intra-locuteur.
- Il doit permettre la meilleure séparation des locuteurs entres eux.

- Il doit avoir la représentation la plus complète possible des paramètres acoustiques des locuteurs.

La plupart des techniques de modélisation du locuteur reposent sur des hypothèses mathématiques de la distribution des caractéristiques (gaussienne, par exemple). Si ces hypothèses ne sont pas vérifiées par les paramètres, des imperfections lors de la phase de modélisation sont introduites. Différentes techniques sont utilisées en RAL pour réaliser les références de locuteurs. Ci-dessous, seront exposées quelques méthodes de modélisation utilisées dans les systèmes de reconnaissance automatique du locuteur.

5.4.1. Approche vectorielle (QV)

Cette méthode consiste à partager l'espace des vecteurs acoustiques en des partitions non chevauchées, telles que chaque vecteur d'une partition est représenté par le vecteur moyen de cette dernière (Soong, Rosenberg, Juang, & Rabiner, 1987). L'ensemble des vecteurs moyens de ces partitions qui représentant le modèle du locuteur, est appelé dictionnaire (codebook) (Figure 1.7). La création de dictionnaire est faite de manière à ce que la distorsion moyenne des données d'apprentissage soit minimale. Lors de la phase de test, une simple métrique de distance (euclidienne, Mahalanobis, cosine, etc.) est utilisée pour mesurer la distance entre la donnée test (vecteurs acoustiques) et le modèle du locuteur (dictionnaire ou vecteurs moyens des partitions). Cependant, cette méthode élimine beaucoup d'information sur les locuteurs et elle nécessite des paroles très longues pour avoir des informations statistiques stables.

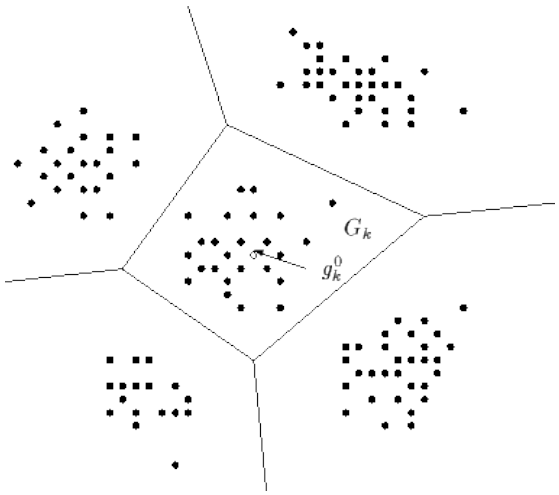

Figure 1.7 Définition des classes de données en quantification vectorielle[1].

[1] http://users.polytech.unice.fr/~leroux/parolehmm/node10.html

5.4.2. Le Modèle de Markov Caché (HMM)

Les Modèles de Markov Caché (Hidden Marcov Model, HMM) ont été initialement introduits en reconnaissance de la parole (Rabiner, 1989), puis leur utilisation s'est étendue peu à peu au domaine de la reconnaissance du locuteur. Cette approche suppose que les caractéristiques forment une succession d'états distincts et considère le locuteur comme source probabiliste. Les probabilités initiales se trouver dans chaque état, les probabilités des transitions décrivant les passages possibles d'un état à un autre, ainsi que la probabilité de sortie représentant les distributions conditionnelles des paramètres observés en fonction des états du modèle; représentent les caractéristiques de ce dernier (Figure 1.8).

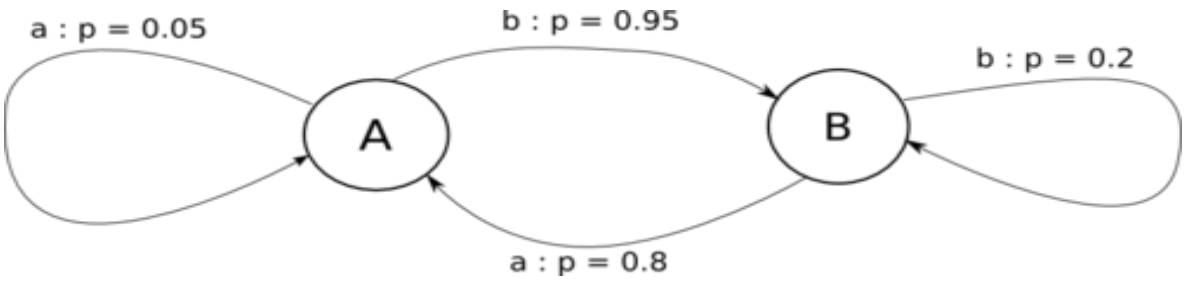

Figure 1.8 Un modèle de Markov caché à deux états.

5.4.3. Les Réseaux de Neurones Artificiels (ANN)

Un Réseau de Neurones Artificiels *(*Artificial Neural Network, ANN), est un modèle discriminant, formé d'un grand nombre de neurones fortement interconnectées. La sortie de chaque neurone est fonction de ses entrées (McCulloch & Pitts, 1943). En RAL, les locuteurs sont modélisés par un réseau de neurones appris sur l'ensemble des données d'apprentissage de tous les clients.

5.4.4. Les Machines à Vecteurs de Support (SVM)

Les Machines à Vecteurs de Support (Support Vector Machine, SVM) ont été développés dans les années 1990 à partir des considérations théoriques de Vladimir Vapnik sur le développement d'une théorie statistique de l'apprentissage : la Théorie de Vapnik-Chervonenkis (Burges, 1998). L'idée d'une classification par SVM est de calculer l'hyperplan permettant de séparer au mieux deux nuages de points. Il s'agit de reconsidérer le problème dans un espace de dimension supérieure. Dans ce nouvel espace, il existe un séparateur linéaire permettant de classer au mieux les deux nuages de points.

5.4.5. Le Modèle de Mélange de Gaussiennes (GMM)

Un modèle de Mélange de Gaussiennes (Gaussian Mixture Model, GMM) est utilisé pour la modélisation des paramètres du locuteur en fonctions de densité de probabilité (Reynolds, Quatieri, & Dunn, 2000; Reynolds & Rose, 1995) (Figure 1.9). L'évaluation de ces fonctions à différents points de données obtenus à partir d'une séquence de test fournit un score de probabilité pouvant être utilisé pour le calcul de la similarité entre le modèle du locuteur et les données d'un inconnu. Dans les tâches de reconnaissance du locuteur indépendantes du texte, lorsqu'il n'y a pas une connaissance à priori sur le contenu de la parole, l'utilisation de GMM pour modéliser les caractéristiques à court terme a été jugée la plus efficace pour la modélisation acoustique (Reynolds et al., 2000).

5.4.5.1. Le modèle du non locuteur ou Modèle du Monde (UBM)

Le Modèle du Monde (Universal Background Model, UBM) est utilisé comme a priori pour l'estimation des modèles de locuteur. Il représente la couverture exhaustive des paramètres acoustiques d'un grand nombre de locuteurs enregistrés dans diverses conditions. Ce modèle est appris par maximum de vraisemblance avec une très grande quantité de données. Ainsi, les enregistrements collectés pour la création du modèle du monde doivent représenter :

1. les différentes langues de la population de locuteurs à reconnaître,
2. les genres considérés (homme, femme, les deux),
3. les conditions d'enregistrements (type de microphone, type de canal de transmission).
4. le type de parole : lue, spontanée, conversationnelle.

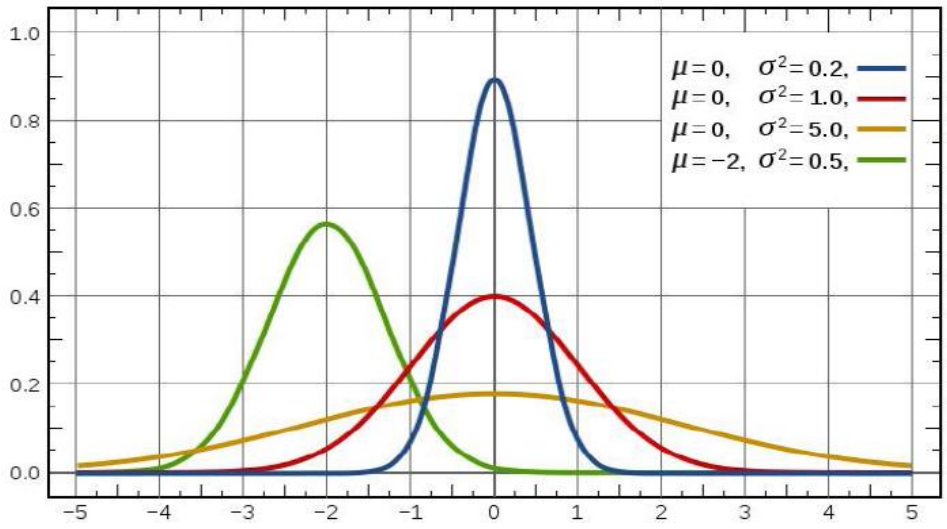

Figure 1.9 Mélanges des gaussiennes.

Nous pouvons observer que le modèle du monde intervient à différents niveaux dans un système de reconnaissance et surtout pour la vérification automatique de locuteur VAL, il représente l'hypothèse inverse dans le test de vérification, et sert d'initialisation pour l'estimation des modèles de locuteur.

5.4.6. Méthodes de compensation de la variabilité inter-session

Le changement des canaux, d'enregistrement et de transmission, entre les données d'apprentissage et celles des tests engendre une dégradation considérable des performances des Systèmes RAL. Le passage à l'espace des super-vecteurs rend possible la compensation de toute variabilité indésirable des supervecteurs GMM. Les formalismes de l'Analyse de Facteur (FA) et de la variabilité totale (I-vecteur) ont été introduits pour compenser ces variabilités dans les systèmes basés sur les GMM.

5.4.6.1. Analyse de Facteur (FA)

FA décrit la variabilité des vecteurs des données de grande dimension observables en utilisant un nombre réduit de variables inobservables/cachées. Pour la reconnaissance du locuteur, de nombreuses variantes de FA ont été employées, commençant par leurs utilisations dans l'espace des super-vecteurs GMM, aboutissant à l'approche actuelle I-vecteur (Dehak et al., 2009; Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2011). Le formalisme de l'analyse jointe des facteurs (Joint Factor Analysis) (JFA) (Kenny, Boulianne, Ouellet, & Dumouchel, 2006, 2007b) tend à modéliser la variabilité inter-locuteurs et à compenser la variabilité inter-sessions en décomposant le modèle du locuteur en deux éléments, l'un dépendant du locuteur et l'autre du canal.

5.4.6.2. Variabilité totale (Dehak et al., 2011)

La JFA définit deux espaces séparés, à savoir l'espace du locuteur (défini par les voix propres) et l'espace du canal (défini par les canaux propres). Ce puissant formalisme a conduit à la définition d'un nouvel espace unique appelé l'espace de la variabilité totale, englobant simultanément les variabilités inter-locuteurs et inter-sessions. Cet espace est défini par la matrice de la variabilité totale qui contient les vecteurs propres (I-vecteurs) associés aux plus grandes valeurs propres de la matrice de covariance de la variabilité totale.

Le lecteur pourra se référer au chapitre 2 pour une description plus complète de GMM-UBM et I-vecteur et leurs importances dans un système de Vérification Automatique du Locuteur.

5.5. Décision et évaluation des performances

Le dernier module dans le système RAL consiste à rechercher l'identité de la personne produisant la séquence définie, en d'autres termes, c'est un processus de comparaison entre une voix inconnue et un ensemble de voix d'une population de référence, et de détermination de l'appartenance ou non de cette personne à la base de données.

5.5.1. Décision et performances du IAL

En identification automatique du locuteur en ensemble fermé, la voix traitée est supposée produite par l'un des locuteurs de la base de données; l'identité recherchée est donc celle qui correspond au plus grand score s'il s'agit d'une mesure de similarité ou au plus petit score s'il s'agit d'une mesure de distance, ces scores sont calculés entre la voix traitée et chaque locuteur de cette base. Cependant en identification en ensemble ouvert, le système ne possède aucune connaissance sur l'appartenance du propriétaire de la voix traitée à la base de données. Dans ce cas, le plus grand/plus petit score doit être comparé avec un seuil prédéfini pour décider si le locuteur existe ou pas dans la base de référence.

En identification en ensemble fermé, les performances se mesurent en termes de taux d'identification correcte I_C ou incorrecte I_i que nous obtenons par les formulations suivantes :

$$I_C = \frac{nombre\ de\ test\ correctement\ identifié}{nombre\ totale\ de\ test} \times 100 \qquad (1.4)$$

$$I_i = \frac{nombre\ de\ test\ mal\ identifié}{nombre\ totale\ de\ test} \times 100 \qquad (1.5)$$

$$avec: I_i + I_i = 100\% \qquad (1.6)$$

Alors qu'en ensemble ouvert, ce sont les faux rejets FR et les fausses acceptations FA qui sont mesurés (Mami, 2003).

5.5.2. Décision et performances du VAL

Dans un système de vérification automatique du locuteur, une décision d'acceptation ou de rejet de l'identité d'un locuteur susceptible d'être la source de l'enregistrement vocal est prise. En d'autres termes, c'est un processus de comparaison entre la voix d'une personne déterminée et l'échantillon de sa propre voix. Cette méthode vérifie si la voix de la personne proclamant une

identité connue correspond bien à la voix de référence de cette personne. Les performances d'un système de VAL s'évaluent par les critères suivants:

5.5.2.1. Les fausses acceptations FAR et les faux rejets FRR

Pour évaluer les performances d'un système de vérification, deux types d'erreurs peuvent être commis. Premièrement, le client est rejeté alors que l'identité proclamée est la sienne. Dans ce cas, on parle de *« Faux Rejet»* (FR). La deuxième erreur survient lorsqu'un imposteur est accepté alors que l'identité annoncée n'est pas la sienne. Ce cas est appelé *« Fausse Acceptation »* (FA) (Figure 1.10). Un système biométrique est évalué par les taux des FR et FA (en anglais, False Rejection Rate, (*FRR*) et False Acceptation Rate (*FAR*)). Un bon système de vérification minimise les deux taux d'erreurs FAR et FRR.

$$FAR = \frac{Nombre\ \ d'imposteur\ acceptées}{nombre\ totale\ d'imposteurs} \times 100 \qquad (1.7)$$

$$FRR = \frac{Nombre\ \ de\ client\ rejetées}{nombre\ totale\ du\ clients} \times 100 \qquad (1.8)$$

5.5.2.2. Taux d'Egal Erreur EER

Le plus souvent, les performances du système RAL sont exprimées en taux d'égale erreur (Equal Error Rate, *EER*), qui correspond à la valeur où le FRR est égal au FAR. Plus la valeur de l'EER est faible, plus le système est meilleur. L'EER est utilisé pour comparer les systèmes RAL entre eux.

5.5.2.3. Courbe DET ou Courbe ROC

À chaque valeur de seuil est associé un couple (FAR, FRR). L'ensemble des couples obtenus peut être représenté sous la forme d'une courbe ROC (Oglesby, 1995) ou d'une courbe DET (A. Martin, Doddington, Kamm, Ordowski, & Przybocki, 1997) qui est la représentation la plus communément utilisée pour évaluer la pertinence du seuil de décision en fonction de ces deux taux d'erreurs (voir Figure 1.11).

5.5.2.4. Fonction de coût de décision DCF

Comme l'EER ne différencie pas entre les deux erreurs (FAR et FRR), ce qui parfois, n'est pas une mesure de performance réaliste. Une pondération est alors introduite pour chacun de ces taux. Une fonction de coût de décision (Decision Cost Function, DCF), qui est une mesure de

rendement présentée par l'Institut National des standards et technologie (NIST) (A. F. Martin & Greenberg, 2010), peut être appliquée. Cette DCF s'exprime sous la forme :

$$DCF(\theta)=C_{miss}P_{miss}(\theta)P_{tar} + C_{fa}P_{fa}(\theta)(1-P_{tar}) \quad (1.9)$$

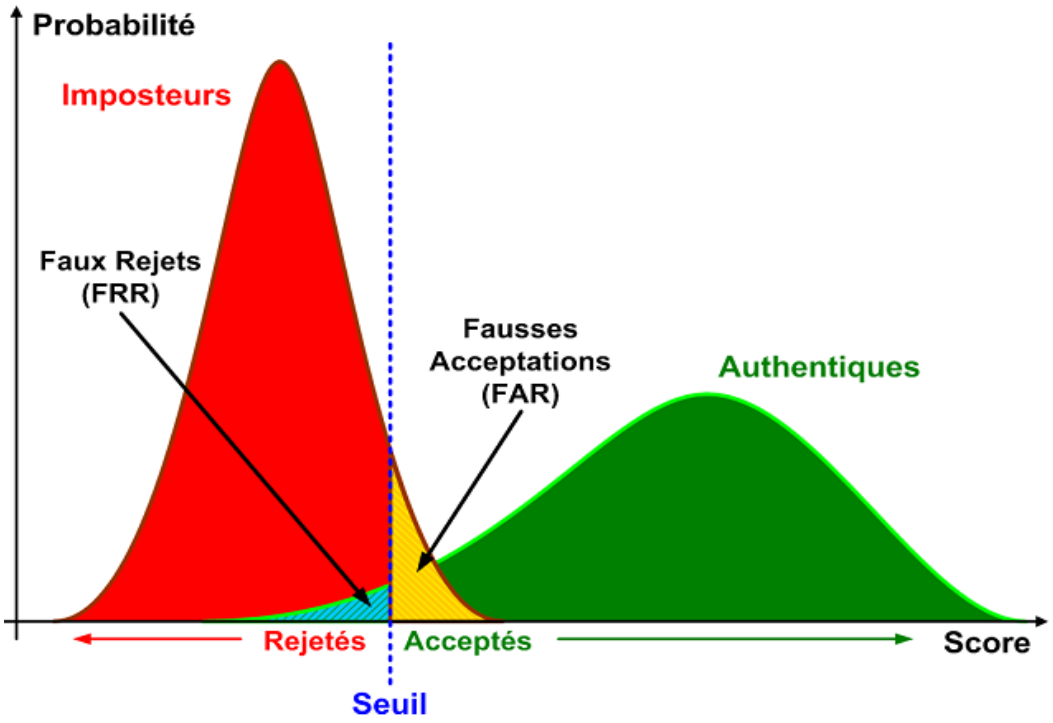

Figure 1.10 : Illustration des distributions des scores et le seuil de décision. Les surfaces sous les courbes avec des couleurs bleu et jaune représentent, respectivement les FRR et FAR (Aloui, Nait-Ali, & Naceur, 2018).

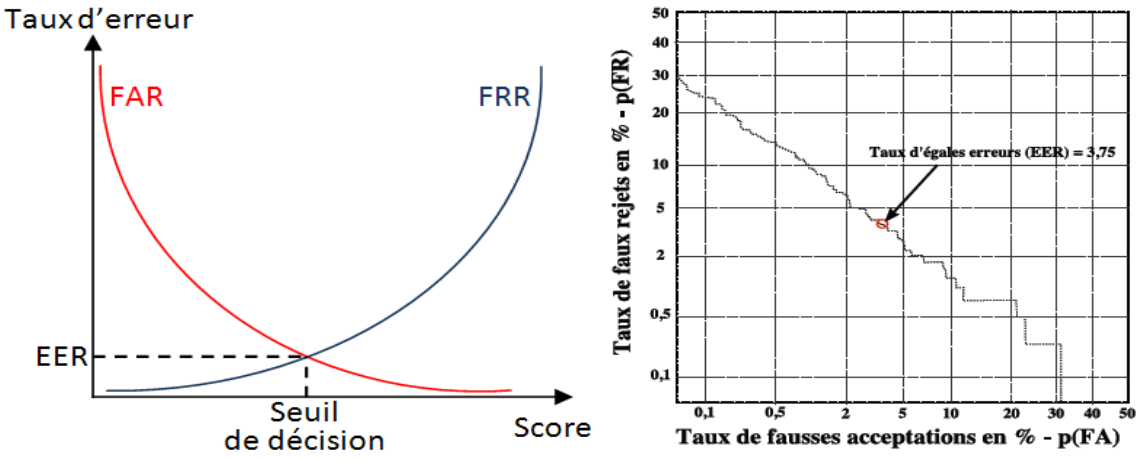

Figure 1.11 Exemple de courbe ROC (à gauche) et de courbe DET (à droite).

Avec $P_{miss}(\theta)$ et $P_{fa}(\theta)$ sont le taux de faux rejet et le taux de fausse acceptation respectivement en fonction du seuil de décision θ, P_{tar} est la probabilité a priori d'un vrai locuteur, C_{miss} est le coût de faux rejet et C_{fa} est le coût de fausse acceptation. L'équation (1.9) est utilisée pour calculer les valeurs minimales (*Min*DCF) proposées par le NIST pour 2008 SRE (DCF_{08}) avec C_{miss} = 10, $C_{fa} = 1$ et $P_{tar} = 0.01$ (A. F. Martin & Greenberg, 2010).

5.5.2.5. HTER

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

$$HTER = \frac{1}{2(FAR+FRR)} \qquad (1.10)$$

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

5.6. Normalisation des scores

Il existe différentes méthodes de normalisation des scores (Jourani, 2012) ayant pour but de rendre le seuil de décision, fixé dans la phase de développement, dépendant du locuteur et adapté aux conditions des séquences de test. La forme générale de ces méthodes est la suivante:

$$\hat{s} = \frac{s-\mu_i}{\sigma_i} \qquad (1.11)$$

Où μ_i et σ_i sont respectivement, la moyenne et l'écart-type des scores imposteurs. Ci-après une brève description de quelques-unes d'entre elles:

5.6.1. Z-Norm

Une technique de normalisation dépendante du locuteur qui ne nécessite pas la connaissance de la séquence test (K.-P. Li & Porter, 1988). μ_i et σ_i sont calculés dans la phase de développement à partir des scores imposteurs du client cible (Equation 1.11).

5.6.2. T-Norm

Une technique qui permet le calcul des scores des imposteurs dans la phase test en comparant l'énoncé de test du client cible avec un ensemble des modèles des locuteurs imposteurs. La T-norm permet de compenser la variation des conditions d'enregistrement entre la données test et la donnée d'apprentissage.

Les méthodes de normalisation les plus utilisées sont Z-Norm et T-Norm et même leur combinaison de ZT-Norm (Z-Norm suivie par T-Norm) et inversement (TZ-Norm) sont appliquées (Jourani, 2012).

$$\hat{s} = \frac{s-\mu_i^t}{\sigma_i^t} \qquad (1.12)$$

Où μ_i^t et σ_i^t sont respectivement, la moyenne et l'écart-type des scores imposteurs calculés dans la phase test.

5.6.3. S-Norm

Une normalisation symétrique des scores proposée par Kenny (Kenny, 2010) qui utilise à la fois les statistiques des scores imposteurs calculées dans la phase de développement et de test.

$$\hat{s} = \frac{s-\mu_i^a}{\sigma_i^a} + \frac{s-\mu_i^t}{\sigma_i^t} \qquad (1.13)$$

Où μ_i^a et σ_i^a sont les statistiques calculées dans la phase d'apprentissage, tandis que μ_i^t et σ_i^t sont celles calculées lors de la phase test.

6. Conclusion

Dans ce chapitre, nous avons présenté un état de l'art et les principaux aspects et concepts utilisés en RAL. Nous avons également présenté la structure générale d'un système de reconnaissance automatique du locuteur à savoir, le prétraitement, l'extraction, la modélisation et la décision et ses composants modulaires. Pour chaque module, nous avons donné les différentes techniques utilisées, souvent en citant leurs avantages et leurs faiblesses qui poussent les chercheurs à d'éventuels travaux d'amélioration dans le but d'augmenter la robustesse de ces systèmes.

Chapitre 2

La modélisation des vecteurs acoustiques

Sommaire

Résumé

Dans ce chapitre, nous allons présenter un module important de la Reconnaissance Automatique du Locuteur. Comme nous l'avons déjà mentionné dans le chapitre précédent, les systèmes RAL nécessitent une étape de modélisation des paramètres acoustiques. En effet, plusieurs méthodes de modélisation ont été utilisées et chacune d'elle présente des avantages et des inconvénients. Dans ce chapitre, nous nous focalisons davantage sur deux méthodes de modélisation à savoir les modèles de Mélanges de Gaussiennes GMM et le I-vecteur. Ces méthodes sont adoptées dans l'implémentation des systèmes de l'état de l'art actuel dans le domaine de la vérification du locuteur.

1. Modèle à Mélanges du Gaussiennes

Les Modèles de Mélanges du Gaussiennes (GMM) sont utilisés pour modéliser un locuteur donné par une somme pondérée de gaussiennes. Cette méthode est la plus utilisée en ce qui concerne la reconnaissance de locuteur en mode indépendant du texte. Les travaux de Reynolds (Reynolds & Rose, 1995) constituent l'état de l'art en la matière. Le modèle GMM est une segmentation en plusieurs classes des paramètres acoustiques représentants l'identité d'un locuteur (Figure. 2.1).

Le modèle de mélange de gaussiennes est le plus adapté à la reconnaissance automatique du locuteur indépendante de texte. Cela s'explique par certaines de ses propriétés, telles que :

✓ La capacité à modéliser des formes complexes de distributions sans être théoriquement limité par sa complexité, en augmentant le nombre de composantes du mélange.

✓ La segmentation implicite des vecteurs acoustiques en *K* classes dans l'espace des paramètres. Chacune d'elles possède sa probabilité d'occurrence a priori, mais la modélisation ne donne aucune information sur leur dynamique.

✓ L'existence d'un outil très puissant pour l'estimation des paramètres: l'algorithme *Expectation-Maximisation (EM).*

Un GMM est un mélange de densités gaussiennes paramétré par un certain nombre de vecteurs moyens, des matrices de covariance et des poids des composants individuels du mélange (Kenny et al., 2007b). Le modèle est représenté par une somme pondérée des M densités gaussiennes (figure 2.2.) donnée dans l'équation (2.1).

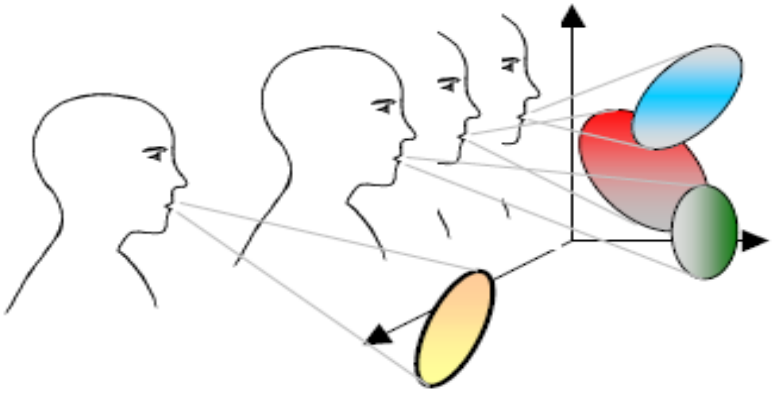

Figure 2.1 Illustration de nuages acoustiques représentants l'identité d'un locuteur.

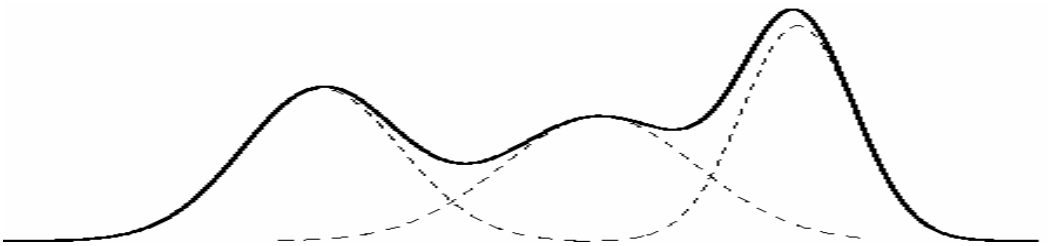

Figure 2.2 Approximation de la distribution d'un paramètre acoustique par une combinaison de gaussiennes.

$$p(x|\theta) = \sum_{m=1}^{M} \pi_m b_m(x) \quad (2.1)$$

Où x est un vecteur aléatoire de dimension d qui représente un vecteur acoustique. $b_m(x)$, $m = 1,2,\dots,M$ sont les densités de probabilités gaussiennes qui composent le modèle. π_m, $m = 1,2,\dots,M$, sont les poids des densités $b_m(x)$.

Chaque densité de probabilité $b_m(x)$ est définie par :

$$b_m(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_m|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu_m)'\Sigma_m^{-1}(x-\mu_m)^T\right\} \quad (2.2)$$

Tel que μ_m est le vecteur moyen et Σ_m est la matrice de covariance. Les poids $\pi_m$vérifient la contrainte donnée par la formule suivante :

$$\sum_{m=1}^{M} \pi_m = 1 \quad (2.3)$$

Ce modèle statistique est composé par des vecteurs moyens μ_m, des matrices de covariances Σ_m et des coefficients π_m de toutes les densités de probabilités $b_m(x)$.
Ces paramètres sont représentés par la notation suivante :

$$\theta = \{\pi_m, \mu_m, \Sigma_m\}, m = 1, \dots, M \quad (2.4)$$

La modélisation par GMM comprend deux phases principales (Morris, Wu, & Koreman, 2004; Reynolds & Rose, 1995) :

- Une phase d'apprentissage sur un ensemble de fichiers audio qui représentent les différents locuteurs du système.

- Une phase d'identification qu'un son quelconque est produit par un locuteur bien déterminé.

1.1. Phase d'apprentissage

Dans cette phase, on estime les paramètres des gaussiennes qui composent un modèle GMM en se basant sur les vecteurs acoustiques déterminés dans l'étape d'extraction de paramètres. L'apprentissage se fait en deux étapes: La première étape est l'initialisation des paramètres du modèle en utilisant l'algorithme K-moyennes (K-means) ou l'algorithme LBG, la deuxième étape est l'optimisation des paramètres obtenus dans la première étape en utilisant l'algorithme EM *(Expectation Maximisation)*(Bilmes, 1998; Dempster, Laird, & Rubin, 1977).

1.1.1. Initialisation des paramètres

1.1.1.1. Algorithme des K-moyennes

L'algorithme des k-moyennes est un algorithme de classification non supervisée ayant pour objectif de trouver k vecteurs $x_1, \dots, x_k$ permettant de représenter au mieux un ensemble de n vecteurs $y_1, \dots, y_n$. Un vecteur représentant une classe est choisi comme étant le représentant de cette classe au sens d'une distance. L'algorithme des K-moyenne n'est que localement optimal, par conséquence, il est influencé par ses conditions initiales(Linde, Buzo, & Gray, 1980).

Étape 1 : La première étape est l'initialisation du dictionnaire.

Étape 2 : La deuxième étape consiste à appliquer deux règles, définies ci-dessous

- **La règle de centroide :** Cette règle exige que tous les centroides soient les moyennes des vecteurs acoustiques des régions représentées par ces centroides. Cela peut être formulé comme suit:

$$c_i = \frac{1}{N_i}\sum_{k=1}^{N_i} \bar{x}_k \qquad (2.5)$$

- **La règle de plus proche voisin :** Le vecteur $\bar{x}_k$ est affecté à la région *i*si la distance euclidienne entre ce vecteur et le centroide de cette région est minimale. La formule suivante décrit explicitement cette règle.

$$region(\bar{x}_k) = i, \quad c_i = \min(d_{j=1}^{K}(\bar{x}_k, c_j)) \qquad (2.6)$$

K est le nombre de régions.

c_i est le centroide de la *i* région.

$d(x_k, c_i)$ est la distance euclidienne entre les vecteurs x_k et c_i.

Étape 3 : S'il y'a une amélioration importante de la distorsion moyenne donnée par la formule ci-dessous, alors aller à l'étape 2 et si la différence de la distorsion est inférieur à un seuil prédéfini alors Fin.

$$D_m = \frac{1}{N}\sum_{k=1}^{N} d(\overline{x}_k, cent(\overline{x}_k)) \qquad (2.7)$$

1.1.1.2. Algorithme LBG

L'algorithme LBG (Linde et al., 1980) est un l'algorithme itératif utilisé pour la création d'un dictionnaire optimal basé sur des vecteurs d'apprentissage.

- **Étape 1:** On initialise le dictionnaire, par exemple, par le vecteur moyen de toute la base d'apprentissage.
- **Étape 2:** Connaissant ce dictionnaire, on étiquette chaque vecteur de base d'apprentissage par le numéro de son plus proche voisin. On détermine la partition optimale (la règle de plus proche voisin).
- **Étape 3:** A partir de tous les vecteurs étiquetés par le même numéro, on en déduit un nouveau représentant par un calcul de la moyenne (la règle de centroide).
- **Étape 4:** Si le nombre de centroides désirés n'est pas atteint, on applique une technique de « Spliting », celle-ci consiste à découper chaque centroide C_i en deux nouveaux vecteurs $C_i + \varepsilon$ et $C_i - \varepsilon$ (ε étant un vecteur de perturbation), avant d'appliquer au nouveau dictionnaire obtenu les itérations 2 et 3.

On arrête cet algorithme si on atteint le nombre de centroides désirés pour représenter les vecteurs acoustiques.

1.1.2. Optimisation des paramètres par l'algorithme EM

L'idée principale de l'algorithme Estimation Maximisation (Expectation-Maximization, EM) est, en commençant par les paramètres initiaux θ du modèle, on estime les nouveaux paramètres θ, telle que la vraisemblance du nouveau modèle soit supérieure ou égale à la vraisemblance du modèle initial. L'algorithme EM fait intervenir à la fois des observations X et des variables manquantes (l'indice de la gaussienne $m = 1, \dots, M$). Cet algorithme maximise, de façon itérative, la fonction de la vraisemblance. Cette maximisation n'est pas directe, elle fait intervenir la fonction auxiliaire $Q(\theta, \theta^{(t)})$ qui est définie comme étant l'espérance mathématique du

logarithme de la vraisemblance jointe (incluant les variables observées et les variables cachées) sur l'ensemble complet des variables d'entraînement, calculée à la base des paramètres courants, à savoir :

$$Q(\theta, \theta^{(t)}) = \sum\sum p(m|x_n, \theta^{(t)}) \times \log p(x_n, m|\theta) \quad (2.8)$$

Où θ désigne l'ensemble des paramètres à estimer et $\theta^{(t)}$ l'ensemble des paramètres estimés à l'itération t. Ce qui donne, après calcul :

$$Q(\theta, \theta^{(t)}) = \sum\sum \gamma_{n,m}^{(t)} \left[\log \pi_m - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_m|\right] - \sum_{m=1}^{M}\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\frac{1}{2}(x_n - \mu_m)\right] \quad (2.9)$$

Où $\gamma_{n,m}^{(t)}$ est une probabilité a posteriori estimée à l'itération *t* :

$$\gamma_{n,m}^{(t)} = \frac{\pi_m^{(t)}\, p(x_n|\mu_m^{(t)}, \Sigma_m^{(t)})}{\sum_{m=1}^{M} \pi_k^{(t)}\, p(x_n|\mu_m^{(t)}, \Sigma_m^{(t)})} \quad (2.10)$$

En supposant que $p(x_m|\theta)$ sont des densités gaussiennes à matrices de covariance diagonales, l'expression de la fonction auxiliaire devient :

$$Q(\theta, \theta^{(t)}) = \sum_{m=1}^{M}\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \log \pi_m - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[Cste + \log \sigma_m^2 + \frac{(x_n - u_m)^2}{\sigma_m^2}\right] \quad (2.11)$$

Où σ_m^2 est un élément diagonal de la matrice de covariance.

Les paramètres sont estimés en annulant la dérivée partielle de la fonction auxiliaire Q par rapport à chacun de ceux-ci. Le cas des poids des composantes de mélange π_m est assez simple puisqu'il s'agit de paramètres scalaires. Ceci dit, il faut tenir compte de la contrainte qui existe sur ces paramètres. La maximisation sous contrainte se résout simplement en introduisant un multiplicateur de Lagrange associé à cette contrainte et l'on obtient :

$$\pi_m^{\{t+1\}} = \frac{1}{N}\sum_{n=1}^{N} \gamma_{n,m}^{\{t\}} \quad (2.12)$$

En ce qui concerne les vecteurs des moyennes, on montre que les formules de ré-estimations sont données par :

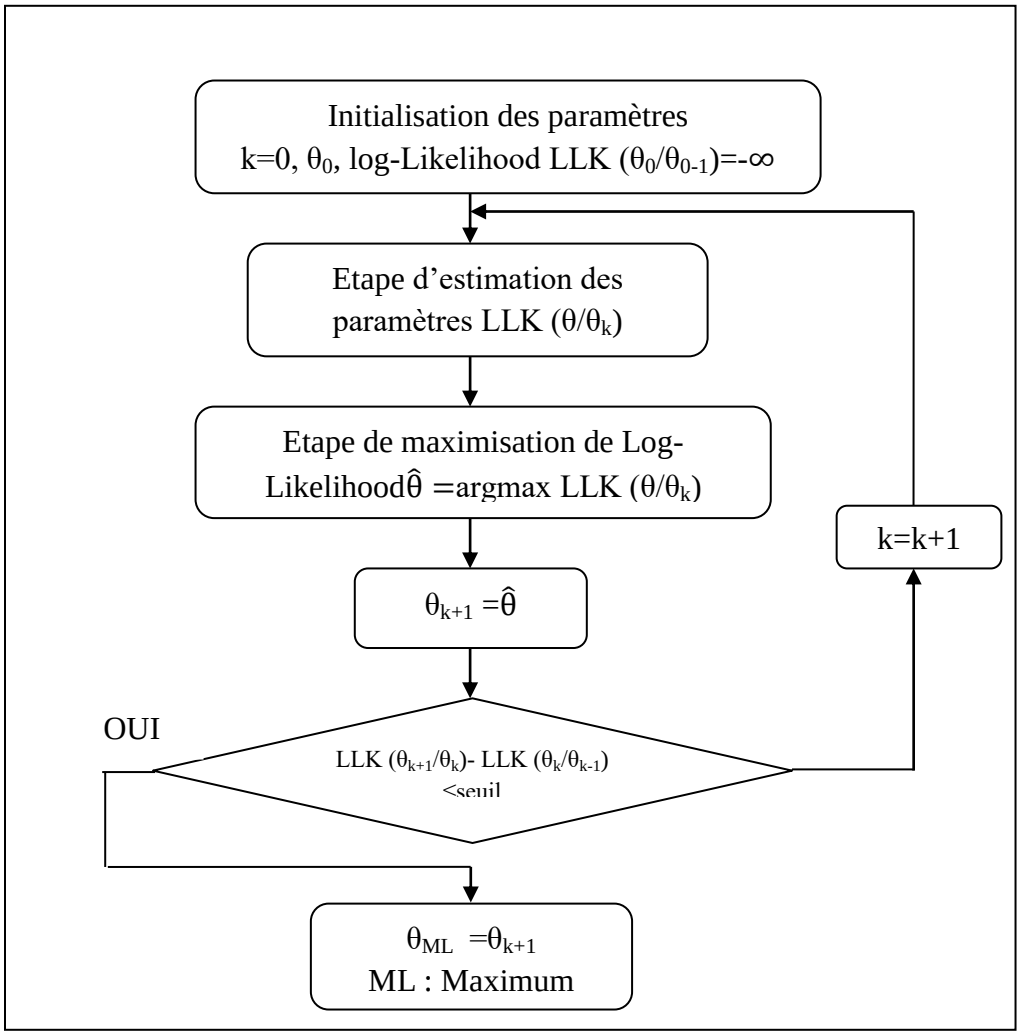

Figure 2.3 Schéma de fonctionnement de l'algorithme EM.

$$\mu_m^{\{t+1\}} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{\{t\}} . x_n}{\sum_{n=1}^{N} \gamma_{n,m}^{\{t\}}} \qquad (2.13)$$

Et pour les variances :

$$\sigma_m^{2\{t+1\}} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{\{t\}} (x_n - \mu_m^{\{t\}})^2}{\sum_{n=1}^{N} \gamma_{n,m}^{\{t\}}} \qquad (2.14)$$

1.2. La phase de classification ou de décision

Dans la phase de classification (la sélection du locuteur le plus probable), on a un groupe de L locuteurs *G*, qui sont représentés, respectivement, par les modèles GMM: $\theta_1, \theta_2, \dots, \theta_L$. Dans une identification du locuteur, l'objectif est de trouver le modèle du locuteur qui a le maximum

de vraisemblance pour une séquence d'observations données. Formellement, cela peut être écrit comme suit :

$$\hat{l} = arg\max_{1\le l\le L}\ p(\theta_l|x) = arg\max_{1\le l\le L}\frac{p(x|\theta_l)Pr(\theta_l)}{p(x)} \qquad (2.15)$$

Supposons que tous les locuteurs ont la même probabilité$p(x) = \frac{1}{L}$, la formule de classification se simplifie comme suit :

$$\hat{l} = arg\max_{1\le l\le L}\ p(x|\theta_l) \qquad (2.16)$$

En utilisant le logarithme et l'indépendance entre les observations, la formule précédente peut être écrite comme suit :

$$\hat{l} = \arg\max_{1\le l\le L}\sum_{t=1}^{T}\log p(x_t|\theta_l) \qquad (2.17)$$

Où $p(x_t|\theta_l)$ est donnée par la formule (2.1).

2. Le modèle du monde UBM et l'adaptation MAP

La modélisation des locuteurs par la méthode GMM-UBM (Reynolds et al., 2000) décrite et schématisée (Figure 2.4) ci-dessous, nécessite un modèle global (modèle du monde UBM) qui représente théoriquement l'ensemble des paramètres de la population en question. Les paramètres du modèle UBM sont entraînés à partir des données de toute la population via l'algorithme EM, et qui sera le modèle initial pour l'algorithme MAP (Maximum A Posteriori).

La technique MAP est une technique de modélisation prometteuse, elle est devenue l'une des techniques les plus utilisées pour estimer le modèle GMM d'un locuteur (Reynolds et al., 2000). L'idée de base de cette méthode est de générer des modèles client à partir de l'adaptation du modèle UBM, cette technique est très utilisée dans les cas où on a un manque de données d'apprentissage.

Tout comme l'algorithme EM, l'adaptation MAP comporte deux étapes de traitement : La première étape est le calcul des paramètres statistiques des données d'apprentissages par rapport au modèle initial. Ces paramètres sont le poids, le moment d'ordre 1 et d'ordre 2 de chaque gaussienne du modèle initial. Etant donné le modèle GMM initial et l'ensemble des vecteurs

d'apprentissage$x = (x_1, \dots, x_D)$, on détermine d'abord le poids de chaque vecteur par rapport à chaque gaussienne k du modèle :

$$\mathrm{P}(k/x_n) = \frac{\omega_k g(x_n, \mu_k, \Sigma_k)}{\sum_{n=1}^{N} \omega_k g(x_n, \mu_k, \Sigma_k)} \qquad (2.18)$$

Cette probabilité est utilisée pour calculer les paramètres suivants :

$$n_k = \sum_{=1}^{N} P\,(k/x_n) \qquad (2.19)$$

$$E_k(X) = \frac{1}{n_k} \sum_{=1}^{N} P\,(k/x_n) x_n \qquad (2.20)$$

$$E_k(X^2) = \frac{1}{n_k} \sum_{n=1}^{N} P\,(k/x_n) {x_n}^2 \qquad (2.21)$$

Dans la seconde étape de l'adaptation, les nouveaux paramètres, estimés dans la première étape, sont combinés avec les paramètres du modèle initial en utilisant des coefficients de pondération qui dépendent des données d'apprentissage. Les coefficients de pondération varient en fonction des données d'apprentissage, de telle sorte que: les composantes gaussiennes comportant une grande quantité de données sont fortement pondérées dans l'estimation des paramètres finaux du modèle, et les gaussiennes avec une petite quantité de données d'apprentissage sont faiblement pondérées (Figure 2.5). L'adaptation se fait par l'utilisation des relations suivantes :

$$\omega_{kadapté} = \left[\frac{\alpha_k^p n_k}{N} + (1 - \alpha_k^p p_k)\right]\gamma \qquad (2.22)$$

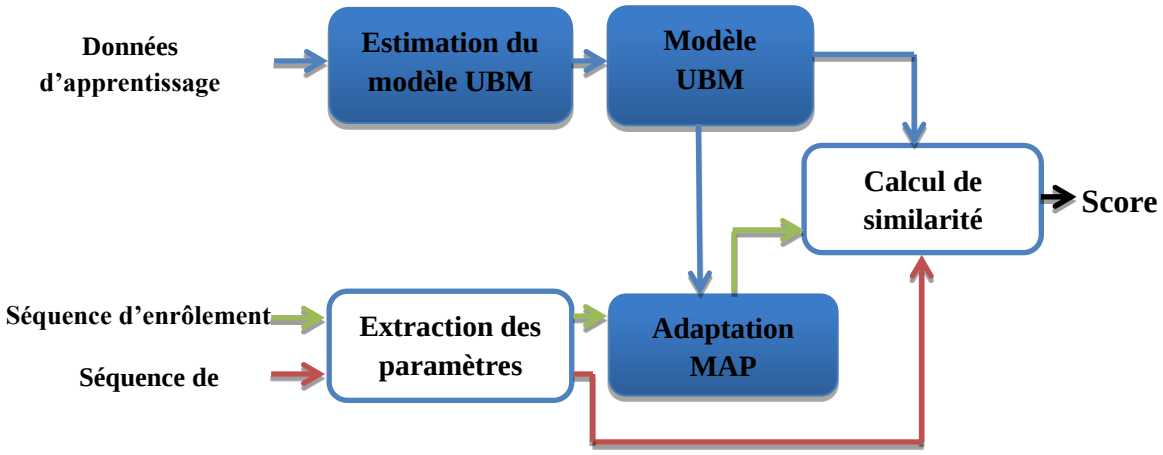

Figure 2.4 Architecture du système RAL à base de GMM-UBM.

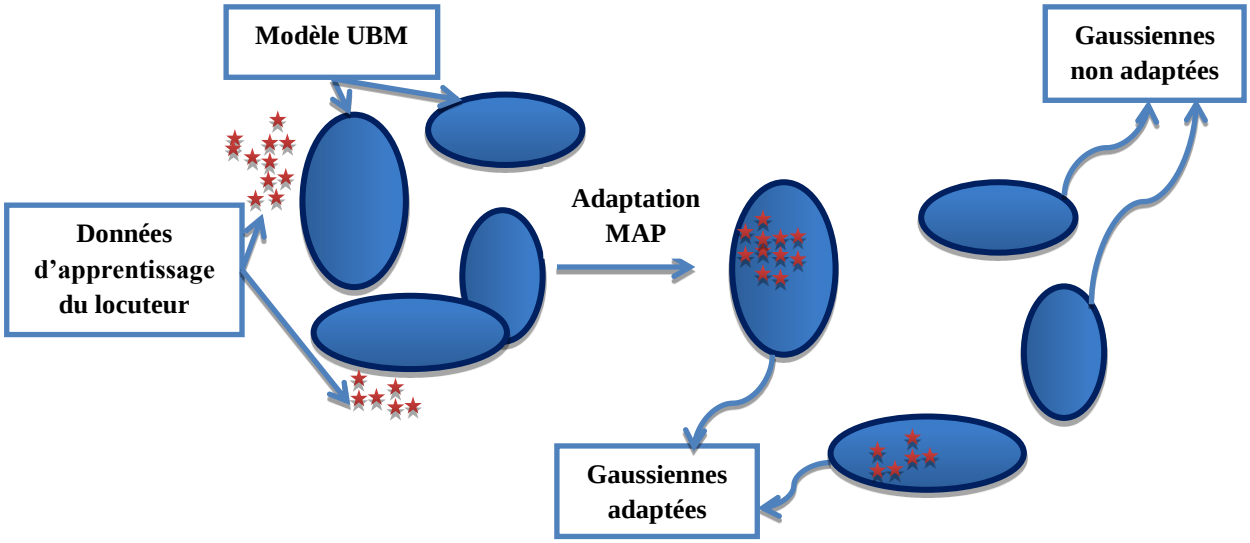

Figure 2.5 Adaptation MAP d'un modèle GMM-UBM.

$$\mu_{kadapté} = \alpha_k^{\mu} E_k(X) + (1 - \alpha_k^{\mu})\mu_k \qquad (2.23)$$

$$\sum_{kadapté} = \alpha_k^{\Sigma} E_k(X^2) + \left(1 - \alpha_k^{\Sigma}\right)(\Sigma_k^2 + \mu_k^2) - \mu_{kadapté}^2 \qquad (2.24)$$

Le facteur γ est calculé après que tous les poids aient été adaptés pour assurer que leur somme soit égale à 1.

Les coefficients d'adaptation α_k^p où $p \epsilon (\omega, \mu, \Sigma)$ servent pour la pondération entre les anciens et les nouveaux paramètres: poids, moyennes et matrice de variances-covariances respectivement. Le coefficient d'adaptation est défini par le facteur :

$$\alpha_k^p = \frac{n_k}{n_k + r_p} \qquad (2.25)$$

Où r_p est un facteur de confiance (relevance factor), fixé pour chaque paramètre p.

3. L'analyse jointe de facteur JFA

L'analyse jointe de facteur proposée par Kenny et al (Kenny, Boulianne, & Dumouchel, 2005; Kenny, Boulianne, Ouellet, & Dumouchel, 2007a) est une technique de modélisation de la variabilité intra-locuteur et de la variabilité inter-session en décomposant le modèle du locuteur en deux composantes, l'une dépendante du locuteur (S) et l'autre dépendante du canal (C) (figure 2.6):

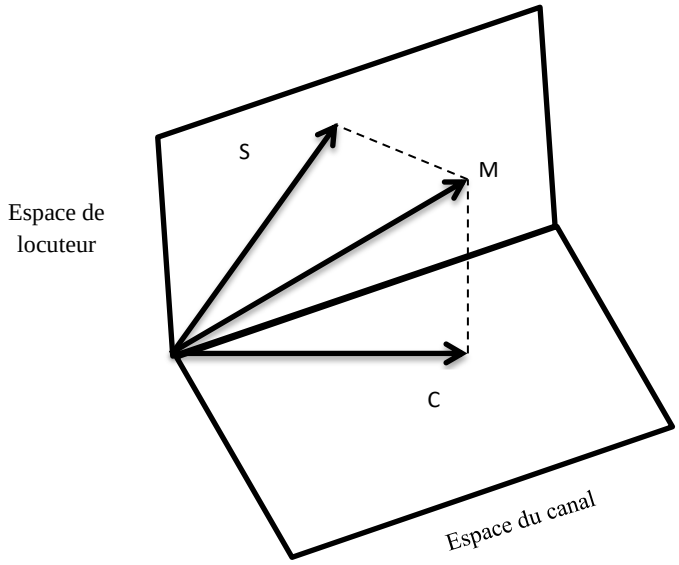

Figure 2.6 Décomposition du supervecteur du locuteur.

$$M_s = S + C \tag{2.26}$$

Où :

$$S = M + Vy_s + Dz_s \tag{2.27}$$

Et :

$$C = Ux_{h,s} \tag{2.28}$$

M_s est le supervecteur du client s de taille $C \times F$ (C est le nombre des composantes GMM et F est la dimension des vecteurs des paramètres). M est le supervecteur du modèle UBM de taille $C \times F$ aussi; Vcaractérisant la variabilité inter-locuteurs et$y_s$sont les facteurs du locuteur (speaker factors) de taille $R_s \times 1(R_s \ll CF)$, il représente les coordonnées du locuteur s dans l'espace formé par la matrice VV^T. Le terme Dz_s décrit les variabilités restantes des locuteurs qui ne sont pas capturées par l'espace du locuteur, et U est la matrice de variabilité du canal (session) de taille $CF \times R_c(R_c \ll CF)$. $x_{h,s}$ est le vecteur des facteurs du canal qui représente les coordonnées de la session h dans l'espace de variabilité du canal formé par la matrice UU^T.

Les matrices V,D et U sont appelées les hyper paramètres du modèle JFA. Elles sont estimées à partir d'un grand ensemble de locuteurs où chaque locuteur dispose de plusieurs enregistrements. L'estimation est faite par l'algorithme EM. On commence tout d'abord par estimer la matrice V en supposant que les deux autres matrices sont nulles. Etant donné

l'estimation de la matrice V, on estime ensuite la matrice U en gardant toujours la matrice D nulle, à la fin on estime la matrice D en considérant l'estimation des deux autres matrices. Nous renvoyons le lecteur à la référence (Kenny, 2005) pour plus de détails sur l'estimation de ces paramètres.

4. Approche I-vecteur

La modélisation JFA classique basée sur les facteurs du locuteur et du canal consiste à définir deux espaces distincts : l'espace du locuteur et l'espace du canal. L'approche I-vecteur (Dehak et al., 2011) est basée sur la définition d'un seul espace, au lieu de deux séparés (Figure 2.7). Ce nouvel espace, que nous appelons l'espace de variabilité totale, contenant les variabilités du locuteur et du canal est défini par une matrice T de variabilité totale contenant à son tour les vecteurs propres avec les plus grandes valeurs propres de la matrice de covariance de cette variabilité. Dans le nouveau modèle, aucune distinction n'est faite entre les effets du locuteur et ceux du canal dans l'espace des supervecteurs GMM.

Cette nouvelle approche est motivée par les expériences menées, montrant que les facteurs du canal du JFA, qui normalement modélisent seuls les effets du canal, contiennent également des informations sur le locuteur aidant à leur discrimination (Dehak et al., 2009). Le nouveau supervecteur GMM dépendant du canal et du locuteur est alors défini comme suit:

$$m(s) = M + Tw(s) \qquad (2.29)$$

Où M est le supervecteur indépendant du canal et du locuteur (pouvant être considéré comme supervecteur UBM de taille $C \times D$, avec C le nombre des composantes GMM et D la dimension des vecteurs des paramètres), T est une matrice rectangulaire représentant la matrice de variabilité totale dont les étapes d'estimation seront cités ci-dessous, et w est le vecteur d'identité (I-vecteur) (Dehak et al., 2011).

4.1. Estimation de la matrice T

Dans l'estimation de la matrice T les séquences de parole du même locuteur sont traitées comme si elles étaient produites par différents locuteurs. Les étapes d'estimation de cette matrice sont données par :

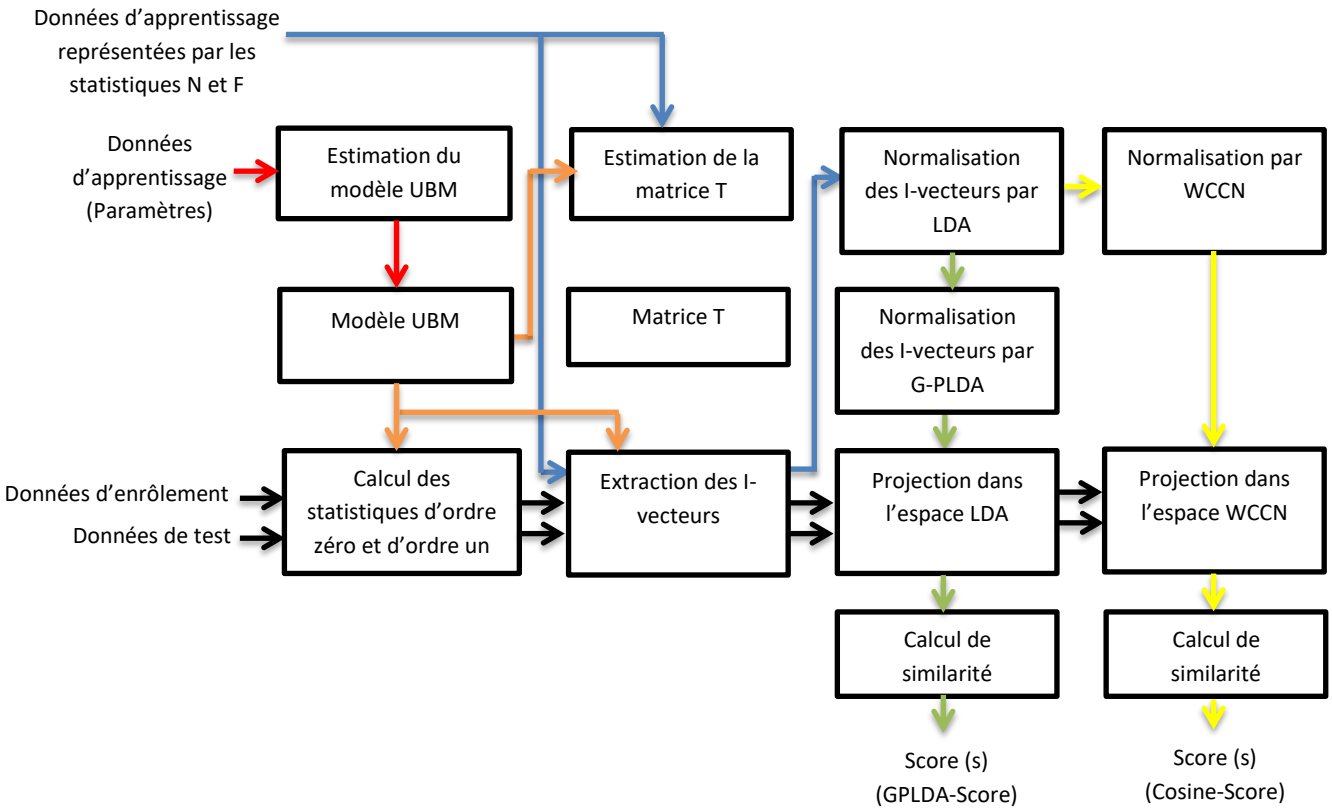

Figure 2.7 Architecture du système RAL à base de I-vecteur.

1. Pour chaque énoncé u :

- Calculer les statistiques d'ordre zéro et d'ordre un $N(u)$ et $F(u)$ respectivement :

$$N(u) = \sum_{t \epsilon u} \gamma_t(c) \qquad (2.30)$$

$$F(u) = \sum_{t \epsilon u} \gamma_t(c) X(t) \qquad (2.31)$$

Avec ; $\gamma_t(c)$ c'est la composante gaussienne cà posteriori pour l'observation $t$du l'énoncé u et $X(t)$ est le vecteur acoustique.

- Centrer la statistique de premier ordre $\tilde{F}(u) = F(u) - N(u)M$.

2. Initialiser aléatoirement la matrice T
3. Répéter pour un certain nombre d'itérations EM.

- Pour chaque énoncé u

$$G(u) = (I + T^T \Sigma^{-1} N(u) T)^{-1}$$

- Calculer la matrice

- Calculer les statistiques du vecteur $w(u)$

$$E_1(u) = E(w) = G(u)T^T\Sigma^{-1}\tilde{F}(u)$$

$$E_2(u) = E(ww^T) = E_1(u)E_1^T(u) + G(u)$$

- Résoudre le système d'équations

$$\Sigma_u N(u)\tilde{T}E_2(u) = \Sigma_s\tilde{F}(u)E_1^T(u)$$

- Remplacer T par $\tilde{T}$.

4.2. Estimation des paramètres du locuteur

Dans le système à base de I-vecteur (Dehak et al., 2011), un locuteur est représenté par son vecteur d'identité w défini par :

$$w(s) = (I + T^T\Sigma^{-1}N(s)T)^{-1}T^T\Sigma^{-1}(F(s) - N(s)M) \quad (2.32)$$

4.3. Normalisation des I-vecteurs

La combinaison de l'analyse discriminante linéaire (Linear Discriminant Analysis, LDA) (Fisher, 1936) avec la méthode de la normalisation par la matrice de covariance intra-classe (Within Class Covariance Normalization, WCCN) (Hatch, Kajarekar, & Stolcke, 2006) est devenue un standard pour l'implémentation des systèmes de vérification du locuteur à base de la similarité angulaire.

4.3.1. Analyse discriminante linéaire LDA

Pour compenser l'effet du canal dans les systèmes I-vecteur, la méthode LDA est utilisée. Cette méthode sert à minimiser la variabilité intra-locuteur et maximiser la variabilité inter-locuteur. La matrice de projection (A) de cette méthode est obtenue en résolvant le problème des valeurs propres suivant :

$$S_B v = \lambda S_W \quad (2.33)$$

Où S_B et S_W sont respectivement la matrice de variabilité inter-classes et la matrice de variabilité intra-classe.

$$S_W = \sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i}(w_{i,j} - \overline{w}_i)(w_{i,j} - \overline{w}_i)^T \quad (2.34)$$

$$S_B = \sum_i J_i\,(\overline{w}_i - \overline{w})(\overline{w}_i - \overline{w})^T \quad (2.35)$$

I est le nombre des locuteurs, J_i est le nombre des I-vecteurs pour chaque locuteur i, $\overline{w}_i$ est la moyenne des I-vecteurs pour chaque locuteur i et $\overline{w}$ est la moyenne de tous les I-vecteurs.

4.3.2. La Matrice de Covariance Intra-Classe (WCCN)

La normalisation de la variance des données en utilisant l'inverse de la matrice de covariance intra-classe est devenue une pratique courante dans le domaine de la reconnaissance du locuteur (Dehak, Dehak, Glass, Reynolds, & Kenny, 2010; Dehak et al., 2011; Senoussaoui, Kenny, Dumouchel, & Dehak, 2013). Cette méthode a été proposée pour la première fois par Hatch (Hatch et al., 2006) qui l'avait utilisée pour améliorer les performances d'un classificateur à vaste marge (SVM). L'idée de base de cette normalisation est de pénaliser les axes contenant une forte variance intra-classe. Cette pénalisation est réalisée par une rotation de données via la matrice triangulaire inférieure K'_w obtenue par la décomposition de Cholesky de l'inverse de la matrice de covariance intra-classe :

$$\widehat{w} = K'_w w \qquad (2.36)$$

$c_w = \frac{S_w}{R}$ est la matrice de covariance intra-classe, R représente le nombre total des données d'apprentissage et $C_W^{-1} = K_w K'_w$.

Afin d'être plus précis, la procédure standard de la vérification du locuteur via la similarité angulaire consiste à estimer la matrice C_w à partir des données préalablement projetées dans l'espace de dimension réduite de la LDA.

4.4. Analyse Discriminante Linéaire Probabiliste PLDA

Afin de modéliser directement la variabilité de la session et du locuteur dans l'espace I-vecteur, une technique d'analyse discriminante linéaire probabiliste PLDA a été proposée par Kenny (Kenny, 2010). Cette approche peut être vu comme étant très similaire à l'approche JFA, mais en utilisant les I-vecteurs qui sont dépendants du locuteur et du canal, plutôt que les super-vecteurs GMM comme base pour la modélisation des facteurs. Le modèle G-PLDA (Garcia-Romero & Espy-Wilson, 2011), suppose que chacun de ces I-vecteurs se décompose comme suit:

$$w = m + U_1 x_1 + U_2 x_2 + \varepsilon \qquad (2.37)$$

Où U_1 est la matrice des voix propres et U_2 la matrice des canaux propres, x_1 et x_2 sont respectivement les facteurs du locuteur et du canal, et ε est le résiduel du locuteur supposé

gaussien. La distribution normale est utilisée pour x_1,x_2 et ε. Ce modèle est donc formé de deux parties: $m + U_1 x_1$ étant la composante qui ne dépend que du locuteur et ne décrit que la variabilité inter-locuteurs, et celle du canal $U_2 x_2 + \varepsilon$ qui ne dépend que de ce dernier et ne décrit que la variabilité intra-locuteurs.

4.5. Calcul des scores

4.5.1. via la similarité angulaire du cosinus

La vérification du locuteur à base de la similarité angulaire du cosinus consiste à calculer la similarité entre les deux I-vecteurs composant un essai de vérification. Étant donné deux I-vecteurs w_{client}, w_{test}, l'équation de la similarité angulaire du cosinus S s'écrit comme suit:

$$S(w_e, w_t) = \frac{{w_{client}}' . w_{test}}{\|w_{client}\| \|w_{test}\|} \qquad (2.38)$$

Où $\|.\|$ est la norme euclidienne d'un vecteur.

4.5.2. Via le modèle PLDA

Le calcul des scores pour I-Vecteur basé sur GPLDA utilise le rapport de vraisemblance (Kenny, 2010). Pour deux I-vecteurs w_{client}, w_{test}, le calcul du rapport de vraisemblance est donné par:

$$\ln\left(\frac{P(w_{client}, w_{test}|H_1)}{P(w_{client}|H_0)P(w_{test}|H_0)}\right) \qquad (2.39)$$

Où :

H_1: les deux locuteurs sont les mêmes.

H_0: les deux locuteurs sont différents.

5. Conclusion

Dans ce chapitre, nous avons étudié deux méthodes de modélisation des paramètres acoustiques. Nous avons commencé par détailler le modèle statistique le plus utilisé dans les systèmes RAL en mode indépendant du texte, à savoir, le modèle GMM-UBM. Par la suite, nous avons donné une définition de l'espace des i-vecteurs ainsi la procédure détaillée de leurs extractions.

Chapitre 3

Nouveaux paramètres acoustiques pour la reconnaissance robuste

Sommaire

Résumé

La variation de la nature du signal acoustique rend le traitement des données brutes issues de ce dernier très difficile. En effet, ces données contiennent des informations complexes, souvent redondantes et mélangées au bruit. Le module de la paramétrisation, qui traite le signal acoustique reçu, doit remplir plusieurs objectifs: séparer le signal du bruit; extraire l'information utile à la reconnaissance et convertir les données brutes à un format directement exploitable par le système. Chacune de ces taches pose des problèmes complexes et influe fortement sur les résultats des systèmes automatiques de reconnaissance. Dans le présent chapitre, nous présentons quelques paramètres utilisés en reconnaissance automatique du locuteur. Nous introduisons par la suite des nouveaux paramètres acoustiques fondés sur le système auditif humain en utilisant les filtres Gammatones et exploitant le module et la phase du spectre de fréquence.

1. Extraction des paramètres acoustiques

1.1. Etat de l'art

La plupart des systèmes RAL utilisent les coefficients cepstraux à fréquence Mel (MFCC) (Davis & Mermelstein, 1980) en raison de leurs propriétés de représentation, notamment la décorrélation des coefficients. Cependant, ces paramètres souffrent de plusieurs limitations, en particulier, leurs sensibilités aux conditions d'acquisition du signal et à l'environnement acoustique (Davis & Mermelstein, 1980). Nous citons:

(i) l'estimation spectrale à court terme est basée sur l'utilisation des fenêtres symétriques qui ne sont pas robustes au bruit (Kinnunen et al., 2012).

(ii) l'échelle Mel standard n'est pas un modèle auditif optimal (Zhou, Garcia-Romero, Duraiswami, Espy-Wilson, & Shamma, 2011).

(iii) la non-linéarité auditive performe la non-linéarité logarithmique adoptée dans les MFCC standards (Kim & Stern, 2012).

(iv) la négligence des informations issues du module de phase.

La négligence de l'information de phase a été motivée par des expériences d'écoute humaines indiquant que le spectre de phase à long terme transmet une quantité négligeable d'information (Alsteris & Paliwal, 2007). Mais récemment, des études suggèrent que l'incorporation de l'information de phase peut améliorer les performances de nombreux systèmes du traitement du signal vocal (Gerkmann, Krawczyk-Becker, & Le Roux, 2015; Mowlaee, Saeidi, & Stylianou, 2016; Vijayan, Reddy, & Murty, 2016).

La raison principale pour laquelle la phase n'a pas été pleinement exploitée réside dans la difficulté de traiter sa nature (Ying, 2006). Cependant, les chercheurs ont essayé d'utiliser d'autres représentations. Les plus communs sont la fréquence instantanée et le retard de groupe du spectre de phase. Ces deux représentations sont obtenues respectivement en prenant la dérivée du spectre de phase sur la fréquence et le temps. Dans (Paliwal & Atal, 2003), les fréquences instantanées proposées comme caractéristiques pour la reconnaissance de la parole, ont obtenu des performances comparables avec les coefficients MFCC. La représentation de la fonction de retard de groupe (Group Delay Function, GDF) a été longtemps considérée comme une méthode alternative pour l'estimation du spectre. Certains auteurs (Hegde, Murthy, & Gadde, 2007) ont proposé la fonction de retard de groupe modifiée (Modified Group Delay Function, MGDF) en

remplaçant le spectre de puissance par le spectre de puissance cepstral lissé. De plus, ils ont appliqué la transformée en cosinus discrète (Discrete Cosine Transformation, DCT) sur le MGDF pour obtenir les coefficients cepstraux du retard de group modifié (Modified Group Delay Cepstral Coefficients, MGDCC). Récemment, Madikeri et al. (Madikeri, Talambedu, & Murthy, 2015) ont utilisé les paramètres MGDCC pour la vérification du locuteur dans l'espace de variabilité totale. Les résultats obtenus montrent une amélioration significative de la performance en termes de taux de vérification sur l'ensemble de données de référence 2010 du NIST SRE.

Donglai et Paliwal dans (Zhu & Paliwal, 2004) ont proposé des paramètres qui dépendent du produit du spectre de puissance et de la fonction de retard de groupe, et qui résultent des coefficients cepstraux du produit spectral à fréquence Mel (Mel Frequency Product Spectrum Cepstral Coefficients, MFPSCC). Le spectre résultant est une combinaison entre le spectre d'amplitude et le spectre de phase. Les mêmes auteurs ont également mené une étude comparative entre le produit du spectre, le retard de groupe modifié et le MFCC pour la tâche de reconnaissance de la parole. Ils arrivent à la conclusion que le produit du spectre fournit les meilleures performances.

Outre ces efforts d'exploiter l'information de la phase, et inspiré de la robustesse de système auditif humain au bruit et à leur grande capacité de séparation de source, des recherches ont été menées sur la conception et le développement de nouveaux systèmes basés sur les filtres auditifs et plus particulièrement l'utilisation des filtres Gammatones (Q. Li & Huang, 2011; Z. Li & Gao, 2016). Le filtre Gammatone proposé dans (Bonastre et al., 2003) a attiré plusieurs chercheurs travaillant sur la reconnaissance de la parole et de locuteur. Récemment des nouveaux paramètres appelée coefficients cepstraux à fréquence Gammatone (Gammatone Frequency Cepstral Coefficient, GFCC) pour la RAL ont démontré une grande robustesse au bruit et performent mieux que les coefficients MFCC (Z. Li & Gao, 2016; Shao & Wang, 2008; Zhao & Wang, 2013)

Fedila et al. dans (Meriem, Farid, Messaoud, & Abderrahmene, 2014) ont proposé de nouveaux paramètres qui dépendent des filtres Gammatones et les fenêtres de pondération ou Multitaper et qui résultent des coefficients cepstraux Mulitaper Gammatone (Mulitaper Gammatone Cepstral Coefficients, MGCC). Ils ont montré que ces paramètres MGCC dérivés de filtres Gammatone sont plus performant que ceux dérivés de filtres Mel classiques. Les même auteurs, et afin de coupler la robustesse de filtres Gammatone aux bruits et la capacité de

représentation du produit du spectre, ont proposé récemment des nouveaux paramètres appelés coefficients cepstraux du produit spectral et Gammatone (Gammatone Product Spectral Cepstral Coefficients, GPSCC) (Fedila, Bengherabi, & Amrouche, 2015). Ces paramètres GPSCC ont démontré une grande robustesse aux différents types de bruit. Toutes les étapes d'extraction de ces deux nouvelles méthodes à savoir, MGCC et GPSCC seront détaillées dans la section 2 de ce chapitre.

1.2. Extraction des paramètres MFCC

Toutes les étapes d'extraction des coefficients MFCC à partir d'une trame de parole seront décrites dans cette section (Figure 3.1) (Davis & Mermelstein, 1980).

- Une première étape de fenêtrage qui consiste à découper le segment de parole continue en trames pendant lesquelles le signal est supposé quasi-stationnaire. Chaque trame a habituellement une durée identique d'environ 10 à 30 ms. Le décalage entre deux trames consécutives est d'environ 10 à 15 ms. Ensuite, on applique une fenêtre à chaque trame afin de limiter le nombre d'échantillons et de réduire les effets de bords (phénomène de Gibbs). Parmi les différentes fenêtres de pondération, on retrouve: la fenêtre rectangulaire, la fenêtre de Hamming, la fenêtre de Hanning et la fenêtre de Blackmann. En traitement de la parole, la fenêtre de Hamming est la plus utilisée (Figure 3.2).

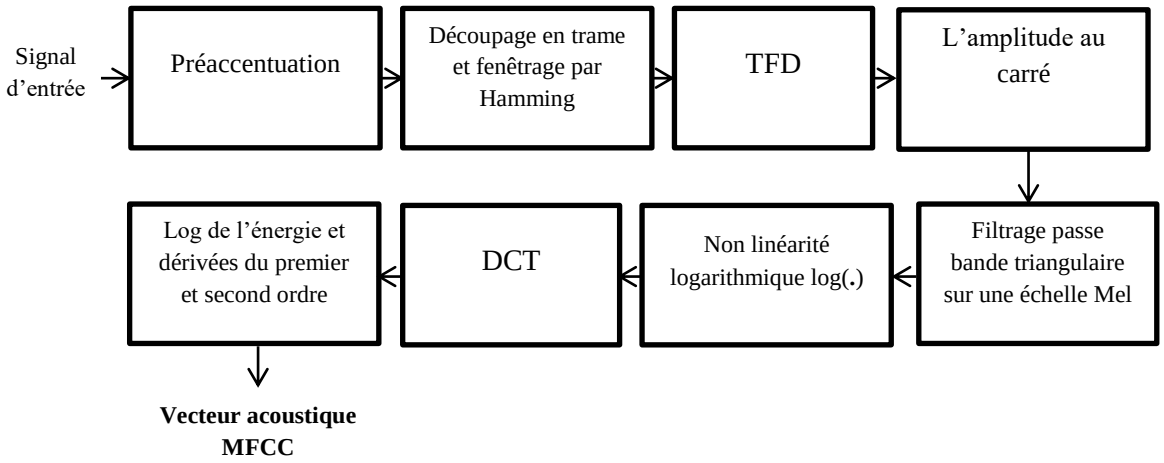

Figure 3.1 Extraction des paramètres MFCC.

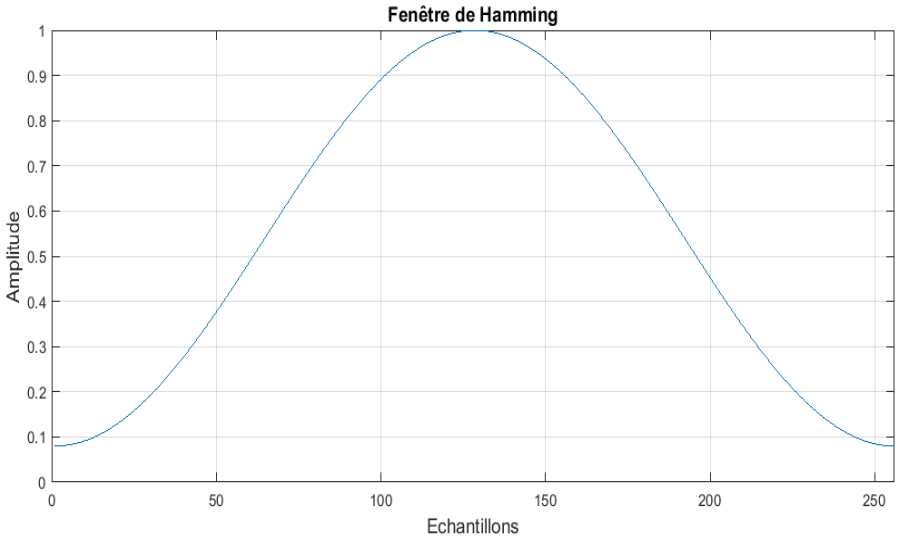

Figure 3.2 Fenêtre de Hamming sur 256 échantillons.

Le signal de la trame résultante à *N* échantillons est donc :

$$s(n) = \sum_{n=1}^{N} x(n)w(n) \qquad (3.1)$$

Avec $s(n)$ Signal de la trame, $x(n)$ Signal à fragmenter et $w(n)$ Fenêtre de longueur $N, \mathrm{n} = 1, \ldots\ldots N$.

L'équation de la fenêtre de Hamming est donnée par l'expression suivante (Figure 3.2):

$$w(n) = 0.54 - 046\ cos\left(\frac{2\pi n}{n-1}\right);\ 0 \leq \mathrm{n} \leq\ \mathrm{N} - 1 \qquad (3.2)$$

- La deuxième étape consiste au passage du domaine temporel au domaine fréquentiel par la transformée de Fourier rapide tel que :

$$S[k,t] = \frac{1}{N}\sum_{n=0}^{N-1} x[n,t]e^{-j2\pi nk/N}, k\ =\ 0,1,2,\ldots N-1\ ;\ t\ =\ 0,1,2,\ldots,T-1 \qquad (3.3)$$

Dans le cas complexe, seulement la valeur absolue est considérée. Où la gamme de fréquence $0 \leq \mathrm{f} \leq \frac{f_e}{2}$ correspond à $0 \leq \mathrm{k} \leq \frac{\mathrm{N}}{2} - 1$ et la gamme de $-\frac{f_e}{2} \leq \mathrm{f} \leq 0$correspond à$\frac{\mathrm{N}}{2} + 1 \leq \mathrm{k} \leq \mathrm{N} - 1$.

- L'élévation du module du spectre résultant de la seconde étape au carré représente le spectre de puissance de chaque trame et constitue alors la troisième étape de ce processus.
- Comme l'échelle de la perception fréquentielle de l'oreille humaine n'est pas linéaire, l'étape de traitement qui suit est celle du filtrage du spectre de puissance du signal vocal par une banque de k filtres (Figure 3.3) positionnés selon une échelle Mel. Cette échelle est construite à partir d'une série de filtres passe-bandes de formes triangulaires positionnés d'une façon linéaire pour les basses fréquences (<1000 Hz) et logarithmique pour les hautes fréquences. Le premier filtre qui donne une indication sur la quantité d'énergie existante près de 0 Hertz est très étroit, en passant vers les hautes fréquences cette largeur augmente, ceci simule le fonctionnement de l'oreille humaine qui n'est pas sensible dans ces fréquences et qui ne peut pas discerner entre deux fréquences très proches. Les points de frontières des filtres en échelle de fréquence Mel sont calculés à partir de la formule:

$$B_m = B_b + m\,\frac{B_h - B_b}{M+1}\ ,\quad 0 \le m \le M+1 \qquad (3.4)$$

Avec M est Le nombre de filtres, B_h la fréquence la plus haute du signal et B_b est la fréquence la plus basse du signal.

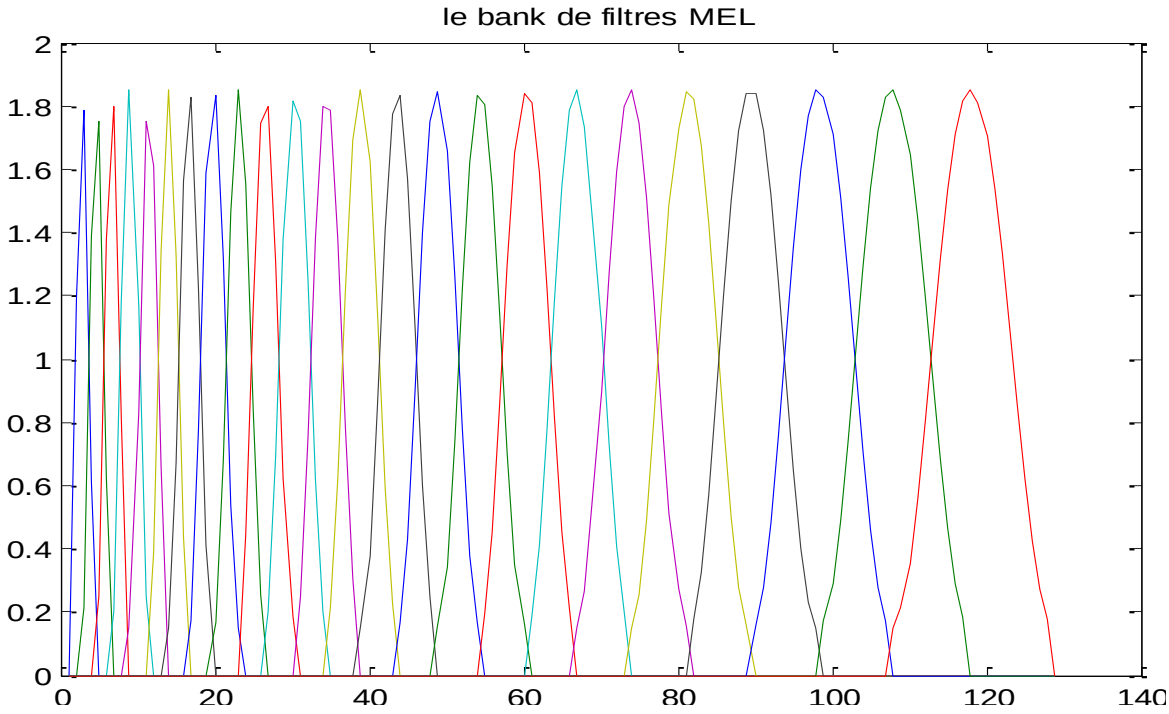

Figure 3.3 La répartition des filtres triangulaires sur les échelles Mel.

Par conséquent, on fait correspondre à chaque fréquence F, mesurée en Hz, une valeur F_{mel} mesurée en Mel, tel que :

$$B(f) = 2595 * log_{10}(1 + \frac{f_{Hz}}{700}) \qquad (3.5)$$

Où f_{Hz} est la fréquence en Hz et $B(f)$ est la fréquence en échelle-Mel de f_{Hz}. Cette transformation non linéaire peut être vue dans la Figure 3.4**.**

A l'issue de cette étape, un vecteur donnant une indication sur la quantité d'énergie contenue dans chaque filtre résulte.

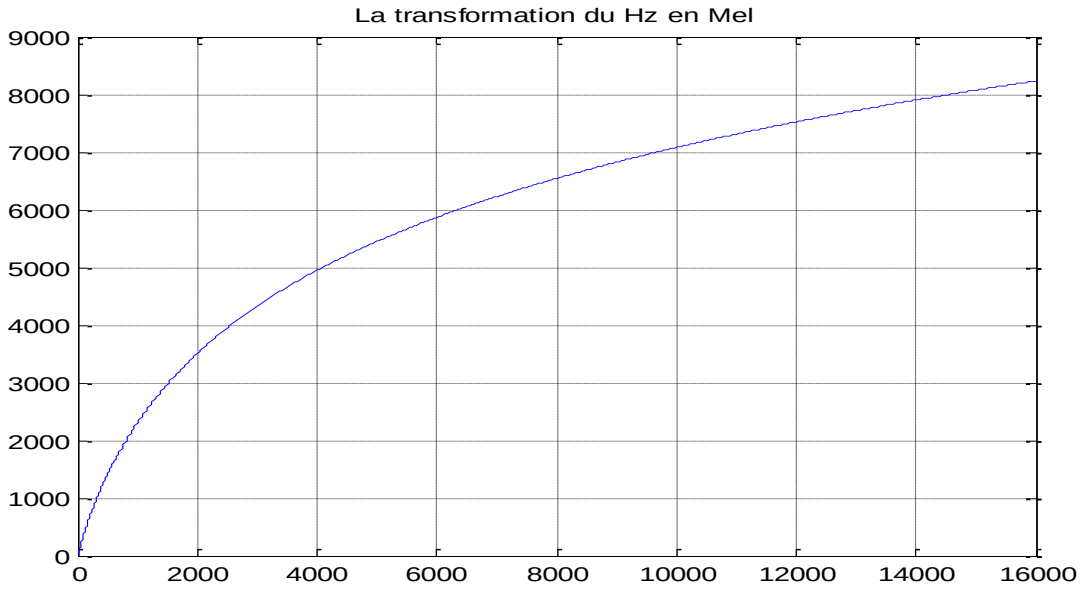

Figure 3.4 La transformation du Hz en Mel.

- Le problème de la reconnaissance est la convolution de la source et du conduit vocal. Afin d'effectuer une déconvolution pour éliminer l'effet de la source, l'analyse cepstrale est effectuée en appliquant le logarithme dans le domaine fréquentiel pour transformer la multiplication en une sommation.
- La dernière étape de ce processus est l'application de la transformée en cosinus discrète (DCT) sur le vecteur d'énergies logarithmiques E_i pour une éventuelle compression de l'information. L'ensemble de ces coefficients cepstraux produits est appelé "vecteur acoustique". L'équation de la DCT est la suivante :

$$c_k = \sum_{i=1}^{M} \log E_i \cos\left[\frac{\pi k}{M}(i - \frac{1}{2})\right]. \qquad 1 \leq k \leq d \qquad (3.6)$$

Avec M le nombre de filtres passe-bande triangulaires, dle nombre de coefficients cepstraux à échelle de Mel.

- Un éventuel calcul des coefficients dynamiques peut être effectué avant de produire le vecteur final de paramètres MFCC, ces coefficients représentent la vitesse et l'accélération tel que : Les coefficients de régression du premier ordre sont calculés par l'équation de régression suivante :

$$d_i = \frac{\sum_{n=1}^{N} n(c_{n+1} - c_{n-1})}{2 \sum_{n=1}^{N} n^2} \qquad (3.7)$$

Où $d_i$est le coefficient Δ de la trame $i$calculé en fonction des coefficients basiques c_{n+1} et c_{n-1}. La même équation est utilisée pour le calcul des coefficients d'accélération (ΔΔ) en remplaçant les coefficients basiques par les coefficients Δ.

2. Proposition de nouveaux paramètres acoustiques

Dans cette section, nous allons donner un détail de deux méthodes nouvelles d'extraction de paramètres acoustiques fondées sur le système auditif humain en utilisant les filtres Gammatones à savoir le MGCC et le GPSCC. Toutes les étapes d'extraction de ces deux paramètres sont détaillées dans la Figure 3.5.

2.1. Détail d'extraction des paramètres GPSCC

Les coefficients GPSCC sont calculés avec quatre étapes.

2.1.1. Etape 1 : Calcul du produit spectral

- Soit la trame du signal parole $x(n)$, n=0,…, N-1 sa transformée de Fourier est donnée par :

$$X(\omega) = |X(\omega)| e^{j\theta(\omega)} \qquad (3.8)$$

On calcule module au carré du spectre d'amplitude de chaque trame de 25 ms en suivant les mêmes étapes que celles citées lors de l'extraction des MFCC,

- On calcul par la suite le produit spectral, défini comme le produit du spectre de puissance par la fonction de retard du Group GDF (Zhu & Paliwal, 2004). L'équation de produit spectral est donnée par :

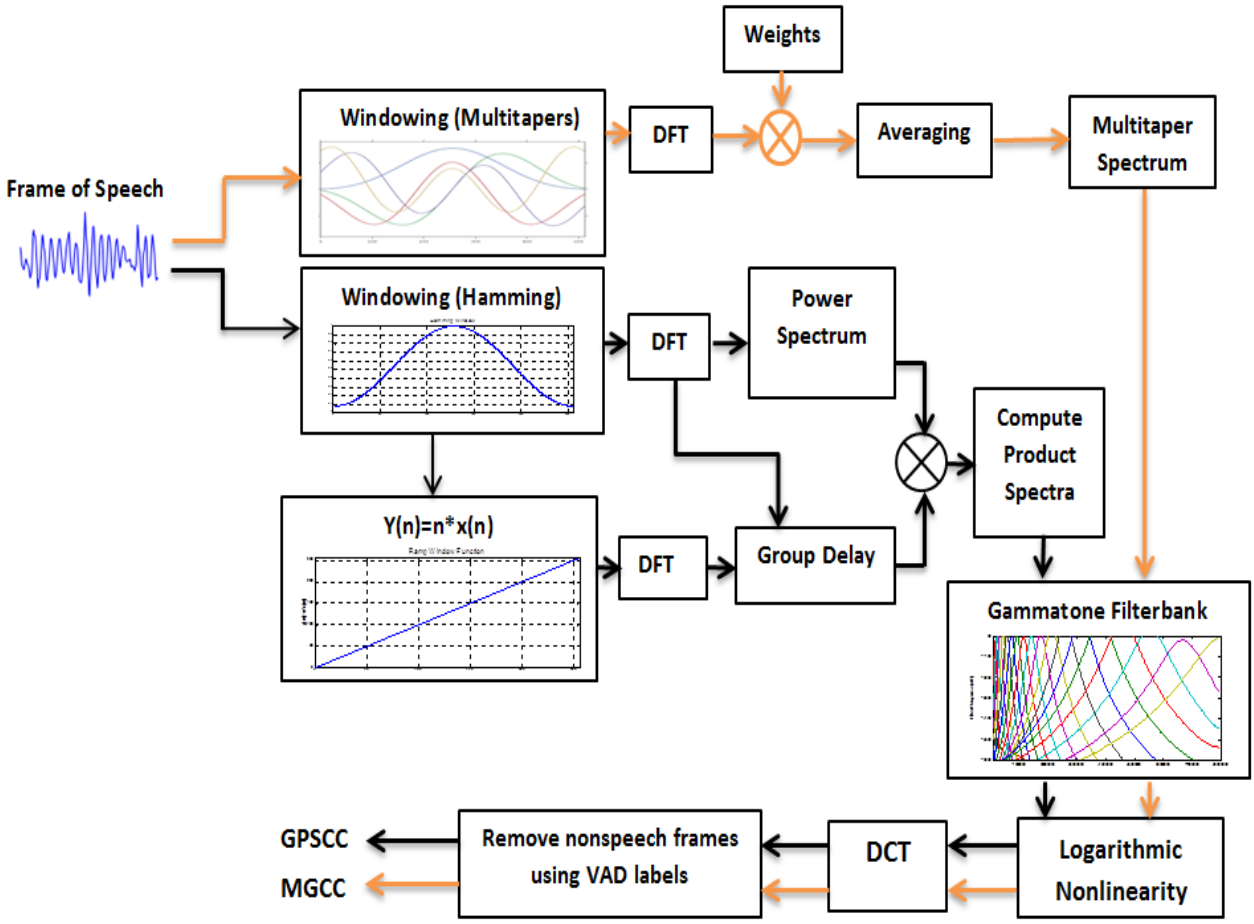

Figure 3.5 Les étapes d'extraction des paramètres GPSCC et MGCC.

$$Q(\omega) = |X(\omega)|^2 \tau_p(\omega) \quad (3.9)$$

$$= X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega) \quad (3.10)$$

Ou $\tau_p(\omega)$est la fonction de retard de groupe (GDF : group Delay function), définie par :

$$\tau_p(\omega) = -\frac{d\theta(\omega)}{d\omega} \quad (3.11)$$

L'équation (3.9) peut être simplifiée comme suit :

$$\tau_p(\omega) = -Im\frac{d(\log(X(\omega)))}{d\omega} \quad (3.12)$$

$$= \frac{X_R(\omega)Y_R(\omega)+X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \quad (3.13)$$

Où $Y(\omega)$ est la transformée de Fourier de $nx(n)$, I et R pour designer respectivement la partie imaginaire et la partie réelle.

- Le spectre résultant du produit spectral est une combinaison entre le spectre d'amplitude et le spectre de phase. Il améliore la région des formants dans la fonction de retard de groupe modifiée et il a une enveloppe comparable à celle du spectre de puissance.

2.1.2. Etape 2 : Utilisation de filtre Gammatone

- La seconde étape consiste à effectuer une intégration fréquentielle en utilisant 40 filtres Gammatone pour obtenir l'énergie issue de chaque banc de filtre.

2.1.2.1. Aperçu sur les systèmes auditif humain et les filtres Gammatones

A. Le système auditif humain

Les systèmes auditifs humains ont toujours été une importante source d'inspiration pour les systèmes de reconnaissance du locuteur. La compréhension du fonctionnement de l'oreille interne, en particulier de la cochlée, a contribué à l'obtention de modèles mathématiques pour ces mécanismes biologiques et à leur application à la reconnaissance du locuteur. La Figure 3.6 montre un diagramme du système auditif humain. Il comporte trois parties :

- L'oreille externe : composée du pavillon et du conduit auditif externe. Le pavillon capte les vibrations sonores et les acheminer via le conduit auditif vers le tympan, située dans l'oreille moyenne.
- L'oreille moyenne : composée du tympan et de la chaîne des osselets. Le tympan convertit les vibrations externes en vibrations mécaniques qui se propagent dans la cavité tympanique. La chaîne ossiculaire de l'oreille moyenne composée du marteau, de l'enclume et de l'étrier reçoit ces vibrations mécaniques qui sont transformées par la suite en ondes de compression et se propagent dans le milieu liquide de la cochlée, située dans l'oreille interne (Pickles, 2008).
- L'oreille interne est composée de la cochlée et ses cellules ciliées. Les cellules ciliées traduisent le message vibratoire en influx nerveux qui sera transmis au cerveau via le nerf auditif où il va être décodé et interprété sous forme de sons.

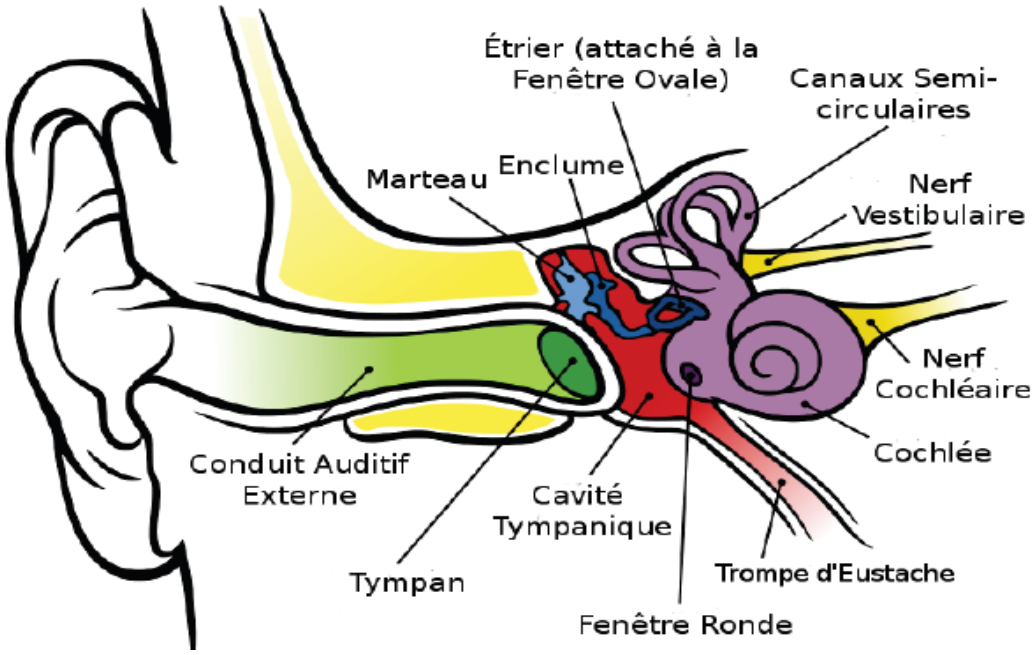

Figure 3.6 Diagramme du système humain (Pinto & Sato, 2016).

La cochlée est un organe enroulé en spirale sur un peu moins de 3 tours et sur une longueur voisine de 35 millimètres et abritant la membrane basilaire. La première extrémité de la cochlée, connectée à la fenêtre ovale et à la fenêtre ronde, est appelée *base* et constitue la partie la plus fine et la plus rigide du tube. Le diamètre de celui-ci (et la largeur de la membrane basilaire) croît progressivement et atteint sa plus grande valeur à l'autre extrémité, plus large et plus souple, appelée *apex* (Gelfand, 2016).

Lorsque une onde traverse la cochlée de la base vers l'apex, son amplitude atteint sa plus grande valeur et décroît ensuite très rapidement à un endroit bien précis de la membrane basilaire qui dépend de la fréquence. Les fréquences aigües (hautes fréquences) agissent à la base de la cochlée, elles perdent donc rapidement en amplitude et ne se propagent pas vers la fin de la membrane basilaire. Les fréquences graves (basses fréquences) quant à elles agissent à l'apex et peuvent continuer à se propager le long de la membrane basilaire. La Figure 3.7 illustre une carte tonotopique (Romani, Williamson, & Kaufman, 1982) montrant les parties activées de la membrane basilaire et les bandes de fréquences correspondantes[2].

[2]http://www.cochlea.eu/cochlee/fonctionnement

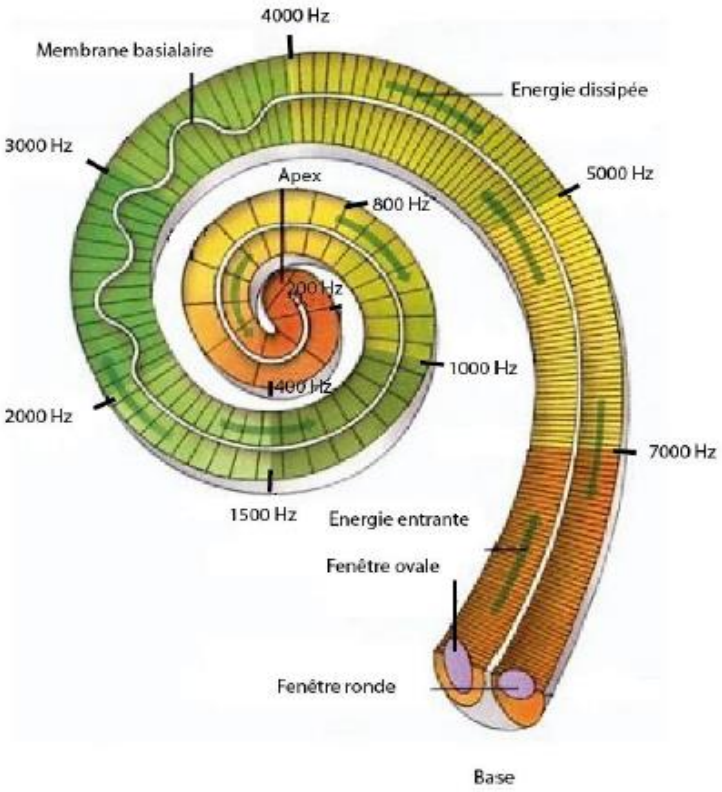

Figure 3.7 Carte tonotopique de la cochlée (Pinto & Sato, 2016).

B. Les filtres auditifs

Depuis de très nombreuses années, plusieurs travaux se sont intéressés à la modélisation du fonctionnement de la cochlée, en proposant un ensemble de filtres auditifs qui tentent de produire un modèle de perception qui s'apparente au système auditif humain (Lyon, Katsiamis, & Drakakis, 2010).

Le rôle d'un filtre auditif, en particulier un filtre passe-bande, est d'accentuer une bande de fréquences donnée tout en atténuant les fréquences au-delà de la bande (Gelfand, 2016). De ce fait la manière la plus simple pour modéliser les variations mécaniques de la membrane basilaire, qui ont pour effet des réponses variables en fonction de la plage de fréquences, est d'effectuer une analyse utilisant un banc de filtres, le modèle le plus simple et le plus réaliste est celui du banc de filtres Gammatones (R. Patterson, Nimmo-Smith, Holdsworth, & Rice, 1987) dont la réponse impulsionnelle s'approche des résultats observés en physiologie.

C. Les bandes critiques

Les bandes passantes du banc de filtres, supposés être idéaux, doivent être disjointes mais leur union doit couvrir la bande utile du signal. La largeur de ces bandes peut être uniforme, ou

non uniforme, elle suit dans ce cas, les caractéristiques des bandes critiques. Une bande critique désigne une plage de fréquences perçues indifféremment par le système auditif humain. La Figure 3.8 illustre le principe de la bande critique ou F_c est la fréquence centrale de la bande. F_1 et F_2 sont les fréquences de coupure inférieure et supérieure respectivement. Les deux fréquences, qui sont choisies de sorte que la différence entre leurs amplitudes et l'amplitude maximale soit inférieure à 3dB, désignent la largeur de la bande critique.

La succession des bandes critiques donne l'échelle fréquentielle qui a pour unité le *Bark*. Par définition, l'unité *Bark* est telle que chaque filtre auditif ait une largeur de 1 Bark. La plage fréquentielle audible est divisée en 24 bandes critiques réparties sur une échelle allant de 0 à 24 Barks (Zwicker & Terhardt, 1980). La largeur des bandes critiques augmente avec la fréquence (Voir le tableau 3.1).

Table 3.1. Normalisation de l'échelle des Barks et découpage en 25 bandes critiques.

Fréquence centrale	Bande Critique	N° de la bande
50	~100	1
150	100	2
250	100	3
350	100	4
450	110	5
570	120	6
700	140	7
840	150	8
1000	160	9
1175	190	10
1370	210	11
1600	240	12
1850	280	13
2150	320	14
2500	380	15
2900	450	16
3400	550	17
4000	700	18
4800	900	19
5800	1100	20
7000	1300	21
8500	1800	22
10500	2500	23
13500	3500	24
19500	-	25

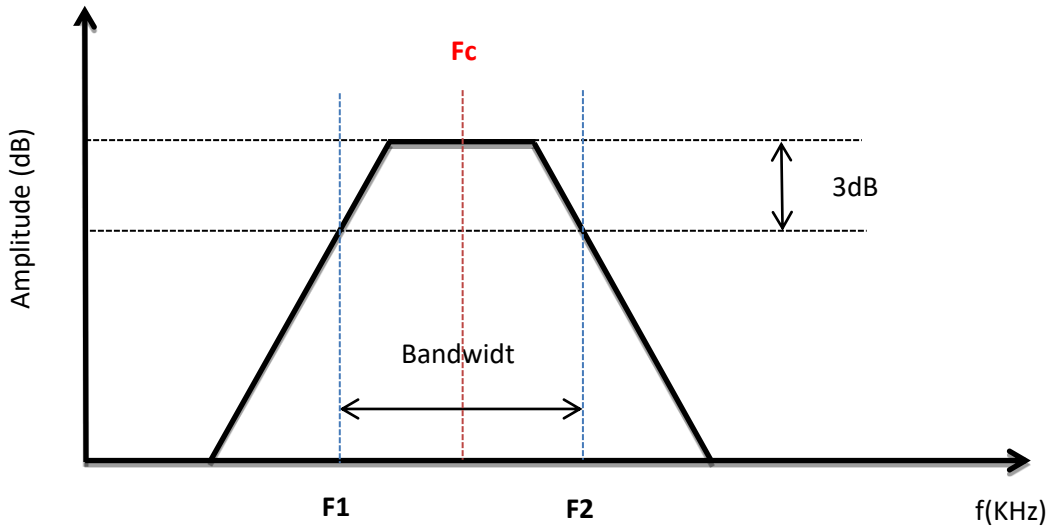

Figure 3.8 Exemple de bande critique.

La formule qui permet alors de convertir une fréquence f en Hertz en sa valeur en Bark est la suivante (Zwicker & Terhardt, 1980):

$$B(f) = 13 * arctan\left(\frac{0.76f}{1000}\right) + 3.5 * arctan\left(\left(\frac{f}{7500}\right)^2\right) Bark \qquad (3.14)$$

D. Echelle ERB

Plus récemment, une autre modélisation de la largeur de bande des filtres auditifs a été proposée par Moore et Glasberg dans (Glasberg & Moore, 1990) appelée largeur de bande rectangulaire équivalente (Equivalent Rectangular Bandwith, ERB). Dans leurs travaux, Moore et Glasberg ont mesuré des largeurs de filtre plus fines que celles obtenues dans (Zwicker & Terhardt, 1980) en utilisant l'échelle de Bark. Pour chaque filtre auditif, la valeur d'ERB est définie comme la largeur d'un filtre passe-bande idéal de même fréquence centrale. La Figure 3.9 montre une comparaison entre les largeurs des filtres décrits en bande critique et en ERB. La mesure des ERB illustre donc une meilleure résolution fréquentielle du système auditif que celle indiquée par les valeurs de bandes critiques.

Plusieurs approximations de l'échelle ERB existent (Smith & Abel, 1999), l'équation (3.15) est l'une d'elles,

$$ERB(f_c) = 1.019 * 24.7 * (4.37 * f_c/1000 + 1) \qquad (3.15)$$

Les valeurs de bandes critiques de la Figure 3.9 sont calculées à partir de cette formule en comparaison avec les valeurs obtenues à partir de l'équation (3.14).

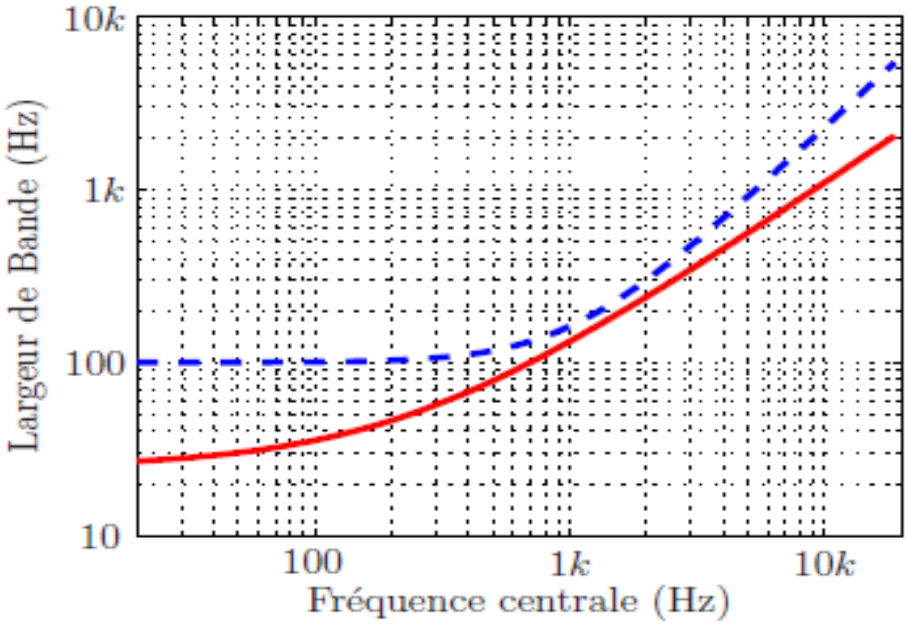

Figure 3.9 Comparaison entre la largeur des bandes critiques en Bark (en pointillé) et en ERB (en trait plein) (Fillon, 2004).

2.1.2.2. Les filtres Gammatones

Les filtres Gammatone (R. Patterson et al., 1987; R. D. Patterson, Allerhand, & Giguere, 1995) sont considérés comme l'une des modélisations du fonctionnement de la membrane basilaire les plus réputées (Bonastre et al., 2003).

Un filtre de Gammatone est un filtre linéaire décrit par une réponse impulsionnelle qui est le produit d'une distribution gamma et d'une tonalité sinusoïdale. Plusieurs implémentations de ces filtres sont connues, toutes proviennent de diverses approximations du filtre Gammatone original.

Un filtre Gammatone est défini dans le domaine temporel par la partie réelle de la fonction $g(t)$ (Figure 3.10) :

$$g(t) = \begin{cases} at^{n-1}\exp(-2\pi b\, ERB(f_c)t)\cos(2\pi f_c t + \varphi)\,, & t > 0 \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\,, & t < 0 \end{cases} \tag{3.16}$$

Avec a paramètre de normalisation d'amplitude, f_c la fréquence centrale de filtre, n l'ordre du filtre qui est généralement pris égal à 4, φ la phase initiale, $ERB(f_c)$ un paramètre définissant l'enveloppe du filtre.

La Figure 3.11 donne la représentation fréquentielle d'un banc de filtres Gammatone à 20 bandes.

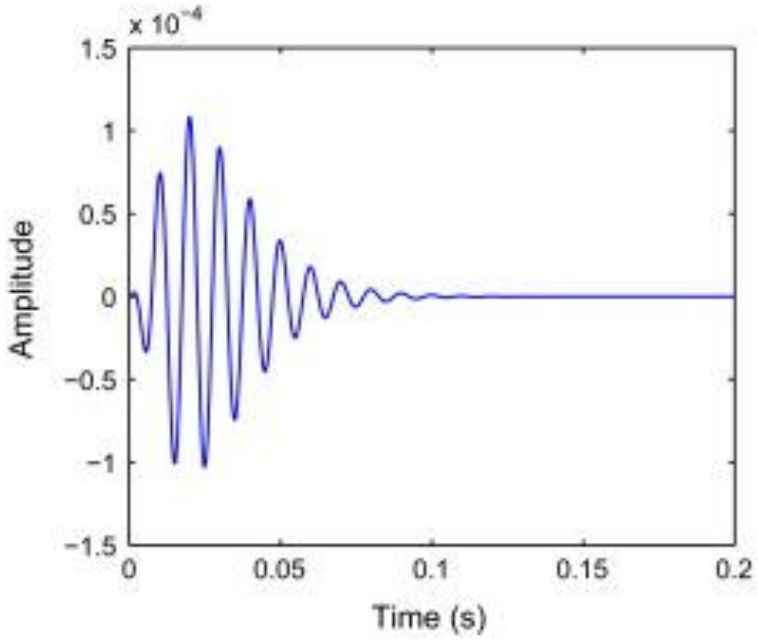

Figure 3.10 Réponse impulsionnelle d'un filtre Gammatone (Venkitaraman, Adiga, & Seelamantula, 2014) .

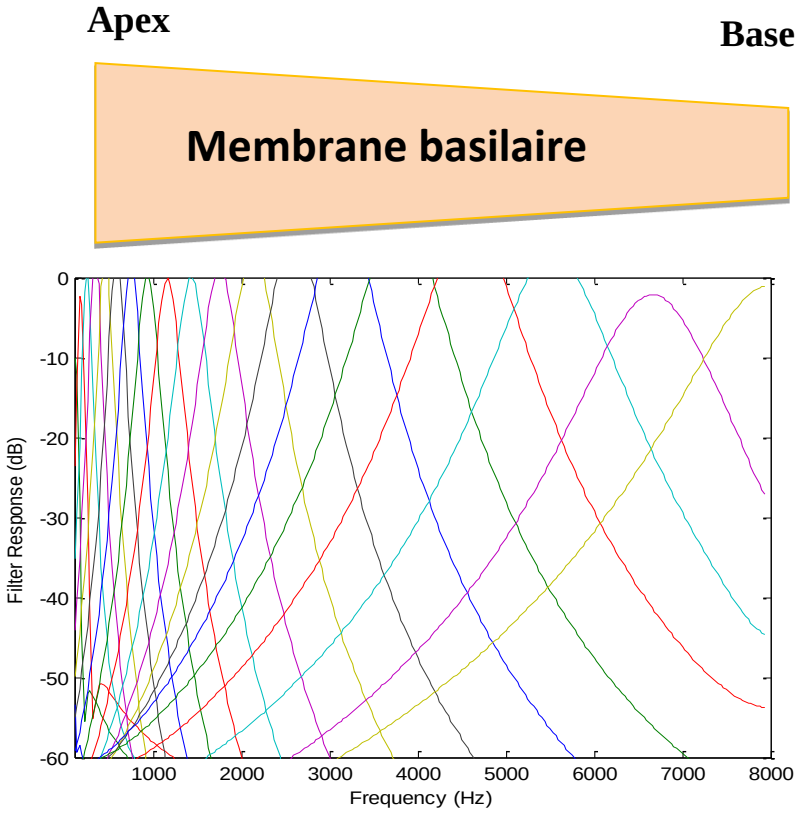

Figure 3.11 Réponse en fréquence d'un banc de filtre Gammatone 20 bandes.

2.1.3. Etape 3 : Application de logarithme

- Le problème de la reconnaissance est la convolution de la source et du conduit vocal, afin d'effectuer une déconvolution pour éliminer l'effet de la source, l'analyse cepstrale est effectuée en appliquant le logarithme dans le domaine fréquentiel pour transformer la multiplication en une sommation.

2.1.4. Etape 4 : Application de la DCT

- La dernière étape, comme celle de l'extraction MFCC, est l'application de la transformée en cosinus discrète (DCT) sur le vecteur d'énergies spectrales logarithmique pour une éventuelle compression de l'information.

2.2. Les étapes d'extraction des Coefficients MGCC

Les coefficients MGCC (Multitaper Gammatone Cepstral Coefficient) sont calculés avec les quatre étapes suivantes:

1- Une première étape de fenêtrage consiste à utiliser les fenêtres de pondération ou Multitaper.

Le fenêtrage réduit le biais (la valeur prévue de la différence entre le spectre estimé et le spectre réel), mais augmente aussi la variance de l'estimation spectrale (Thomson, 1982). Une technique pour réduire cette variance est de substituer l'estimation du périodogramme fenêtré par une estimation spectrale basée sur une multitude de fenêtres connue sous le nom de l'estimation Multitaper du spectre (Thomson, 1982). Dans cette technique un ensemble de fenêtres orthogonales est appliqué au signal de parole de courte durée et les estimations spectrales résultantes sont mises en moyenne, ce qui réduit la variance spectrale. La méthode de Multitaper a été utilisée dans plusieurs domaines, on cite par exemple, l'amélioration de la parole (speech enhancement, (Hu & Loizou, 2004)), la reconnaissance du locuteur et de parole (Alam, Kenny, Ouellet, & O'Shaughnessy, 2011; Alam, Kinnunen, Kenny, Ouellet, & O'Shaughnessy, 2013; Kinnunen, Saeidi, Sandberg, & Hansson-Sandsten, 2010). Les performances obtenues avec cette méthode d'estimation spectrale dépassaient celles du périodogramme fenêtré.

L'estimateur du spectre Multitaper, qui utilise une séquence de M fenêtres orthogonales peut être exprimé comme suit:

$$\hat{S}_{MT}(m,k) = \frac{1}{M}\sum_{p=0}^{M-1}\lambda(p)\left|\sum_{j=0}^{N-1} w_p(j)s(m,j)e^{\frac{-2\pi jk}{N}}\right|^2 \qquad (3.17)$$

Où $s(m,j)$ est le signal du parole dans le domaine temporel, N est la longueur de la trame, M le nombre de périodogramme fenêtré (Taper en anglais), w_p est la p^{ième} périodogramme fenêtré de données (p= 1, 2, ...,. M) utilisée pour l'estimateur spectrale $\hat{S}_{MT}(.)$, et $\lambda(p)$ est le coefficient de pondération associé au p^{ième} périodogramme fenêtré. Les périodogrammes w_p sont choisis pour être orthogonales, i.e.,

$$\sum_j w_p(j)w_q(j) = \delta_{pq} = \begin{cases} 1 & p = q \\ 0 & ailleurs \end{cases} \qquad (3.18)$$

Le spectre Multitaper est donc estimé via une moyenne pondérée de M sous-spectres individuels. Le périodogramme fenêtré (appelé aussi simple Taper) peut être obtenu comme un cas particulier de l'équation. (3.17), quand $p = M = 1$ et $\lambda(p) = 1$:

$$\hat{S}_d(m,k) = \left|\sum_{j=0}^{N-1} w(j)s(m,j)e^{\frac{-i2\pi jk}{N}}\right|^2 \qquad (3.19)$$

Divers périodogrammes fenêtrés ont été proposés dans la littérature pour l'estimation du spectre, on cite par exemple : les périodogrammes fenêtrés Thomson, les périodogrammes fenêtrés Multipeak et les périodogrammes fenêtrés sinus. Ces derniers sont appliqués avec une pondération optimale pour l'analyse de cepstre, appelé estimateur sinusoïdal pondéré de cepstre (Sinusoidal Weighted Cepstrum Estimator, SWCE) dans (Hansson-Sandsten & Sandberg, 2009). Les Figures 3.12, 3.13 et 3.14 illustrent une famille de M (M= 6) périodogrammes fenêtrés de types Thomson, SWCE et Multipeak respectivement dans le domaine temporel et fréquentiel.

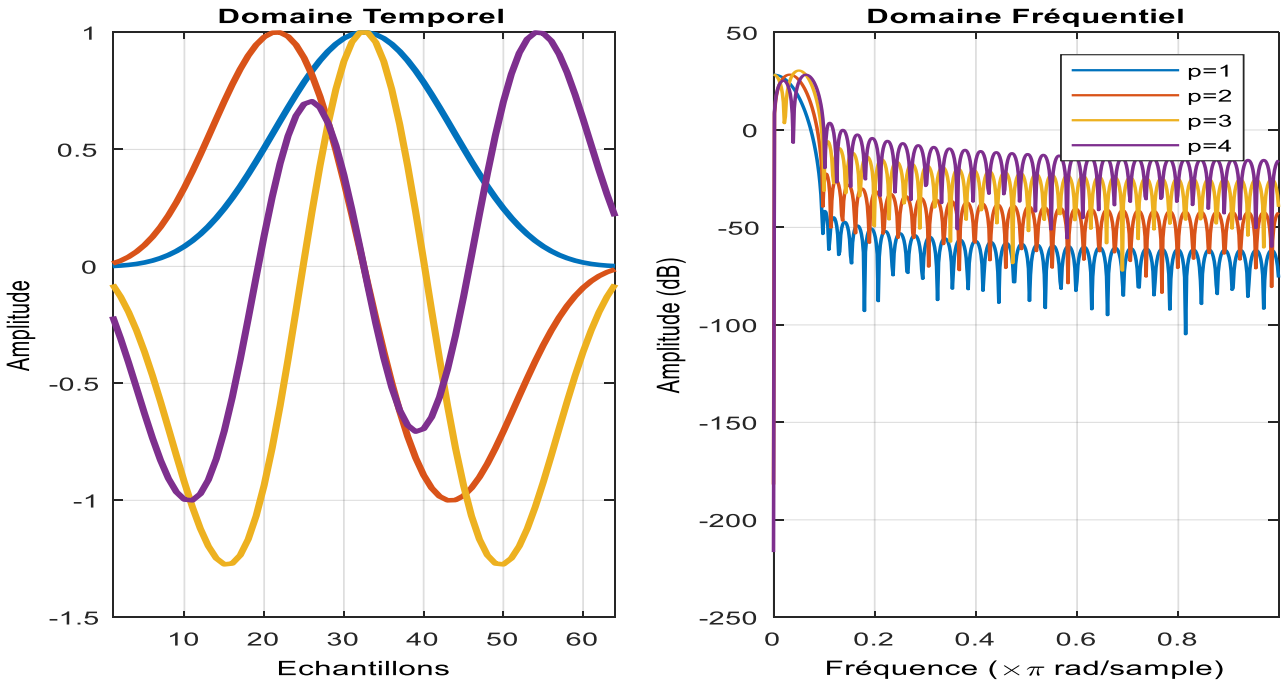

Figure 3.12 Thomson Multitapers pour N= 64, M=4 des domaines du temps et de fréquence.

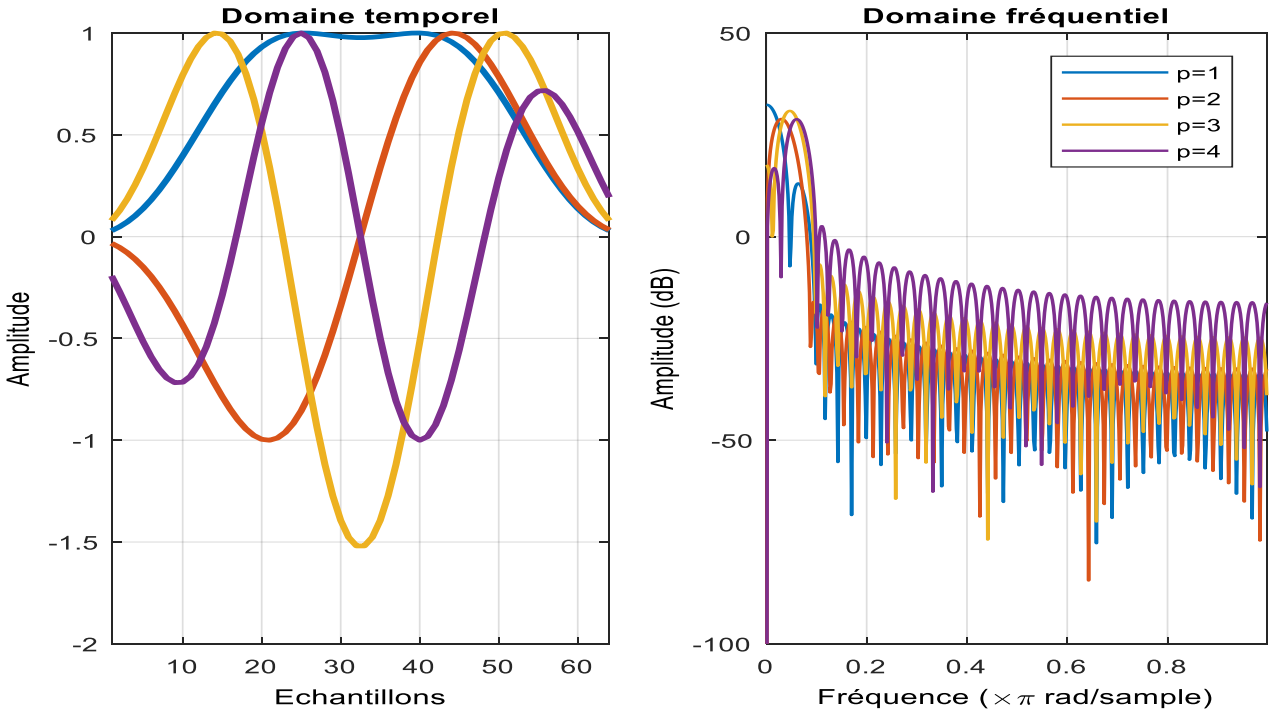

Figure 3.13 SWCE Multitapers pour N= 64, M=4 des domaines du temps et de fréquence.

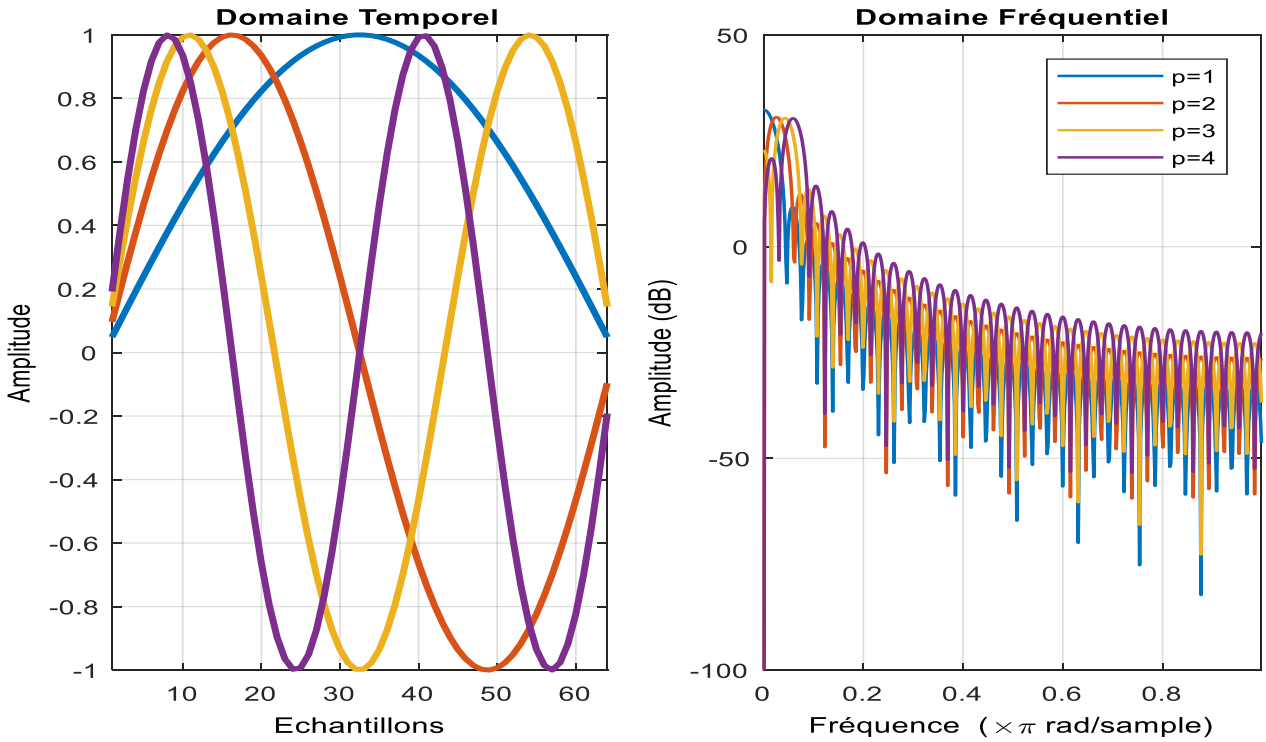

Figure 3.14 Multipeak Multitapers pour N= 64, M=4 des domaines du temps et de fréquence.

2- Une seconde étape consiste à effectuer une intégration fréquentielle en utilisant des filtres de type Mel pour obtenir l'énergie issue de chaque banc de filtre $E(j)$.

3- en appliquant le logarithme dans le domaine fréquentiel pour transformer la multiplication en une sommation.

4- La dernière étape, comme celle de l'extraction MFCC, est l'application de la transformée en cosinus discrète (DCT).

3. Conclusion

Ce chapitre, nous avons présenté un état de l'art sur les paramètres acoustiques les plus utilisés, visant l'amélioration de la robustesse du système de reconnaissance automatique du locuteur dans un environnement réel. Nous avons par la suite proposé deux méthodes nouvelles d'extraction de paramètres acoustiques fondées sur le système auditif humain en utilisant les filtres Gammatones à savoir le MGCC et le GPSCC. Le GPSCC exploite le module et la phase du spectre de fréquence, tandis que le MGCC utilise les fenêtres de pondération ou Multitaper.

Chapitre 4

Expérimentation I: Evaluation du système de vérification du locuteur par GPSCC et MGCC

Sommaire

Résumé

Ce chapitre donne les premiers résultats obtenus suite aux expérimentations réalisées sur la base de données TIMIT. Des expérimentations pour évaluer la pertinence de nouveaux paramètres proposés aux dégradations causées par la présence de bruit d'environnement d'une part et le transcodage de la parole dans les applications mobiles d'autre part.

1. Détails d'implémentation

Avant d'exposer les expériences et les résultats, il est indispensable de présenter tous les détails nécessaires afin de reproduire conformément ces travaux. Pour cela, nous allons suivre les étapes suivantes pour la construction et la validation du système :

- on utilisera tout d'abord une base de données (ou corpus) utile à l'apprentissage du système et ensuite à son évaluation,
- après une étape de séparation de parole/non parole sera nécessaire afin de ne conserver que les zones de paroles correspondantes aux locuteurs,
- on passera, par la suite, à l'étape de l'extraction des paramètres par différentes méthodes à savoir MFCC, MFPSCC, MGCC et GPSCC,
- ensuite on fera l'apprentissage des GMM et la création du modèle du monde UBM,
- finalement on passera à l'étape de décision et calcul de score.

1.1. Description de la base de données TIMIT

Les différentes configurations présentées ont été évaluées sur la base de données TIMIT. Une base de données spécialement conçue pour l'apprentissage et le test des systèmes de reconnaissance automatique de la parole (Garofolo et al., 1993) qui est maintenant largement utilisée dans les systèmes de reconnaissance du locuteur. TIMIT contient des enregistrements à large bande (16000 Hz) de 630 locuteurs, 70 % d'hommes et 30 % de femmes, provenant de huit régions dialectales majeures en Amérique.

Chaque locuteur est enregistré en lisant 10 phrases phonétiquement riches en anglais d'environs 3 secondes chacune. Chaque discours a été enregistré en une seule session, en utilisant un microphone de haute qualité dans une cabine insonorisée. Les 10 phrases prononcées par chaque locuteur sont comme suit :

- 2 phrases (SA) communes à tous les locuteurs de cette base, qui sont destinées à exposer les variantes dialectales des locuteurs.
- 5 phrases (SX) phonétiquement compactes.
- 3 phrases (SI) phonétiquement diverses.

Pour finir, le corpus TIMIT a été divisé en un répertoire d'apprentissage, composé de 530 locuteurs, et un répertoire de test, contenant 100 locuteurs.

1.2. Extraction des paramètres

Nous présentons en bref les différents algorithmes implémentés pour extraire les paramètres robustes pour la tâche de la vérification (Figure 4.1). Comme mentionné précédemment, quatre paramètres ont été considérés: MFCC comme notre référence, MFPSCC, MGCC et GPSCC. Chaque caractéristique a été extraite en utilisant une longueur de trame de 30 ms avec un pas de traitement de 10 ms. Après une pré-accentuation et fenêtrage de type *Hamming*, une analyse de banc de filtres auditifs a été appliquée sur les spectres obtenus. En particulier, dans le cas du MFCC et du MFPSCC, un ensemble de 24 filtres triangulaires à l'échelle de Mel a été utilisé, tandis que pour les MGCC et les GPSCC, un banc de 64 filtres Gammatones dont les fréquences centrales sont linéairement espacées sur l'échelle ERB a été appliqué. Finalement, les coefficients cepstraux sont obtenus à partir de la transformée en cosinus discrète (DCT) du logarithme des énergies issues des filtres. Les coefficients Cepstraux C_1 à C_{13} ont été retenus pour les MFCC, le MFPSCC et le GPSCC. Suite à l'étude précédente (Meriem, Farid, Messaoud, & Abderrahmene, 2017), 30 coefficients cepstraux sont conservés pour les MGCC.

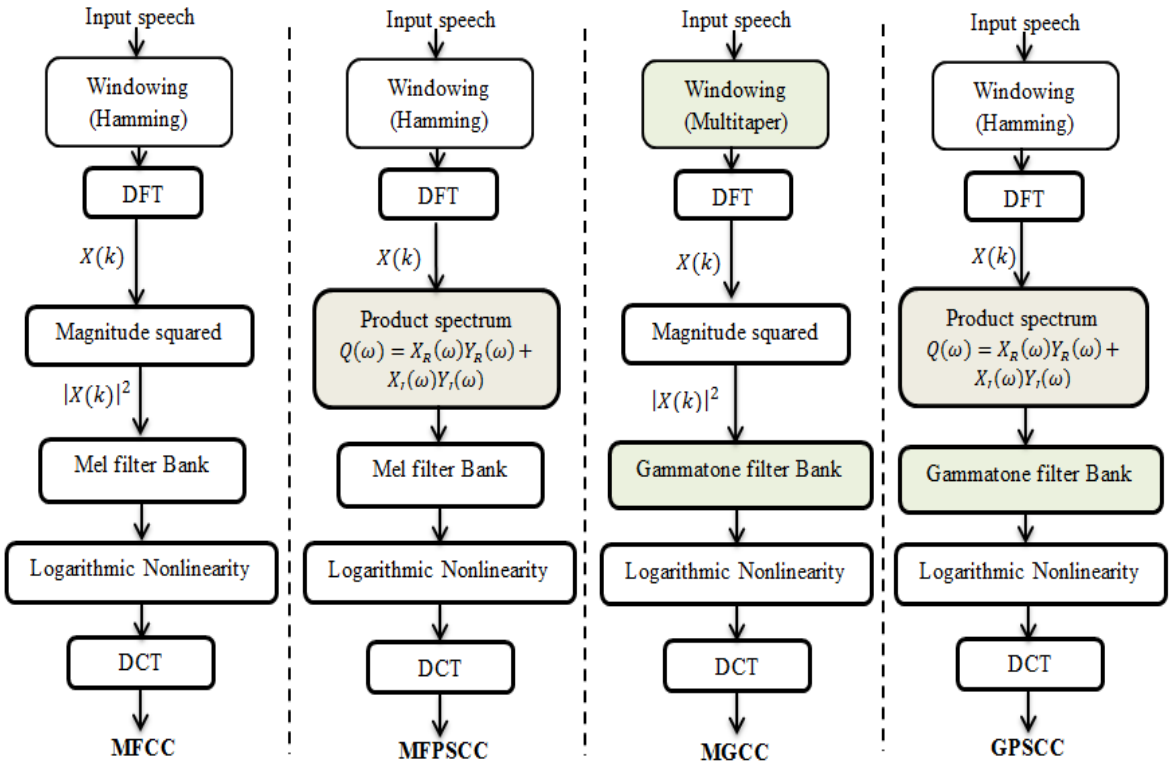

Figure 4.1 Les étapes d'extraction des paramètres MFCC, MFPSCC, MGCC et PSGCC.

1.3. Modélisation GMM-UBM et évaluation de performance

L'utilisation du paradigme de modélisation GMM-UBM dans cette étude est motivée par ses performances supérieures comparées à la technique de modélisation basée sur l'analyse factorielle JFA et I-vecteur. Ceci est dû principalement à l'absence de variabilité de multisession dans la base de données TIMIT et à la courte durée des énoncés d'entraînement et de test. Nous avons utilisé les mêmes hyper-paramètres comme dans la boîte à outils MSR Identity (Sadjadi, Slaney, & Heck, 2013), nous avons entrainé un modèle du monde dépendant du genre, qui contient C = 256 gaussiennes. Ce modèle est estimé à partir des données d'apprentissage de la base de données TIMIT représentées par des vecteurs acoustiques de 5300 phrases (10 énoncés de 530 locuteurs).

Toutes les expériences de vérification des locuteurs sont menées en utilisant les 100 locuteurs restants. Un modèle d'enrôlement est créé par la concaténation de leurs neuf phrases. La dixième phrase (1 énoncé) est conservée pour les tests. Les essais de vérification sont constitués de toutes les combinaisons modèle-test possibles, résultant en un total de 10.000 essais (100 essais client et 9900 essais imposteur).

Les Taux d'Erreur Egal (EER), la Fonction de Coût de Détection Minimum (*Min*DCF) couramment utilisée par des évaluations de reconnaissance des locuteurs de NIST, sont adoptés pour évaluer la performance de notre système de vérification du locuteur (Voir chapitre 1 section 5.5). De plus nous avons tracé la courbe DET (Detection Error Tradeoff) qui décrit le taux de fausses acceptations en fonction du taux de faux rejets.

2. Résultats et discussion

2.1. Performances du système de vérification dans un environnement bruité

Dans cette section, nous procédons à la comparaison de nos paramètres avec celles de l'état de l'art à savoir les MFCC. L'apprentissage du système de reconnaissance est fait dans des conditions non bruitées. Pour mieux évaluer nos paramètres acoustiques proposés plusieurs tests ont été faits lors de la phase d'évaluation du système et dans des conditions environnementales différentes de celle de l'apprentissage. Les signaux de parole obtenus sont corrompus par les bruits extraits du monde réel du corpus Noisex-92 développé (Varga & Steeneken, 1993). Trois

types de bruits ont été sélectionnés: bruit de bavardage (Babble), bruit dans une usine de production de véhicule que nous appelons bruit d'usine (Factory), et bruit blanc.

Il est à souligner que dans cette première partie du travail les trames de parole actives sont déterminées en utilisant l'auto-adaptatif VAD (VQVAD) (Kinnunen & Rajan, 2013).

2.1.1. Influence de type de la Multitaper sur les performances du système VAL

Dans ce paragraphe, nous allons exposé les résultats obtenus par (Meriem et al., 2014) en étudiant l'influence des différents types de fenêtres de pondération (Multitaper) sur les performances du système de vérification automatique du locuteur. D'après les résultats obtenus dans la Figure 4.2, montrant les courbes DET de différentes caractéristiques à savoir MFCC, GFCC[3] (Gammatone Frequency Cepstral Coefficients) et MGCC dans des conditions de test propres, on peut remarquer la supériorité des paramètres MFCC au point de fonctionnement EER. Cependant, les paramètres MGCC proposées surpassent les MFCC dans les régions à faible fausse alarme (MinDCF).

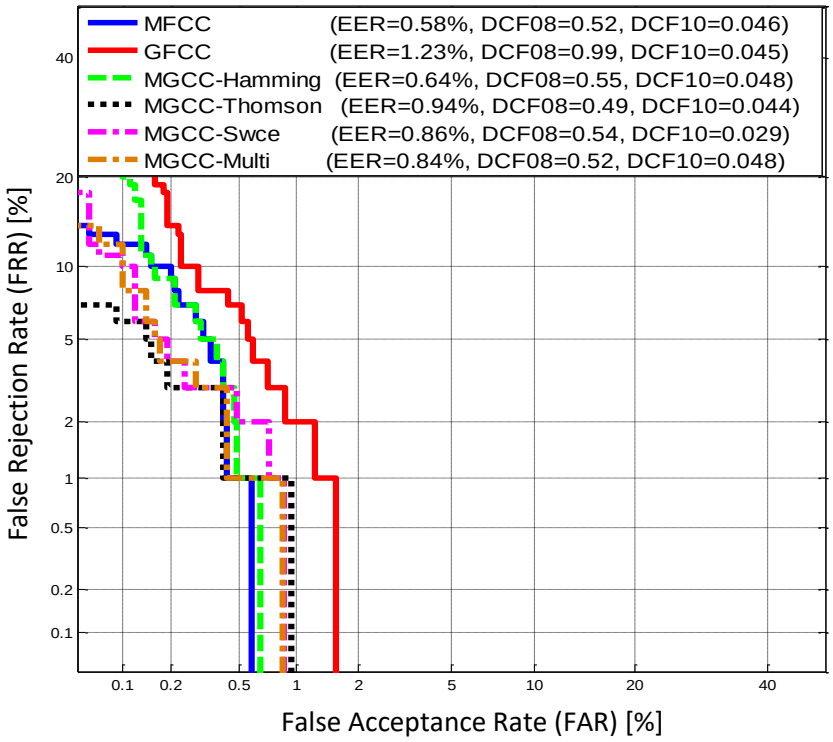

Figure 4.2 Les performances, en courbes DET, du système de vérification en utilisant les MFCC, GFCC, et MGCC dans des conditions propres.

[3] http://www.cse.ohio-state.edu/pnl/shareware/gfcc/

De plus, et en observant les résultats présentés dans les Figures 4.3 et 4.4, nous constatons que les résultats obtenus via les MGCC sont meilleurs que ceux obtenus via MFCC et GFCC dans un milieu bruité.

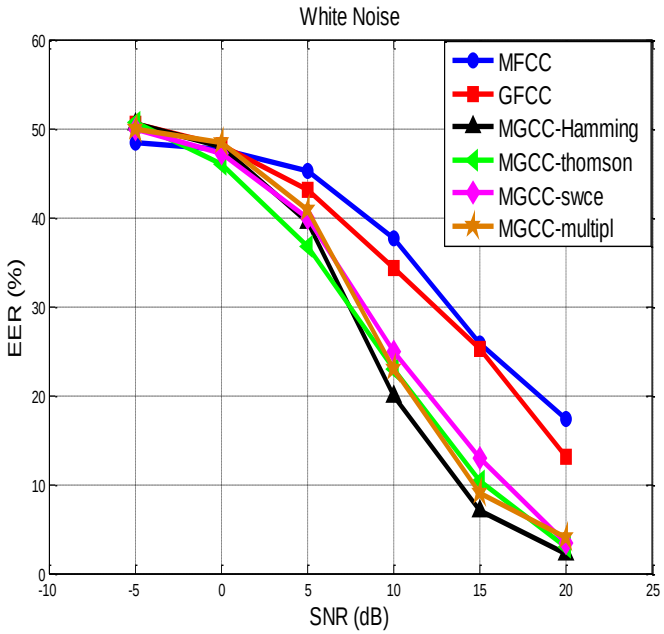

Figure 4.3 Les performances en EER (%) du système de vérification en utilisant les MFCC, GFCC, et MGCC dans un milieu bruité par un bruit Blanc pour différentes valeurs du RSB.

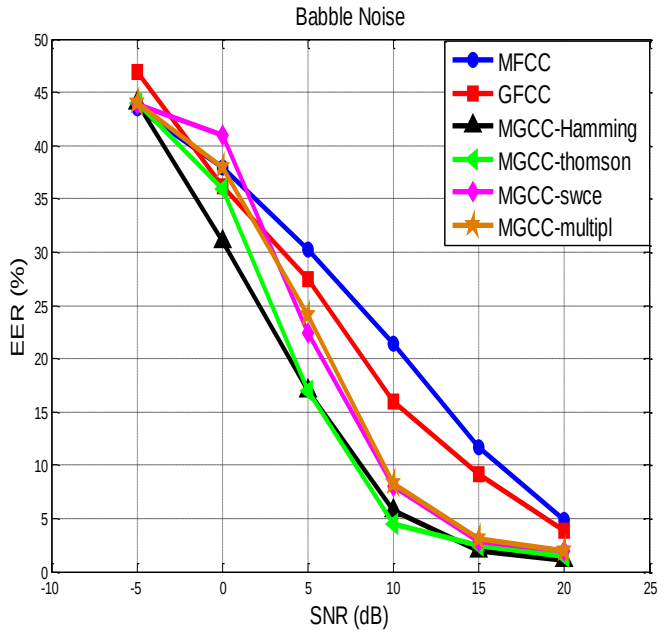

Figure 4.4 Les performances en EER (%) du système de vérification en utilisant les MFCC, GFCC, et MGCC dans un milieu bruité par un bruit de bavardage pour différentes valeurs du RSB.

2.1.2. Etude de la robustesse des paramètres proposés GPSCC et MGCC

Dans cette section, nous présenterons des résultats expérimentaux de la vérification du locuteur en utilisant les paramètres MFCC, MFPSCC, MGCC ainsi GPSCC. À partir de la Figure 4.5, montrant les courbes DET de différentes caractéristiques dans des conditions propres, le taux d'égale erreur EER est minimal avec une extraction à base des MFCC et MFPSCC par rapport à sa valeur avec celles des MGCC et GPSCC, cette différence est de l'ordre de 2%.

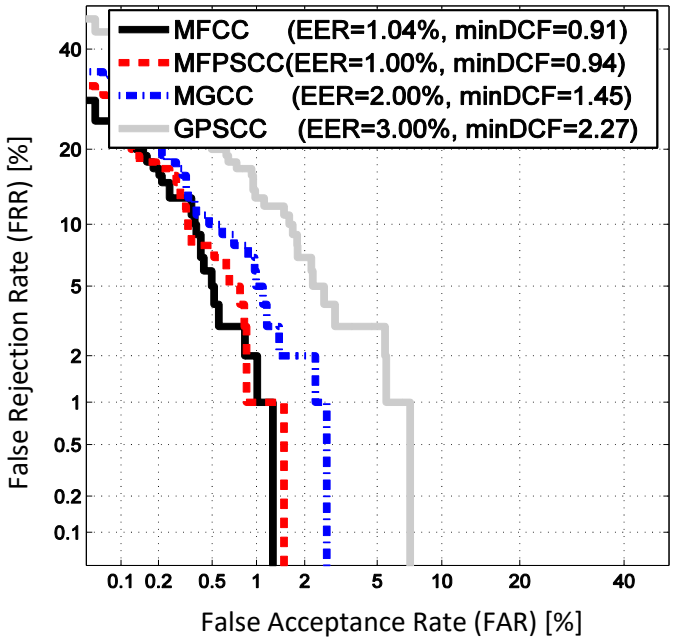

Figure 4.5 Les performances, en courbes DET, du système de vérification en utilisant les MFCC, MFPSCC, MGCC et GPSCC dans des conditions propres.

Table 4.1. Les performances en EER (%) et MinDCF(×100) du système de vérification en utilisant les MFCC, MFPSCC, MGCC et GPSCC dans un milieu bruité par un bruit de bavardage pour différentes valeurs du RSB.

RSB	Evaluation (%)	MFCC	MFPSCC	MGCC	GPSCC
5dB	EER	22.13	16.61	**7.85**	9.58
	MinDCF(×100)	7.92	5.22	**3.97**	5.79
10dB	EER	12.93	9.14	5.60	**4.70**
	MinDCF(×100)	4.00	3.23	**3.03**	3.28
15dB	EER	7.12	5.12	4.55	**3.65**
	MinDCF(×100)	2.64	**2.26**	2.56	2.66
20dB	EER	3.26	**2.30**	3.83	3.28
	MinDCF(×100)	1.58	**1.48**	2.29	2.38

Table 4.2. Les performances en EER (%) et MinDCF(×100) du système de vérification en utilisant les MFCC, MFPSCC, MGCC et GPSCC dans un milieu bruité par un bruit Blanc pour différentes valeurs du RSB.

RSB	Evaluation(%)	MFCC	MFPSCC	MGCC	GPSCC
5dB	EER	42.98	43.91	**32.37**	33.94
	MinDCF(×100)	9.95	9.95	**8.27**	9.62
10dB	EER	37.07	36.90	22.73	**21.55**
	MinDCF(×100)	9.94	9.94	**7.25**	9.08
15dB	EER	27.26	27.21	16.81	**10.66**
	MinDCF(×100)	9.61	9.68	**5.93**	6.77
20dB	EER	19.03	18.62	13.32	**5.38**
	MinDCF(×100)	8.94	8.98	4.94	**3.49**

Table 4.3. Les performances en EER (%) et MinDCF(×100) du système de vérification en utilisant les MFCC, MFPSCC, MGCC et GPSCC dans un milieu bruité par un bruit d'usine pour différentes valeurs du RSB.

RSB	Evaluation(%)	MFCC	MFPSCC	MGCC	GPSCC
5dB	EER	33.58	31.08	16.32	**14.97**
	MinDCF(×100)	9.42	9.37	**6.00**	8.04
10dB	EER	25.09	21.38	10.84	**7.62**
	MinDCF(×100)	7.56	7.46	4.52	**4.49**
15dB	EER	12.94	10.61	8.18	**4.61**
	MinDCF(×100)	5.39	4.76	3.69	**2.85**
20dB	EER	6.40	5.01	6.68	**3.58**
	MinDCF(×100)	3.10	2.56	3.20	**2.45**

Cependant, Dans un milieu bruité, ce sont les faibles valeurs du EER obtenues avec une extraction des MGCC et GPSCC qui prennent le dessus avec une baisse considérable par rapport à celles obtenues avec l'extraction des MFCC tel que décrits dans les Tableaux 4.1, 4.2 et 4.3 (atteignant 14% pour un RSB = 5 dB dans le milieu bruité avec un bruit bavardage, 10% pour le bruit blanc avec un RSB = 5 dB et 16% pour le bruit d'usine avec un RSB = 5 dB dans le cas des MGCC).

Nous pouvons conclure, que dans un milieu propre ce sont les MFCC et MFPSCC à base de filtres Mel qui donnent de meilleurs résultats, cependant dans un environnement bruité, ce sont

les nouveaux paramètres proposés MGCC et GPSCC à base de filtres Gammatones qui prennent le dessus avec une amélioration remarquable des performances du système.

2.1.3. Interprétation des résultats obtenus

L'utilisation des bancs de filtres Gammatone à la place de filtre Mel rend considérablement les caractéristiques plus résistantes au bruit additif. Cette observation n'est pas une surprise, sachant que les bancs de filtres Gammatone exploite les avantages de système auditif humain (R. Patterson et al., 1987) qui semble plus résistant au bruit (l'oreille humaine est robuste dans les environnements bruités). Comparé au filtre Mel, le filtre Gammatone a une forme plus lisse; il est très similaire à la fonction exponentielle arrondie utilisée dans la représentation de la réponse en amplitude des filtres auditifs humains (Gold, Morgan, & Ellis, 2011). De plus, le chevauchement dans les bancs de filtres Mel est fixe, donc si le nombre de filtres augmente, la bande passante de chaque filtre triangulaire diminuera, pour le filtre Gammatone, la bande passante est déterminé par sa fréquence centrale, donc si le nombre de filtres augmente, le chevauchement augmente. Les résultats théoriques et expérimentaux dans (Dimitriadis, Maragos, & Potamianos, 2011) ont montré que la bande passante du filtre est l'un des facteurs les plus importants affectant la performance de la reconnaissance vocale. Afin de mettre en évidence la robustesse au bruit des filtres Gammatones, nous avons tracé trois Représentations Temps-Fréquence (TF) du signal de parole à savoir le *spectrogramme* qui est une visualisation temps-fréquence calculée en utilisant la transformée de Fourier à court terme (STFT), le *Spectrogramme-MFCC* utilisant un ensemble de filtres triangulaires Mel et le *Gammatogramme* qui utilise un banc de filtres Gammatones. La Figure 4.6 illustre ces représentations temps-fréquence d'un signal parole propre, tandis que la figure 4.7 montre l'impact du bruit blanc avec un RSB = 5 dB, sur eux. Lorsque nous regardons de près la Figure 4.6, nous constatons que les MFCC se rapprochent le plus au spectre du signal. La Figure 4.7, reflète la forte perturbation du spectre du signal lors de l'ajout du bruit, et son impact sur les deux spectres. Les MFCC avec l'utilisation de filtres Mel sont fortement influencés (apparition de hautes fréquences) par rapport aux filtres Gammatones qui sont moins bruités et plus fidèles à leur représentation dans le milieu propre. Nous pouvons prédire que dans un milieu bruité, l'utilisation des MGCC et PSGCC (à base de filtres Gammatone) caractérisent le mieux, un même locuteur.

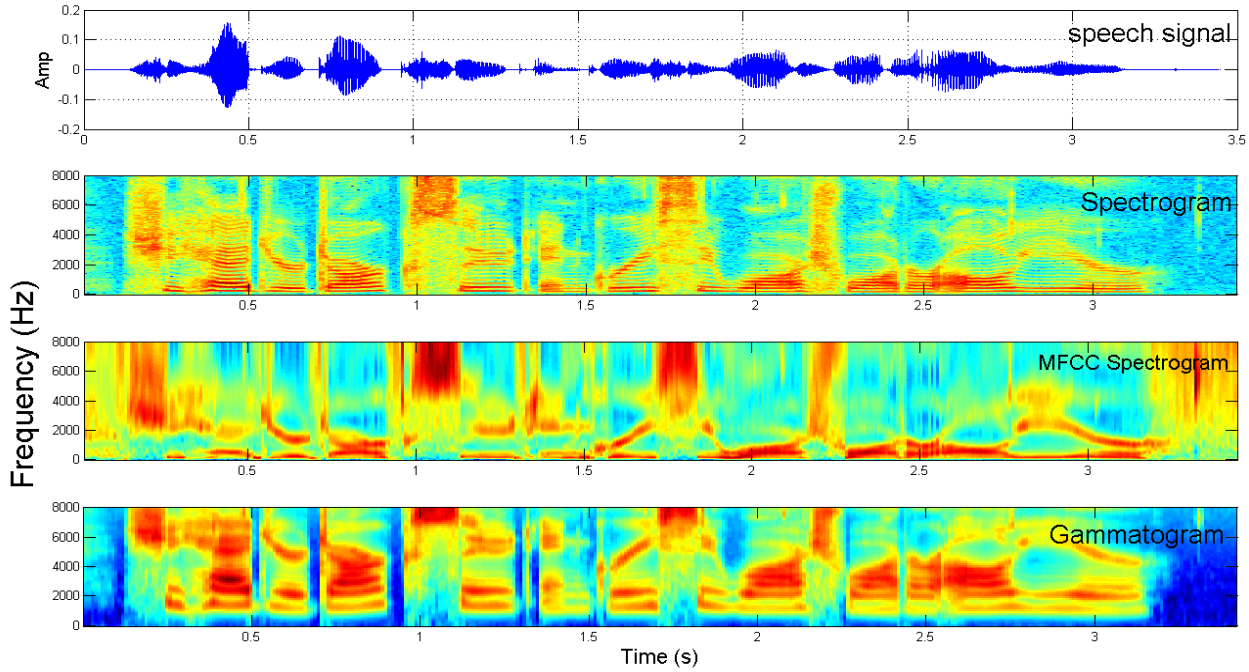

Figure 4.6 Premier panel : signal parole propre, deuxième panel: représentation du spectre; troisième panel: représentation du spectre des MFCC ; quatrième panel : représentation du spectre Gammatone.

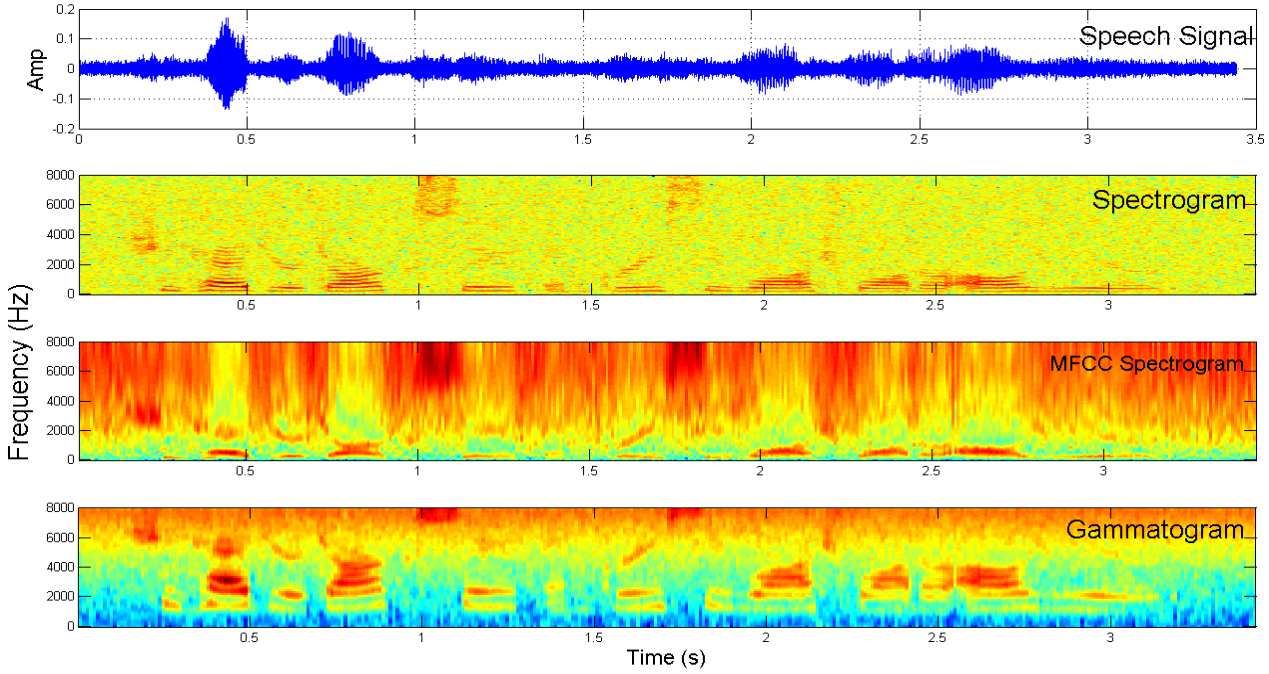

Figure 4.7 Premier panel: signal parole bruité avec un bruit blanc à un RSB= 5 dB, deuxième panel: représentation du son spectre; troisième panel: représentation du spectre des MFCC ; quatrième panel : représentation du spectre Gammatone.

Une explication possible du fait que les GPSCC et les MGCC résistent mieux au bruit blanc peut être trouvée lors du calcul de la corrélation linéaire entre le signal d'origine et le signal bruité. Une corrélation parfaite égale à 1 (unité) indique une correspondance parfaite. D'un autre côté, une corrélation négative ou nulle indique que les caractéristiques du signal propre et celle du signal bruité sont indépendantes. Pour illustrer cela, les diagrammes de dispersion du premier coefficient des quatre caractéristiques étudiées avant et après l'ajout du bruit blanc à 5 dB sont représentés sur la Figure 4.8. Il est clair que les GPSCC et MGCC ont les coefficients de corrélation les plus élevés donnant lieu à une distribution des caractéristiques plus stable et meilleure généralisation des performances dans des conditions bruitées. Par contre, l'effet de mismatch est plus prononcé dans MFCC et MFPSCC.

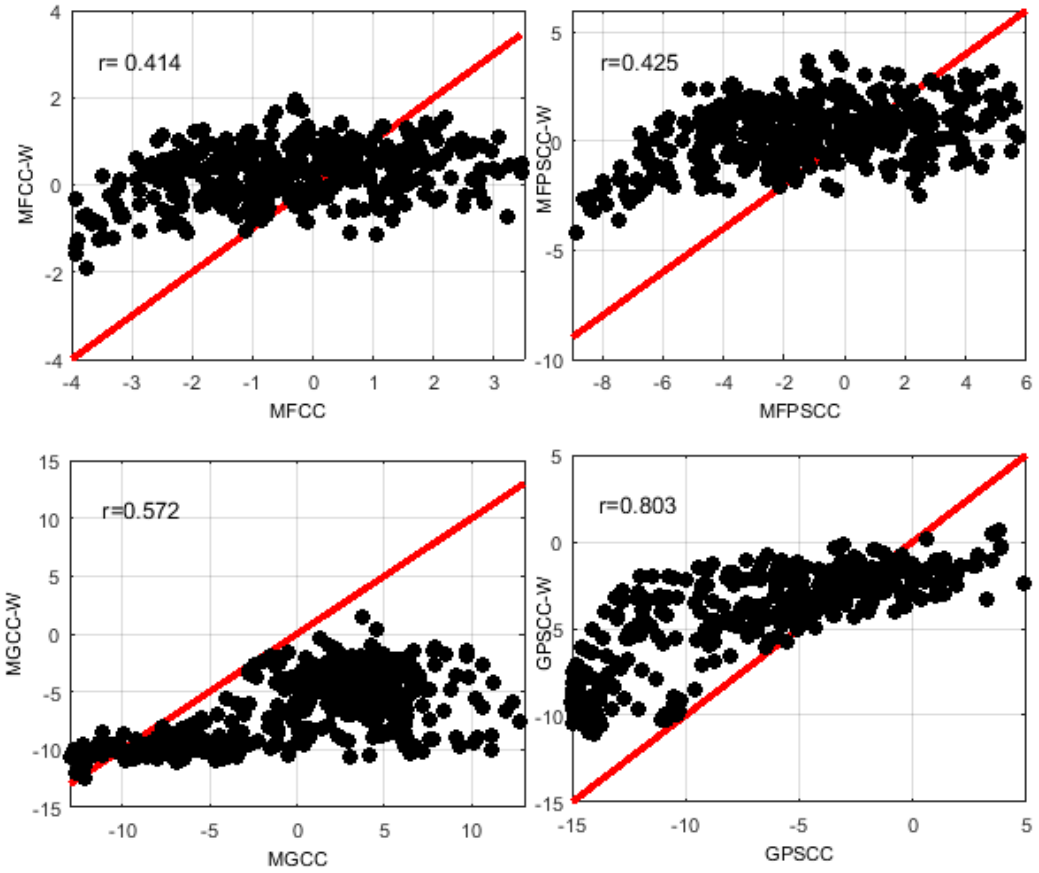

Figure 4.8 La corrélation linéaire et la dispersion du premier coefficient d'un signal parole propre en fonction du signal bruité avec un bruit blanc à 5 dB pour différentes caractéristiques (signal parole d'un homme MWSH0 SX436).

2.1.4. Fusion des scores de GPSCC et MGCC

Dans le but d'améliorer davantage nos résultats dans un environnement réel, nous avons utilisé un algorithme de fusion des scores de GPSCC et MGCC. L'objectif est de combiner les informations de spectre d'amplitude ainsi que le spectre de phase, où les deux caractéristiques sont dérivées en utilisant les bancs de filtres Gammatones. La fusion au niveau du score est mise en œuvre en utilisant une régression logistique (Brummer et al., 2007). De plus, la boîte à outils FoCal (Brümmer, 2005) est utilisée pour optimiser les poids de fusion. Les résultats illustrés dans le tableau (4.4), montrent l'apport de la fusion sur les performances de notre système, où toutes les valeurs de l'EER ont chuté pour les différentes valeurs du RSB.

Table 4.4. Performances en EER (%) du système de vérification en effectuant la fusion des scores des deux caractéristiques MGCC et GPSCC.

Bruit	Caractéristiques	5 dB	10 dB	15 dB	20 dB	clean
bavardage	GPSCC	9.58	4.70	3.65	3.28	3.00
	MGCC	7.85	5.60	4.55	3.83	2.00
	Fusion	**7.18**	**3.14**	**2.22**	**1.76**	**1.15**

2.2. Performances dans un environnement mobile et l'influence de la parole transcodée

Nous poursuivons cette étude en réalisant une série d'expériences qui vise à explorer l'impact de la parole transcodée sur les performances de système de vérification des locuteurs (Fedila & Amrouche, 2012)les signaux de parole de la base de données TIMIT sont passé à travers le codec G722.2 (AMR-WB) que nous avons simulé pour aboutir à la base de données transcodée. Le codeur AMR_WB (Adaptatif multi-débit à large bande) a été sélectionnée comme recommandation par UIT-T G.722.2. Il fonctionne sur une bande passante étendue allant de 50 Hz à 7 kHz et comporte 9 modes dont seuls 5 sont obligatoire dans les terminaux: 6,6 kbit/s, 8,85 kbit/s, 12,65 kbit/s, 15,85 kbit/s, 23,85 kbit/s**.** Sur les réseaux 2G, seuls les 3 débits inférieurs peuvent être employés. Sur les réseaux 3G, les 5 débits sont utilisables. Le codec AMR_WB est fondé sur le modèle de codage prédictif linéaire avec excitation par code (CELP, *code excited linear prediction*) et fonctionne sur des trames de 20 ms.

2.2.1. Mismatch condition

Pour évaluer les performances du notre système, nous avons comparé les différentes techniques de paramétrisation standards (MFCC et MFPSCC) avec les paramétrisations proposées (MGCC et PSGCC) dans des conditions incompatibles (apprentissage avec parole propre / test avec parole transcodé ou décodé). A partir de la Figure 4.9, on peut constater que la dégradation des paramètres MFCC apparaît nettement pour les bas débits (6,60 kbps et 8,85 kbps), alors que les coefficients MGCC et PSGCC résistent mieux à ces bas niveaux de débits. On peut voir clairement pour un débit de 6,60 kbps, les paramètres PSGCC et MGCC sont les plus robustes et les plus adaptés à cette environnement par rapport aux paramètres MFCC (une réduction de taux de 12% et 6% respectivement par rapport au MFCC).

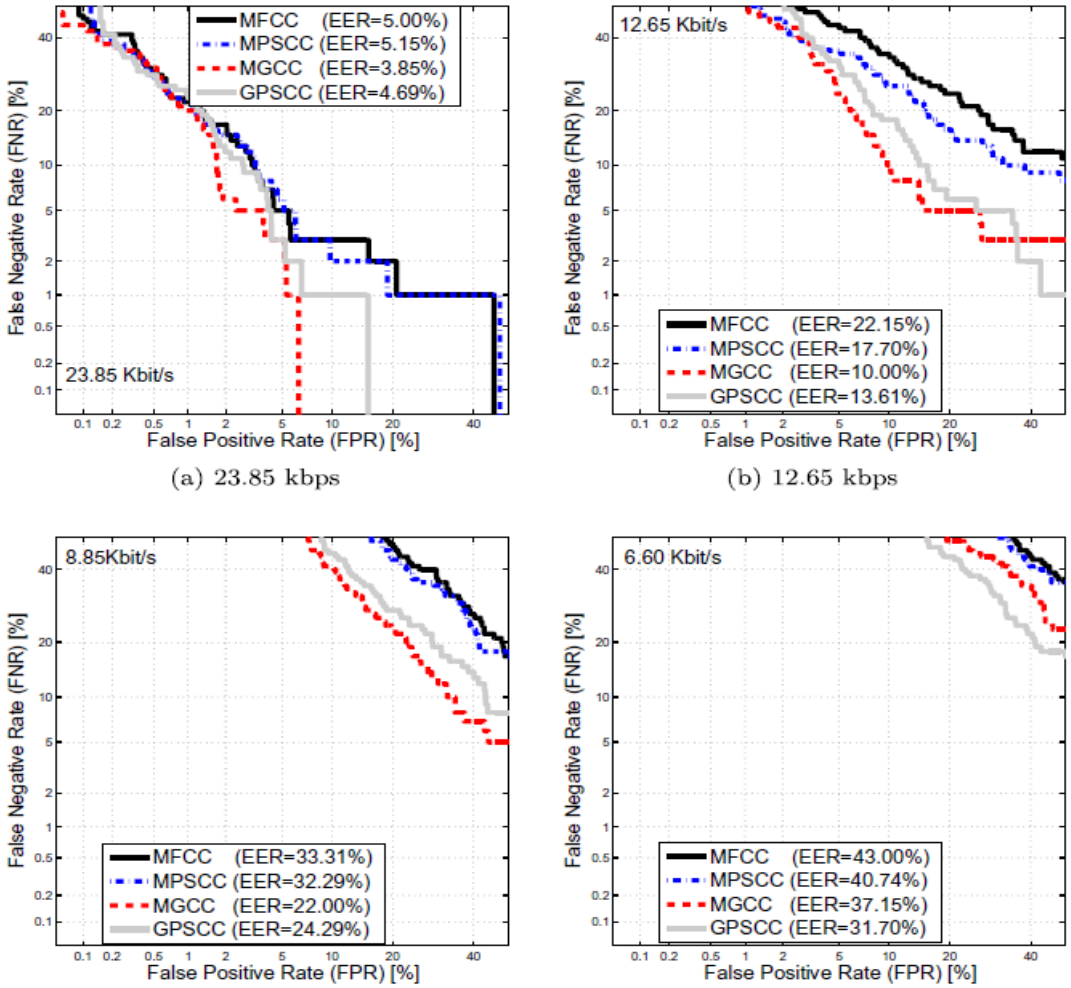

Figure 4.9 Les performances, en courbes DET, du système de vérification en utilisant les MFCC, MFPSCC, MGCC et GPSCC pour différentes valeurs de débits.

Une fusion au niveau du score du GPSCC et du MGCC est étudiée à différents débits. Nous remarquons à partir des résultats illustrés dans le tableau 4.5 l'efficacité de la fusion à bas débit.

Table 4.5. Performances en EER (%) du système de vérification en effectuant la fusion des scores des deux caractéristiques MGCC et GPSCC pour différentes valeurs de débits.

Features	6,60 kbps	8,85 kbps	12,65 kbps	15.85 kbps	23,85 kbps
GPSCC	31.70	24.29	13.61	**9.05**	4.69
MGCC	37.15	22.00	10.00	10.10	**3.85**
Fusion	**31.01**	**21.68**	9.68	9.60	4.15

2.2.2. L'impact de la parole transcodée sur la robustesse des paramètres proposés

En suivant la même méthodologie de la section précédente, nous avons étudié l'impact de la parole transcodée sur les performances du notre système en traçant les diagrammes de dispersion de la corrélation linéaire entre le signal d'origine et le signal transcodé du premier coefficient des quatre caractéristiques (Figures 4.10, 4.11, 4.12 et 4.13). Les Figures 4.12 et 4.13 montrent une distribution plus stable des paramètres MGCC et GPSCC. MFCC et MFPSCC ont montré des changements de données significatifs, en particulier pour le faible débit de 6,60 kbps, où le changement se manifeste par des coefficients de corrélation négatifs. Ces résultats révèlent clairement l'efficacité des paramètres MGCC et PSGCC et l'utilisation de filtres Gammatones.

De plus, nous avons calculé l'erreur relative (E) proposée dans (Ireland, Knuepffer, & McBride, 2015) pour quantifier la discordance associée à la compression:

$$E = \frac{C-\hat{C}}{C} \times 100 \qquad (4.1)$$

OùC est une caractéristique calculée à partir de données propres, alors que $\hat{C}$est la caractéristique calculée à partir de la parole codée-décodée en utilisant G722.2 à différents débits binaires. Ce critère a été utilisé par Irlande et al. (Ireland et al., 2015) pour évaluer l'efficacité d'un paramètre particulier dans le domaine compressé.

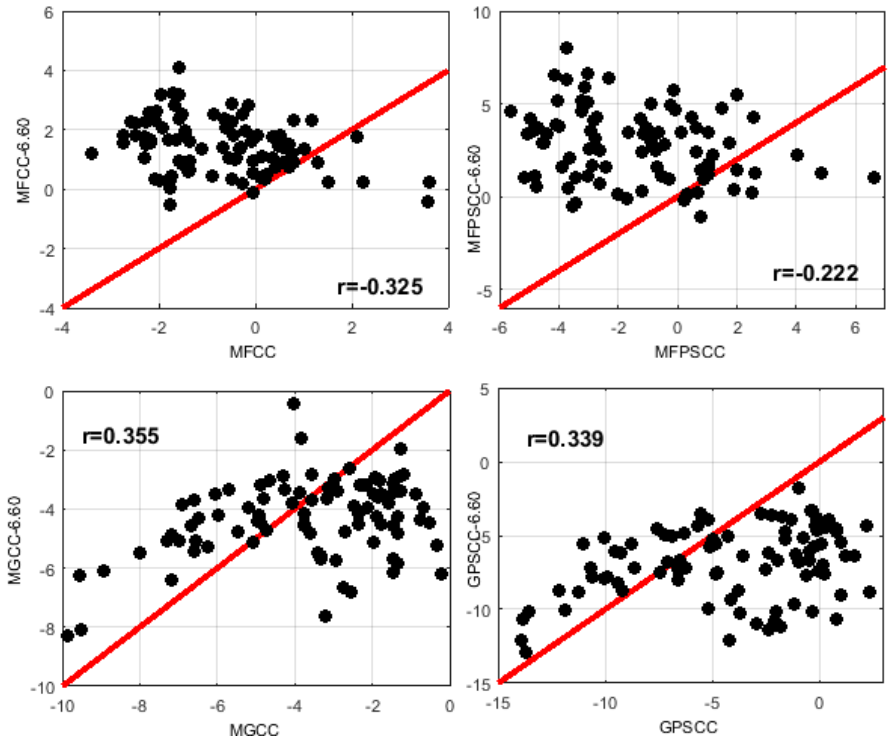

Figure 4.10 La dispersion du premier coefficient des différentes caractéristiques d'un signal de parole propre en fonction du signal décodé pour un débit de 6.60 kbps.

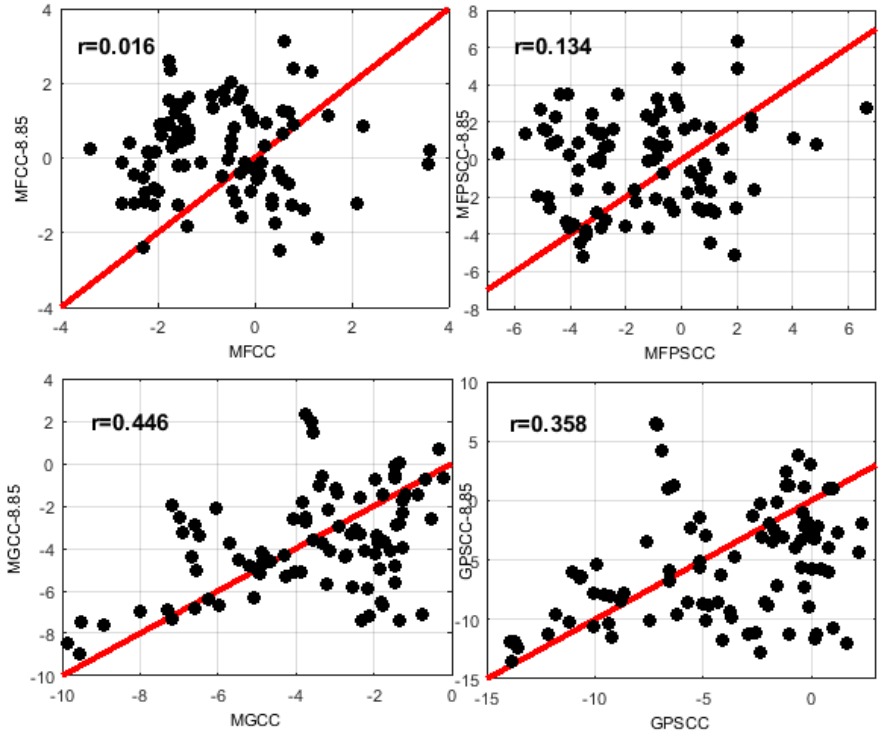

Figure 4.11 La dispersion du premier coefficient des différentes caractéristiques d'un signal de parole propre en fonction du signal décodé pour un débit de 8.85 kbps.

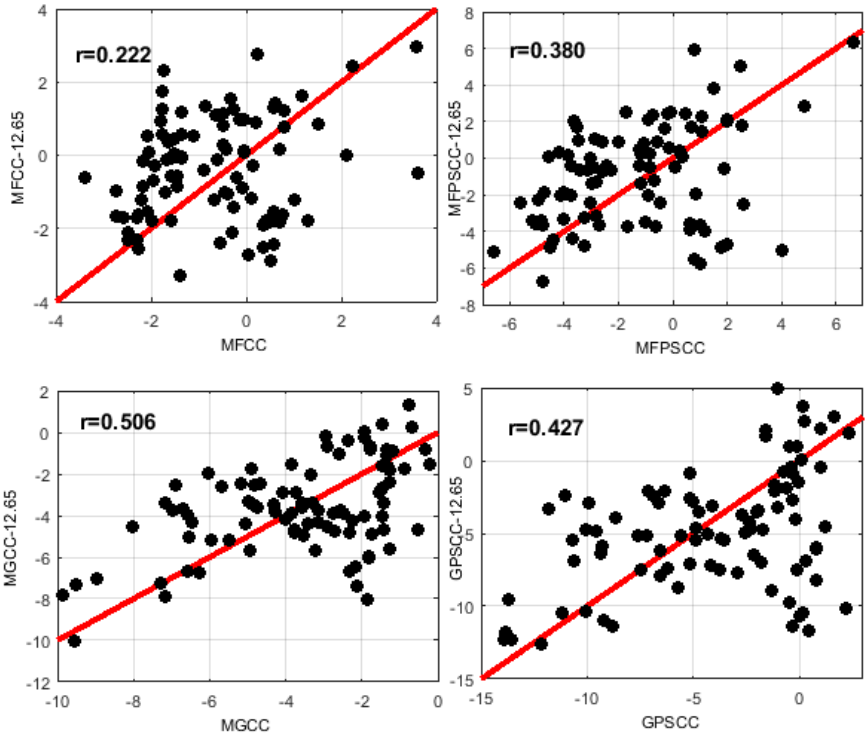

Figure 4.12 La dispersion du premier coefficient des différentes caractéristiques d'un signal de parole propre en fonction du signal décodé pour un débit de 12.65 kbps.

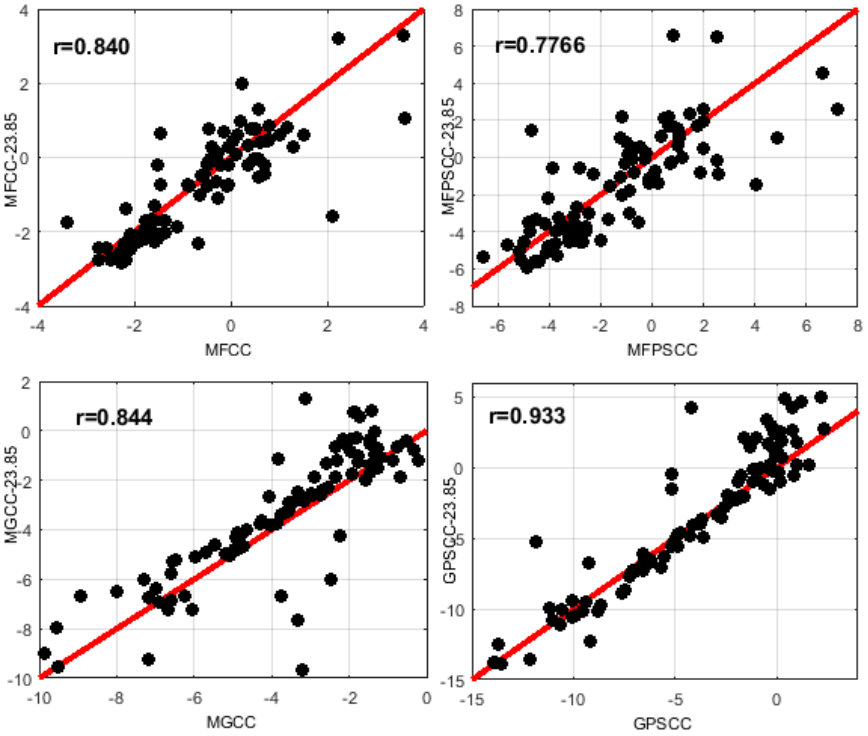

Figure 4.13 La dispersion du premier coefficient des différentes caractéristiques d'un signal de parole propre en fonction du signal décodé pour un débit de 23.85 kbps.

Les erreurs relatives de différents paramètres pour différents débits sont résumées sous la forme de boxplot (Figures 4.14 et 4.15). Le boxplot ou la boîte à moustaches (aussi appelée diagramme en boîte), est un outil statistique efficace, la boîte fermée représente le quartile inférieur, la médiane et le quartile supérieur tandis que les bras de la boîte (moustaches) renseigne sur la dispersion des valeurs situées au début ou à la fin de la série ordonnée. Il est clair qu'une propriété souhaitable de boxplot dans notre cas est principalement une médiane absolue près de zéro et des intervalles interquartiles inférieurs.

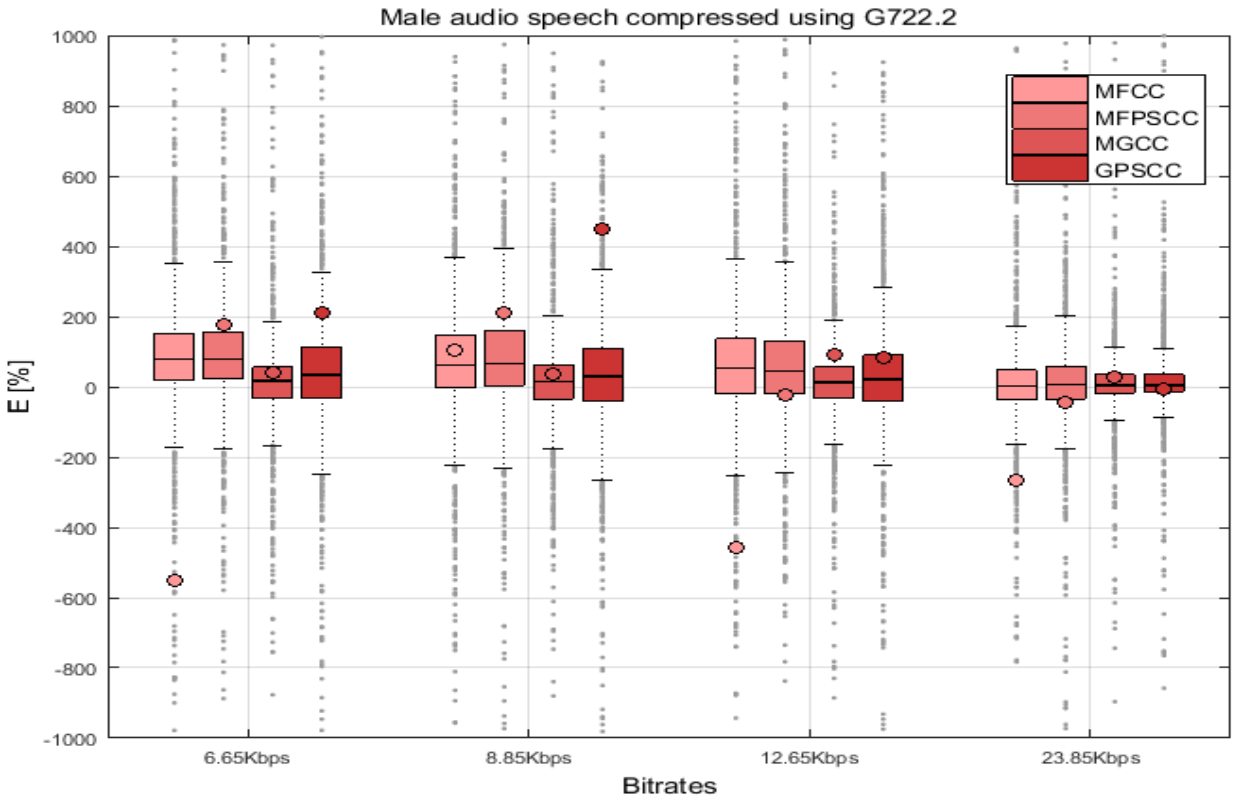

Figure 4.14 Représentations en boxplot de l'erreur relative en % pour les différentes caractéristiques utilisant un discours audio masculin compressé à différents débits.

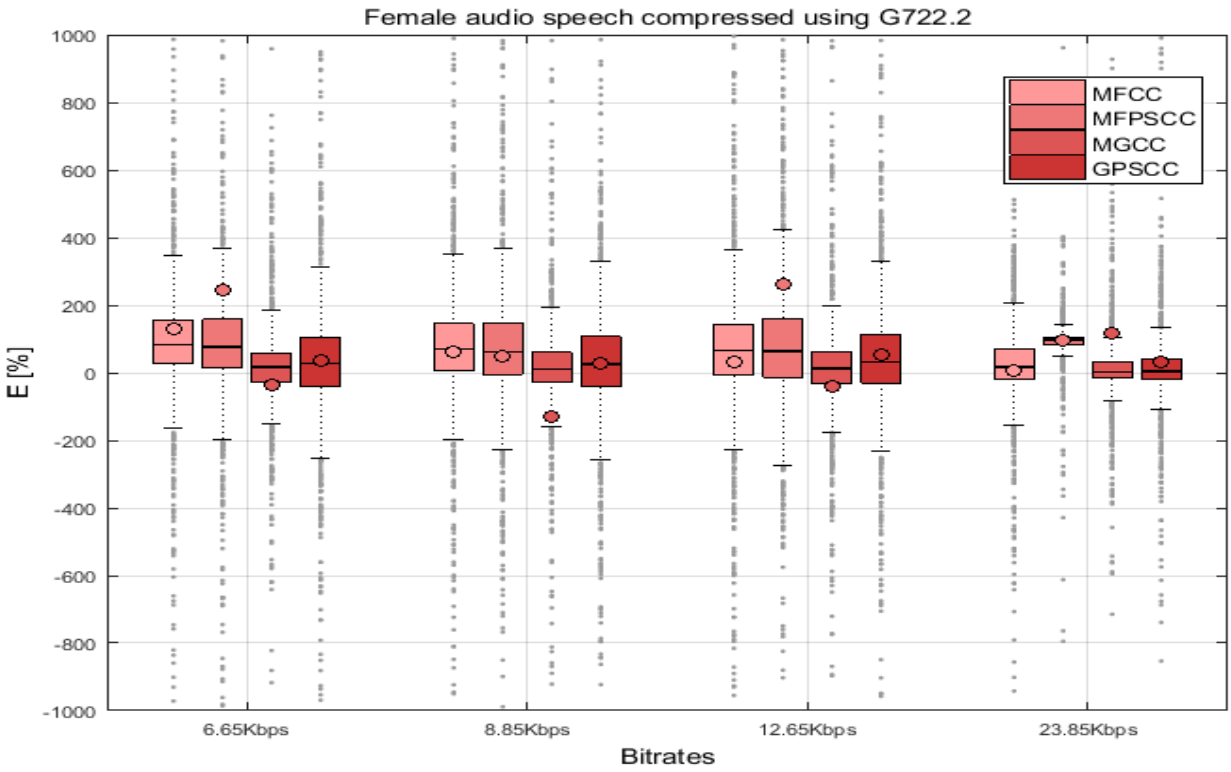

Figure 4.15 Représentations en boxplot de l'erreur relative en % pour les différentes caractéristiques utilisant un discours audio féminin compressé à différents débits.

L'analyse des boxplot représentée sur les Figures 4.14 et 4.15 approuve les résultats de EER de la Figure 4.9. Malgré que toutes les caractéristiques ont une valeur médiane proche de zéro pour 23,85 kbps (débit le plus élevé), les boxplots de MGCC et PSGCC sont caractérisés par une distance interquartile inférieure (dispersion inférieure). L'efficacité des deux paramètres MGCC et PSGCC devient plus prononcée à des débits inférieurs et cela se manifeste par des boxplot avec des médianes inférieures par rapport à MFCC et MFPSCC.

3. Conclusion

Dans ce travail de thèse, nous avons exploité la relation symbiotique entre la phase à court terme et le spectre d'amplitude et aussi les bancs de filtres Gammatones pour proposer de nouveaux paramètres. Dans ce chapitre, nous avons montré que les nouveaux paramètres proposés présentent une grande robustesse à des conditions de bruit sévères et de même pour un environnement mobile.

Chapitre 5

Expérimentation II: Nouvelle technique de détection d'activité vocale

Sommaire

Résumé

La détection d'activité vocale VAD utilisée pour détecter la présence de la parole dans un signal audio représente notre centre d'intérêt dans la seconde partie de cette thèse. Dans ce chapitre nous répertorions et étudions les méthodes de détection d'activité vocale performantes dans des environnements difficiles. Nous porterons plus d'intérêt aux méthodes récentes. Nous retrouvons par la suite le détail de la nouvelle technique de détection basée sur l'utilisation des Support vecteur machine SVM.

1. Etat de l'art sur la détection d'activité vocale

La détection d'activité vocale se définit de manière générale comme étant le processus qui consiste, à partir d'un signal sonore, à différencier les portions qui contiennent de la voix à celles qui n'en contiennent pas. L'acronyme VAD[4] signifie « Voice Activity Detector », ou détecteur d'activité vocale, définit le système qui exécute le processus de détection d'activité vocale.

Le VAD s'applique dans de nombreux systèmes de communications de la voix actuels. Parmi eux, nous citons la téléphonie mobile, la téléphonie par Internet, les systèmes de communications sans fils et la reconnaissance vocale. Plusieurs techniques ont été utilisées pour le développement d'un VAD. Chaque technique possède ses qualités et ses limites, et des recherches dans le domaine apportent régulièrement des résultats de plus en plus performants.

On rencontre plusieurs problématiques lors de la conception d'un détecteur d'activité vocale. Celles que l'on rencontre le plus souvent sont reliées au bruit. Un niveau de bruit trop élevé dégrade la qualité de la voix qu'on désire écouter, ce qui empêche le VAD de bien fonctionner. Pour cette raison, plusieurs méthodes ont été développées pour essayer d'améliorer la qualité du VAD dans des environnements bruités tels que le bavardage, le trafic routier, une usine,... etc.

L'objectif de ce travail est d'implémenter un nouvel algorithme robuste de VAD qui soit capable de travailler en temps réel pour les applications de transmission de la parole telles que la téléphonie et de performer dans les conditions où l'on n'a pas d'informations a priori sur la nature ou le niveau du bruit pour les applications de reconnaissance de locuteur et/ou parole.

Dans les systèmes de reconnaissance de locuteur actuels, les VAD à base d'énergie sont principalement utilisés. Par exemple, dans les évaluations de reconnaissance de locuteurs NIST (indépendantes du texte) (SRE), les sites participants utilisaient typiquement un VAD basé sur l'énergie avec / sans technique de soustraction spectrale comme prétraitement (Brummer et al., 2010). Les auteurs (Kinnunen & Li, 2010) ont également utilisé un VAD basé sur l'énergie qui s'avère utile pour les données vocales du NIST.

En dehors de cela, les détecteurs d'activité vocale statistiques proposés par Sohn et al. sont également utilisés pour la détection d'activité vocale dans les systèmes de reconnaissance de locuteurs (Sohn, Kim, & Sung, 1999).

[4] L'acronyme anglophone VAD, plus largement utilisé comme terme technique, sera utilisé dans le mémoire

Dans ce chapitre, nous passons rapidement en revue quelques techniques VAD populaires qui sont couramment utilisées dans les systèmes de reconnaissance du locuteur.

1.1. VAD à base de l'énergie

Habituellement, un simple VAD basé sur le seuil d'énergie fonctionne correctement dans des conditions propres. L'énergie de la parole est donnée par l'expression suivante :

$$E_i = 10 \log_{10}\left(\frac{1}{N-1}\sum_{n=1}^{N}(x_i[n] - \mu_i)^2 + \varepsilon\right) \quad (5.1)$$

Où $x_i[n]$ désigne le $n^{ième}$ échantillon de la première trame de parole dans un énoncé. $\mu_i = \left(\frac{1}{N}\right)\sum_{n=1}^{N} x[n]$ est la moyenne de l'échantillon de la trame, N est la longueur de trame et ε est une constante arbitraire pour éviter le log de zéro. Ensuite, l'énergie maximale $E_{max} = \max_{i=1,\dots,I}(E_i)$ sur les I trames de l'énoncé est calculée. Enfin, la décision VAD est une simple comparaison de ce niveau maximum, et une contrainte d'énergie minimale à un seuil donné. La règle de décision VAD à base d'Energie est donnée par (Kinnunen & Li, 2010) :

$$(E_i > E_{max} - \theta_{main}) \wedge (E_i > \theta_{min}) \quad (5.2)$$

Où θ_{main} et θ_{min}, désignent les seuils d'énergie primaire et minimum respectivement.

1.2. VAD statistique

Des algorithmes VAD plus récents utilisent des modèles statistiques pour distinguer la parole/ non-parole. Sohn et al (Sohn et al., 1999) ont proposé une technique VAD efficace basée sur le calcul de la moyenne des rapports de Log-Vraisemblance (Log-Liklihood, LR) entre le signal observé et le bruit de fond dans les différents intervalles de fréquence.. Les modèles statistiques peuvent être gaussiens (Sohn et al., 1999). Cependant, pour traiter une grande variété de conditions de bruit, d'autres modèles plus appropriés (Laplacien et Gamma) ont été utilisés (Chang, Kim, & Mitra, 2006).

1.3. VAD Auto-Adaptatif (VQVAD)

Afin de détecter la présence ou l'absence de la parole dans de courts segments de signaux vocaux bruités, les auteurs (Kinnunen & Rajan, 2013) ont proposé une méthode VAD auto-adaptative (Self Adaptive VAD, VQVAD), basée sur les caractéristiques MFCC extraites du signal bruité. Les auteurs ont procédé à une amélioration du signal bruité en appliquant la

soustraction spectrale. La soustraction spectrale est une méthode de débruitage (Amehraye, 2009), qui opère dans le domaine fréquentiel et a pour principe de soustraire une estimé du bruit à partir du signal observé. Après le calcul des énergies du signal amélioré, un tri de ces derniers est effectué. Un pourcentage choisi apriori est utilisé, par la suite pour indexé les trames à minimum et à maximum d'énergie qui sont supposées correspondre, respectivement, aux trames de silence et de parole faiblement marquées. Les modèles de parole et de silence sont formés en utilisant les MFCC correspondants aux indices de ces trames. Enfin, toutes les trames sont marquées en utilisant les modèles formés, avec une contrainte supplémentaire de minimum d'énergie. L'algorithme détaillé de cette méthode est donné par *Algorithme 1*.

1.4. GMM-MAP VAD

Les auteurs Asbai et al (Asbai et al., 2015) proposent une modification de VQVAD pour améliorer ses performances pour les énoncés de courte durée. Ceci est fait en créant deux modèles du monde (UBM) parole et non parole, en mode hors ligne. Ces modèles UBM sont formés via l'algorithme EM, en utilisant la parole d'un grand nombre de locuteurs. Par la suite, ces modèles sont adaptés à un énoncé spécifique en utilisant l'adaptation MAP maximale a posteriori. Toutes les étapes de cette méthode sont schématisées dans la Figure 5.1.

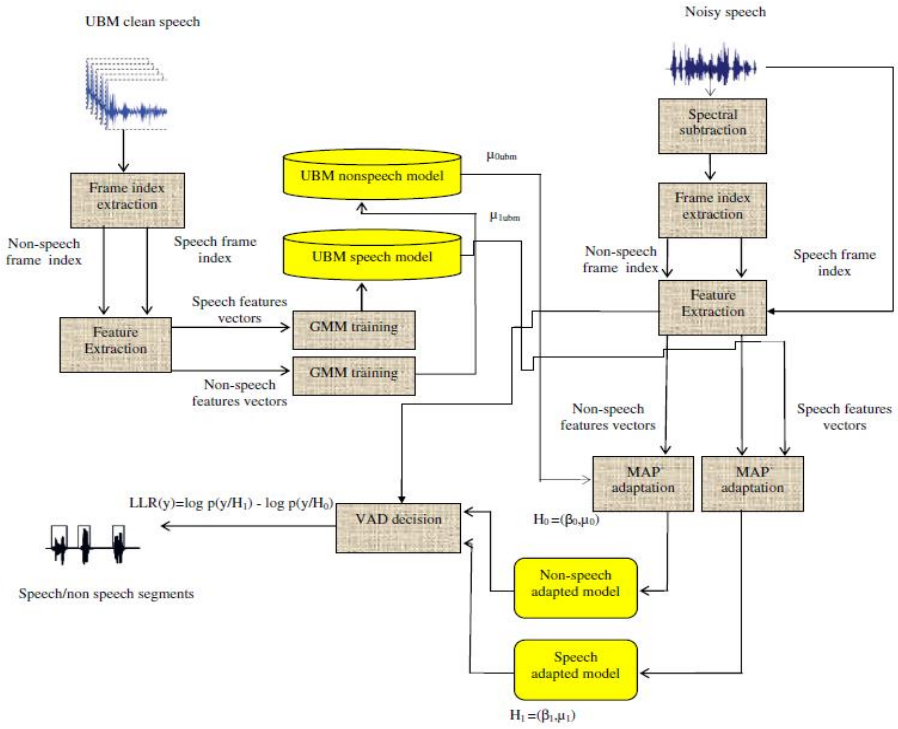

Figure 5.1. Schéma fonctionnel de l'approche GMM-MAP-VAD (Asbai et al., 2015).

Algorithme1 : VAD Auto-adaptive (VQVAD)

***Entrées*:** Signal parole $S(n)$, taille de la trame L et le chevauchement H.

***Sorties*:** Valeurs binaires (0 ou 1) faisant référence respectivement aux segments de silence et de parole.

Début

1. Extraction des MFCC du signal original (bruité)
   $X \leftarrow Extraction\_MFCCs\ (S, L, H, Paramètres\_MFCC)$;
   avec: $Paramètres\_MFCC$ == inclusion ou non du C_0, nombre de coefficients, nombre de filtres Mel, taille de la transformée de Fourier rapide.

2. Débruitage du signal vocal
   $S_{propre} \leftarrow Soustraction\_Spectrale(S, Paramètres\_de\_soustraction\_spectrale)$
   avec: $Paramètres\_de\_soustraction\_spectrale$ == paramètres du filtre de Wienner

3. Calcul des énergies des trames à partir du signal débruité S_{propre} en utilisant l'équation :$E_t = 10log_{10}(\frac{1}{L-1}\sum_{l=1}^{L}(x_t(n) - \mu_t)^2 + \varepsilon)$,
   Où E_t est l'énergie de la $t^{iéme}$ trame, L sa taille et μ_t sa moyenne. Donc :

   $E \leftarrow Calcul\_Energie(S_{propre}, L, H)$;

4. Recherche des indices des trames de haute et basse énergie (un pourcentage de 10% est fixé)
   $[i_{bas}, i_{haut}] \leftarrow Recherche\_indices(E, pourcentage)$;

5. Création des modèles de parole et de silence à partir des trames adéquates
   $\lambda^{parole} \leftarrow Création(x_t \in X | t \in i_{Haut}, Paramètres\_modèle)$;
   $\lambda^{silence} \leftarrow Création(x_t \in X | t \in i_{bas}, Paramètres\_modèle)$;
   avec: $Paramètres\_modèle$ == nombre maximal des itérations de l'algorithme k_means, taille du dictionnaire pour la parole et le silence.

6. Pour toutes les trames, choisir l'hypothèse la plus probable tel que :
   $VAD(t) \leftarrow \left(log\ p(xt|\lambda^{parole}) \geq log\ p\left(xt|\lambda^{silence}\right)\right) \Lambda\ Et \geq \theta_{min}$;
   avec la contrainte de l'énergie minimale ($\theta_{min} = -75$db).

Fin

2. La nouvelle technique LSVM-VAD proposée

Dans cette section, nous allons présenter un nouveau détecteur d'activité vocale destiné à la fois aux applications en temps réel et à la vérification robuste des locuteurs. Le VAD proposé, appelé LSVM-VAD, est basé sur l'utilisation des paramètres MFCC et un simple classificateur linéaire SVM ou Machine à Vecteur du Support formé en mode hors ligne. Toute la procédure d'implémentation de la technique LSVM-VAD est décrite dans la Figure 5.2 et les *Algorithmes 2* et *3*.

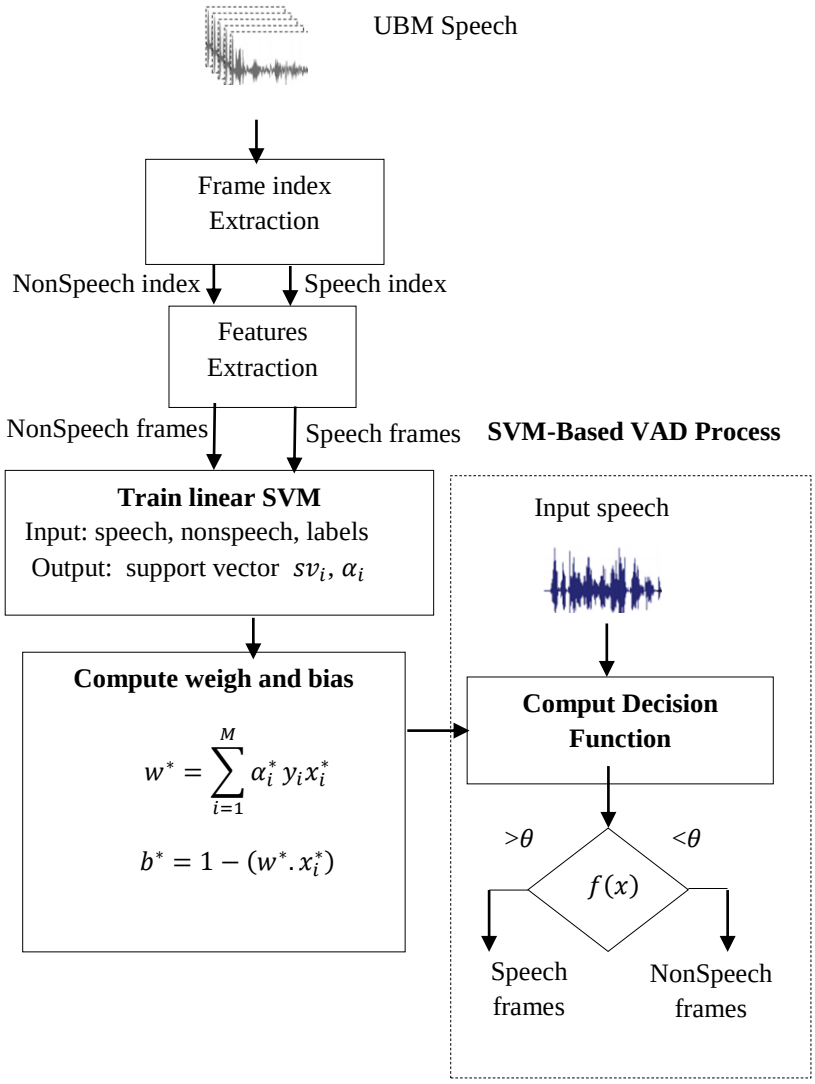

Figure 5.2. Diagramme de la technique LSVM-VAD proposée.

Algorithme 2 : Processus hors-ligne de LSVM-VAD

Entrées: **signal de parole *UBM:*** signal de parole S[n] de tous les locuteurs.

Sorties: **les paramètres de SVM:** le poids: w, le bais :b

Début

1. Débruitage du signal vocal
   $S_{propre} \leftarrow Soustraction\_Spectrale(S, Paramètres\_de\_soustraction\_spectrale)$
   avec: $Paramètres\_de\_soustraction\_spectrale$ == paramètres du filtre de Wienner

2. Calcul des énergies des trames à partir du signal débruité S_{propre} en utilisant l'équation :$E_t = 10log_{10}(\frac{1}{L-1}\sum_{l=1}^{L}(x_t(n) - \mu_t)^2 + \varepsilon)$,
   Où E_t est l'énergie de la $t^{ième}$ trame, L sa taille et μ_t sa moyenne. Donc :

   $E \leftarrow Calcul\_Energie(S_{propre}, L, H)$;

3. Recherche des indices des trames de haute et basse énergie (un pourcentage de 10% est fixé)
   $[i_{bas}, i_{haut}] \leftarrow Recherche\_indices(E, pourcentage)$;

4. Recherche des trames parole et non-parole de modèle du monde UBM
   $S^{UBMSpeech} \leftarrow \{frame(\{x_t \in X \setminus t \in i_{high}\})$
   $S^{UBMnonSpeech} \leftarrow \{frame(\{x_t \in X \setminus t \in i_{low}\})$

5. // Extraction des coefficients MFCC à partir des trames parole et non-parole de modèle du monde UBM
   $X^{Speech} \leftarrow$ ExtractMFCC($S^{UBMSpeech}$);
   $X^{nonSpeech} \leftarrow$ ExtractMFCC($S^{UBMnonSpeech}$);

6. // Estimation des paramètres SVM
   svmstruct←trainsvm ($[X^{Speech}, X^{nonSpeech}], labels$)

7. // Calculer le poids et le bais

$$w^* = \sum_{i=1}^{M} \alpha_i^* y_i x_i^*$$
$$b^* = 1 - (w^*.x_i^*)$$

Fin

Algorithme 3: Processus VAD-LSVM

Entrées : les paramètres de *SVM:* le poids: w, le bais :b, signal de la parole (S), t : trame de parole.

Sorties: les indices parole et non-parole

Début

1. // Extraction des paramètres MFCC à partir du signal de la parole
   $x \leftarrow$ ExtractMFCC(S);
2. // Calcul de la fonction de décision
   $$f(x) = w.x + b$$
3. // Calcul des indices parole et non-parole
   $$\text{VAD}[t] \leftarrow f(x) > seuil$$

Fin

Dans cet algorithme, les paramètres du classificateur SVM sont estimés en mode hors ligne. La règle de décision pour la classification binaire parole/non-parole est apprise sur les paramètres MFCC. Les trames parole et non-parole étiquetées de manière fiable sont obtenues en utilisant la même procédure dans VQVAD. On applique la soustraction spectrale pour améliorer le signal de parole, après on classe les trames en fonction de leur niveau d'énergie et on marque les trames d'énergie les plus élevée comme des trames de parole et les trames d'énergie les plus faible comme des trames de non-parole. Enfin, une trame audio est marquée comme trame de parole si la décision calculée est au-dessus d'un seuil. Dans notre cas, la décision est prise en prenant simplement le signe de l'équation du SVM après le calcul de poids et de biais donnés par les équations 5.3 et 5.4:

$$w^* = \sum_{i=1}^{M} \alpha_i^* \, y_i x_i^* \qquad (5.3)$$

$$b^* = 1 - (w^*.x_i^*), \quad y_i = 1 \qquad (5.4)$$

Avec x les coefficients MFCC (13 MFCC), α_i^*et x_i^* sont les coefficients et les vecteurs de support calculés dans la phase d'entrainement du SVM. La fonction de décision est définie par :

$$w.x + b = 0 \qquad (5.5)$$

Il convient de mentionner à ce stade que les auteurs (Kinnunen, Chernenko, Tuononen, Fränti, & Li, 2007) ont utilisé les SVM et les MFCC pour la détection d'activité vocale, et que leur technique VAD proposée a démontré d'excellentes performances. Néanmoins, notre nouvelle technique LSVM-VAD proposée diffère de celle mentionnée ci-dessus, principalement en raison de la nature non supervisée de notre problème où nous n'avons pas de données étiquetées *apriori* pour former notre système.

4. Expériences et résultats

4.1. Traitement VAD

Dans la première partie des expériences, un test objectif est effectué pour évaluer la performance de la technique LSVM-VAD proposée. Nous comparons cette nouvelle technique à la VAD standard basée sur l'énergie (VAD-Energie) et à l'algorithme récemment publié VQVAD.

Dans le processus LSVM-VAD et afin d'entrainer le SVM linéaire (processus hors-ligne) nous avons utilisé un discours d'une durée moyenne égale à 600 secondes en concaténant les énoncés de 10 fichiers parole hommes et 10 fichiers parole femmes de la base de données TIMIT et un autre discours de 1960 sec en moyenne de la base de données MOBIO en concaténant les énoncés de 15 fichiers parole hommes et 13 fichiers parole femmes. Nous avons appliqué par la suite la soustraction spectrale en utilisant les mêmes paramètres optimisés dans (Kinnunen & Rajan, 2013). Après cela, nous avons trié les valeurs d'énergie et nous avons conservé les trames d'énergie les plus élevées (trames fiables de parole) et les plus basses (trames fiables de non-parole) avec un pourcentage fixe de 5%. Les MFCC correspondant à ces indices de trames sont finalement utilisés pour former un SVM linéaire. Noter que nous avons utilisé des coefficients statiques (y compris c_0) sans aucune normalisation des caractéristiques. La seconde étape du processus LSVM-VAD consiste à l'évaluation de la règle de décision $w.x + b$ en calculant le poids w^* et le biais b^* définis par les équations 5.3 et 5.4.

Tout d'abord, les performances de l'algorithme LSVM-VAD sont évaluées sur la base de données TIMIT en comparant les étiquettes du LSVM-VAD et VQVAD prédites avec une segmentation de référence propre.

Pour choisir la segmentation de référence nous avons utilisé un signal de parole propre de la base de données TIMIT (signal parole féminin) d'une durée de 3.44s. Nous avons obtenu 78.55% trames paroles et 21,45% trames silences en utilisant l'algorithme VAD-Energie. Avec les même conditions (signal parole propre) nous avons obtenu seulement 71,33% trames paroles et 28,67% trames silences en utilisant VQVAD. De plus, comme le montre les histogrammes de la Figure 5.3, VAD-Energie est bon à trouver les régions de parole et de silence comparé au VQVAD. Par la suite, nous avons choisis VAD-Energie comme segmentation de référence pour le calcul de l'erreur qui sera utilisée dans l'évaluation des algorithmes VAD.

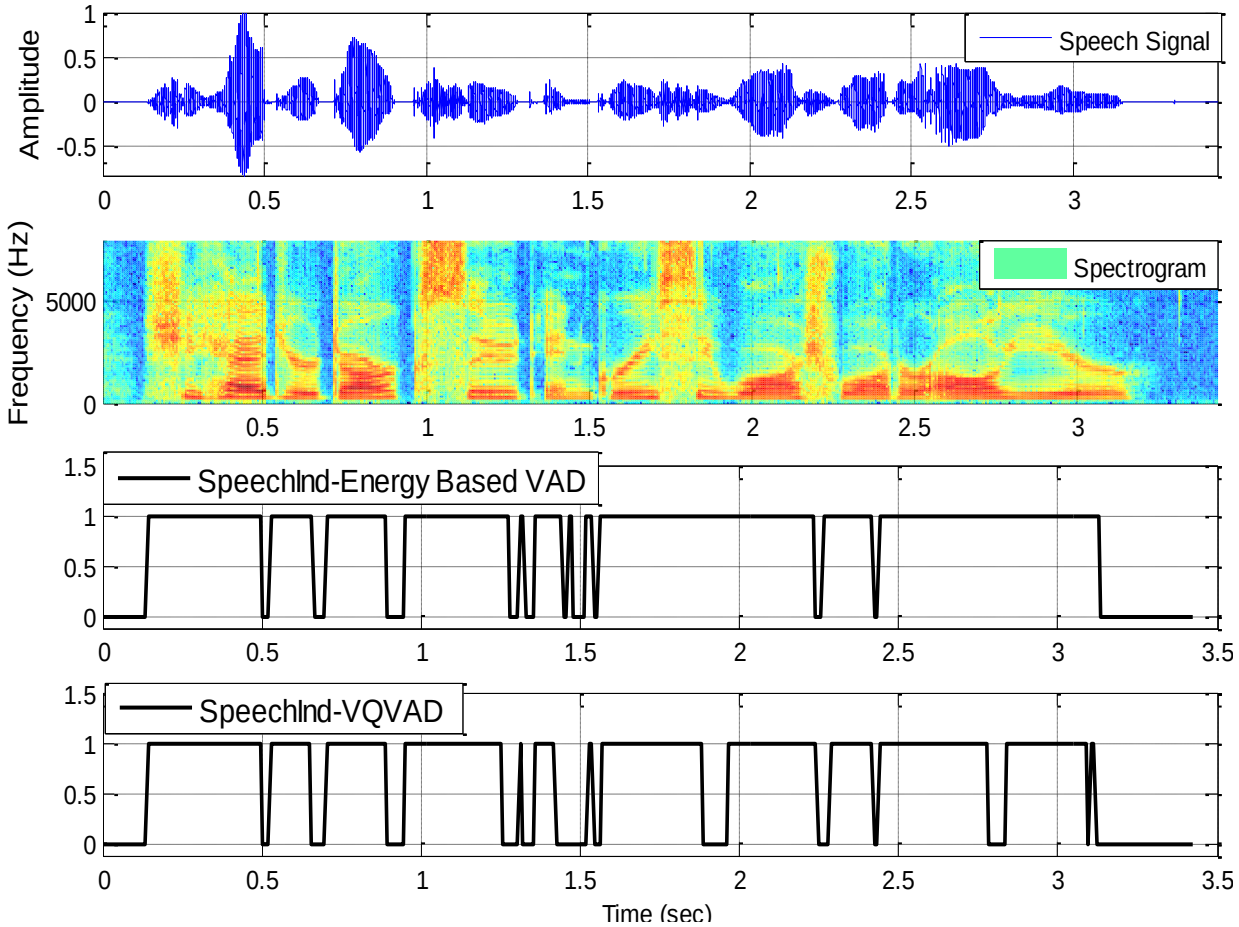

Figure 5.3. Signal de parole (FADG0_SA1.wav) de la base de données TIMIT, Spectrogramme et étiquettes utilisant VAD à base d'énergie et VQ-VAD.

Comme nous l'avons mentionné ci-dessus, notre métrique pour l'évaluation du VAD est le Taux d'Erreur Total Moyen (Average Total Error Rate, ATER). Soit $y_t(n)\ \epsilon\ 0{,}1$ et $\hat{y}_t(n)\ \epsilon\ 0{,}1$ dénotent, respectivement, l'étiquette du VAD originale et prédite de la trame t d'un signal de parole n, et soit T(n) le nombre total de trame. Le Taux d'Erreur Total Moyen est donné par:

$$\varepsilon = \frac{1}{N_{ut}} \sum_{n=1}^{N_{ut}} \frac{1}{T(n)} \sum_{t=1}^{T} I(\hat{y}_t(n) \neq y_t(n)) \qquad (5.6)$$

où I(.) est une fonction d'indicateur.

Nous avons également évalué la performance de LSVM-VAD proposée en termes de taux de fausse acceptation et de faux rejet correspondant à $(y_t(n) = 0\ \Lambda\ \hat{y}_t(n) = 1)$ et $(y_t(n) = 1\ \Lambda\ \hat{y}_t(n) = 0)$, respectivement dans des conditions bruitées. Nous constatons d'après les résultats illustrés dans le tableau 5.1 que le LSVM-VAD surpasse le VQVAD.

De plus, il y a une autre observation qui devrait être soulignée; Nous avons analysé la complexité algorithmique des trois techniques VAD à savoir VAD-Energie, VQVAD et LSVM-VAD, le tableau 5.2 indique que l'algorithme LSVM-VAD proposé nécessite 2.791 ms pour traiter un enregistrement femelle de 3.44 s de la base de données TIMIT. Le facteur temps réel (rt_factor) est 1232.53. Le temps de traitement de LSVM-VAD est presque 100 fois plus petit que la méthode VQVAD et 12 fois plus petit que le VAD-Energie. Ce gain en performance de calcul est plus prononcé dans les énoncés plus longs de la base de données MOBIO comme l'illustre le tableau 5.2.

Table 5.1. Erreur VAD (ε %), taux de fausse alarme *(FAR %)* et taux de faux rejet (*FRR %)* des techniques VQ-VAD et LSVM-VAD par rapport VAD-Energie pour un RSB =5dB.

Bruit (RSB=5dB)	**Evaluation**	**VQVAD/ VAD-Energie**	**LSVM-VAD/ VAD-Energie**
Bavardage	*Erreur*ε(%)	25.96	**19.72**
	FAR (%)	0.30	7.06
	FRR (%)	25.66	12.60
Blanc	*Erreur*ε(%)	45.31	**42.00**
	FAR (%)	*0.13*	7.52
	FRR (%)	45.17	34.47
Usine	*Erreur*ε(%)	40.25	**30.16**
	FAR (%)	0.69	8.90
	FRR (%)	39.55	21.25

Table 5.2. Temps d'exécution des techniques VAD-Energie et LSVM-VAD pour un énoncé donné en utilisant un ordinateur de bureau HP LV2011 avec processeur: Intel Core i3-3220 CPU @ 3.30GHz, RAM: 8Go et disque dur HDD: 465.76 Go.

SystèmeVAD	**TIMIT (durée= 3.44sec)**		**MOBIO (durée=12.78 sec)**	
	Temps de traitement (ms)	*Facteur du temps réel*	*Temps de traitement (ms)*	*Facteur du temps réel*
VAD-Energie	29.68	115.90	24.988	511.44
VQVAD	281.48	12.221	425.85	30.01
LSVM-VAD	**2.791**	**1232.53**	**4.289**	**2979.71**

Avec : $rt\_factor = \frac{la\ durée}{\text{Temps de traitement}}$

4.2. Evaluation de la technique LSVM-VAD pour la tâche de vérification

La technique LSVM-VAD a été d'abord évaluée sur la base de données TIMIT dans des conditions de test bruitées. Pour évaluer les performances, nous comparons la technique LSVM-VAD à la technique VAD-Energie. Les tableaux 5.3 et 5.4 présentent le taux d'erreur égal EER et le MinDCF_{08} respectivement, obtenus en utilisant les algorithmes LSVM-VAD et VAD-Energie dans la présence des bruits de bavardage, blanc et d'usine. A partir de ces tableaux, on peut constater que la dégradation des performances du système de vérification en présence de bruit apparaît nettement en utilisant la technique VAD-Energie, alors que la technique LSVM-VAD résiste mieux.

Table 5.3. Les résultats de la vérification du locuteur en terme de EER dans des conditions bruitées obtenues en utilisant les algorithmes VAD-Energie et LSVM-VAD.

Bruit		**EER (%)**	
	RSB (dB)	*VAD-Energie*	*LSVM-VAD*
Bavardage	*5*	27.16	**21.95**
	10	16.30	**12.62**
	15	7.71	**6.85**
	20	3.92	**3.00**
Blanc	*5*	**42.97**	43.26
	10	35.21	**33.32**
	15	25.30	**18.99**
	20	17.26	**10.57**
Usine	*5*	36.48	**35.32**
	10	29.00	**23.79**
	15	17.38	**13.11**
	20	6.83	**5.70**

Table 5.4. Les résultats de la vérification du locuteur en terme de $MinDCF_{08}$ dans des conditions bruitées obtenues en utilisant les algorithmes VAD-Energie et LSVM-VAD.

Bruit		**$MinDCF_{08} \times 100$**	
	RSB (dB)	*VAD-Energie*	*LSVM-VAD*
Bavardage	*5*	6.84	**6.21**
	10	4.52	**3.94**
	15	2.87	**2.65**
	20	**1.79**	2.05
Blanc	*5*	9.97	**9.91**
	10	9.90	**9.85**
	15	9.70	**9.28**
	20	8.97	**6.99**
Usine	*5*	**9.55**	9.65
	10	7.85	**7.77**
	15	5.68	**4.99**
	20	3.34	**2.84**

4.3. Efficacité de la méthode LSVM-VAD sur une base de données mobile

Dans cette section, la méthode proposée LSVM-VAD est évaluée sur la base de données MOBIO (McCool et al., 2012) en utilisant l'approche I-Vecteur (Dehak et al., 2011).

4.3.1. Description de la base de données MOBIO

La base de données MOBIO contient des enregistrements audiovisuels de 152 personnes (100 hommes et 52 femmes). Cette base de données est enregistrée dans 5 pays européens en deux phases. Chaque phase est composée de 6 sessions. Dans la première phase les locuteurs ont été invités à répondre à une série de 21 questions, les réponses à ces question sont de type : réponses courtes, réponses libres et courtes, réponses fixes et réponses libres. Dans la deuxième phase les locuteurs ont répondu à 11 questions, les réponses à ces questions varient entre : réponses libres, réponses courte et réponse fixes.

Deux appareils mobiles ont été utilisés pour enregistrer la base de données MOBIO, un téléphone portable NOKIA N93i et un micro-portable MacBook 2008. Comme cette base est enregistrée avec de tels appareils, elle contient une quantité importante de bruit. 10% des enregistrements ont un rapport signal sur bruit (Signal Noise Ratio SNR) inférieur à 5 dB et 60% des enregistrements ont un SNR entre 5 et 10 dB. Aussi la durée de ces enregistrements est très

limitée. 25% des enregistrements ont une durée de parole inférieure à 2 secondes et 35% ont des durées entre 2 et 3 secondes.

La base de données MOBIO est partitionnée en trois ensembles.

- Un ensemble d'apprentissage : Les enregistrements de la partie d'apprentissage sont utilisés pour générer le modèle UBM et calculer les matrices de projection.
- Un ensemble de développement : utilisé pour fixées les paramètres comme par exemple le nombre des gaussiennes, et la dimension des espaces. Cette partie est divisée aussi en deux sous-ensemble: ensemble d'enrôlement et ensemble de test. L'ensemble d'enroulement est utilisé pour générer les modèles des locuteurs et l'ensemble de test est utilisé pour calculer les scores.
- Un ensemble d'évaluation: cet ensemble a la même structure de la partie développement. Cette partie est utilisée pour donner les performances globales des systèmes.

Le tableau 5.5 donne un aperçu détaillé de la partition de la base de données entre l'ensemble d'apprentissage, l'ensemble de développement et l'ensemble d'évaluation, y compris le nombre de clients (identités) dans les ensembles, le nombre de fichiers utilisés pour l'enrôlement et le test et les scores qui doivent être calculés pour les expériences.

Plus de détails sur la base de données MOBIO peuvent être trouvé dans (McCool et al., 2012)et sur la page web[5], qui contient des instructions sur la façon d'obtenir les données.

Table 5.5. Table Partitions of the MOBIO database.

	Apprentissage		Développement				Evaluation			
			Enrôlement		**Test**		**Enrôlement**		**Test**	
	Clients	***Fichier***	***Clients***	***Fichier***	***Fichier***	***Scores***	***Clients***	***Fichier***	***Fichier***	***Scores***
Homme	37	7104	24	120	2520	60480	38	190	3990	151620
Femme	13	2496	18	90	1890	34020	20	100	2100	42000
Total	50	9600	42	210	4410	94500	58	290	6090	193620

[5] http://www.idiap.ch/dataset/mobio

4.3.2. Détail d'implémentation

- **Modèle du monde (UBM):** nous avons entrainé un modèle du monde indépendants du genre, contient 512 gaussiennes. Ce modèle est estimé à partir des données téléphoniques de la base de données MOBIO représentées par des vecteurs acoustiques de type MFCC de dimension égale à 60 (19 coefficients MFCC + l'énergie + les premières et les secondes dérivées) en utilisant une fenêtre de type Hamming de 25 ms avec un chevauchement de 10 ms. Les trames parole/ non-paroles sont calculées en utilisant VAD-Energie et LSVM-VAD.
- **Extracteur des i-vecteurs:** la matrice de la variabilité totale T avec 400 facteurs, est estimée à partir des mêmes données utilisées pour l'estimation de l'UBM.
- **Compensation des effets du canal:** Afin de compenser les variabilités, nous avons fait appel à la combinaison de l'analyse discriminante linéaire (LDA) de taille égale à 200 et la normalisation de covariance intra-classe (WCCN) décrite en Section 4.3.2 du chapitre 2.
- **La normalisation des scores:** pour calculer la similarité entre la donnée test et la donnée d'enrôlement, nous avons adopté à une simple mesure de distance cousine. Pour la normalisation des scores, nous avons utilisé C-Norm.

Pour évaluer les performances nous comparons également notre système (*LDA-Cousine)* à un système de base (*NIST Baseline*) implémenté par NIST. Ce système se base sur la distance cousine et sans l'utilisation de la technique de compensation LDA (Figure 5.4).

4.3.3. Résultats et interprétation

Les performances du notre système *LDA-Cousine* à base de LSVM-VAD sont représentées par les valeurs de EER pour l'ensemble de développement et les valeurs HTER (Half Total Error Rate) pour l'ensemble d'évaluation.

A partir des tableaux 5.6 et 5.7, nous observons que notre système *LDA-Cousine* donne des bons résultats par rapport au système *NIST-Baseline*.

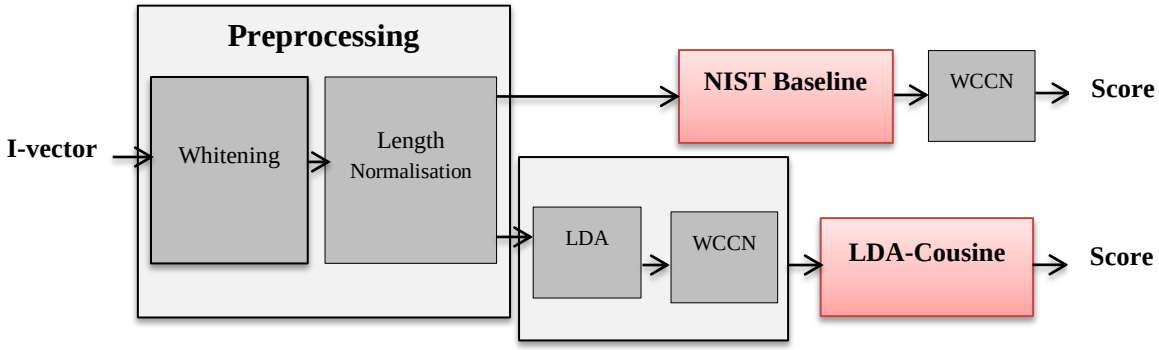

Figure 5.4. Extracteur I-Vecteur.

Les systèmes à base de LSVM-VAD donnent des meilleurs résultats par rapport aux systèmes à base de VAD-Energie. Pour les essais femmes, le HTER de la fusion des deux systèmes a diminué de 22.21% avec la méthode VAD-Energie à 11.96% avec la méthode LSVM-VAD soit une réduction de 10.25% sur l'ensemble d'évaluation et une réduction de 10.70% sur l'ensemble de développement. Pour les essais hommes une réduction de 2.66% a été observée sur l'ensemble d'évaluation.

Table 5.6. Taux d'erreur égal (EER%) sur l'ensemble de développement (DEV) et taux d'erreur total (HTER%) sur l'ensemble de l'évaluation (EVAL) définis sur les essais Homme de la base de données MOBIO.

Ensemble	Système	VAD-Energie	LSVM-VAD
Dev	*NIST Baseline +WCCN+C-norm*	13.9820	**11.8269**
	LDA+ C-norm	**8.7202**	9.9363
	Fusion	**7.9308**	8.0612
Eval	*NIST Baseline +WCCN+C-norm*	14.1960	**11.8692**
	LDA+ C-norm	11.0628	**9.6884**
	Fusion	10.2459	**7.5794**

Table 5.7. Taux d'erreur égal (EER%) sur l'ensemble de développement (DEV) et taux d'erreur total (HTER%) sur l'ensemble de l'évaluation (EVAL) définis sur les essais Femme de la base de données MOBIO.

Ensemble	Système	VAD-Energie	LSVM-VAD
Dev	*NIST Baseline +WCCN+C-norm*	16.1676	15.6387
	LDA+ C-norm	16.8014	16.8430
	Fusion	24.1677	**13.4612**
Eval	*NIST Baseline +WCCN+C-norm*	20.0727	**14.6604**
	LDA+ C-norm	21.2995	**13.3434**
	Fusion	22.2180	**11.9612**

Il est intéressant de noter que notre système de fusion représenté dans le tableau 5.8 atteint un HTER de 7,58% et 11,96% pour les essaies hommes et femmes, respectivement. Comparé aux résultats obtenus par les concurrents dans l'ICB 2013 SRE (Khoury et al., 2013), notre système atteint le deuxième résultat pour les essais hommes après le meilleur système dans l'évaluation, i.e., *Alpineon* qui est basé sur la fusion de 9 sous-systèmes I-vecteur, chacun avec un ensemble de caractéristiques différent, et le troisième résultats pour les essais femmes après *Alpineon* et *Mines-Telecom* (voir Figures 5.5 et 5.6).

Table 5.8. Résultats d'ICB. Ces tableaux répètent les résultats de (Khoury et al., 2013)donnés par EER sur l'ensemble de développement et par HTER sur l'ensemble d'évaluationdes deux protocoles homme et femme de la base de données MOBIO. (Les systèmes marqués par * sont en fait la fusion de différents systèmes, alors que les systèmes marqués d'un + sont ceux qui ont utilisé des données d'entraînement externes / supplémentaires).

ID	SYSTEM	Male		Female	
		EER (%)	HTER (%)	EER(%)	HTER(%)
S-1	*Alpineon****	5.04	**7.08**	7.98	**10.68**
S-2	*L2F-EHU****	7.89	8.14	11.01	13.59
S-3	*Phonexia+*	9.60	10.78	8.36	14.18
S-4	*GIAPSI*[2]	9.68	8.86	11.59	12.81
S-5	*IDIAP*	9.96	10.03	12.01	14.27
S-6	*Mines-Telecom+*	10.20	9.11	11.43	**11.63**
S-7	*L2F****	10.60	11.05	13.48	14.73
S-8	*EHU*	11.31	10.06	17.94	19.51
S-9	*CPqD****	11.82	10.21	14.35	15.99
S-10	*CTDA*	12.74	19.40	19.47	22.64
S-11	*RUN+*	13.73	12.13	13.39	14.09
S-12	*ATVS+*	14.88	15.43	16.84	17.86
S-13	*Méthode proposée*	8.06	**7.57**	13.46	**11.96**

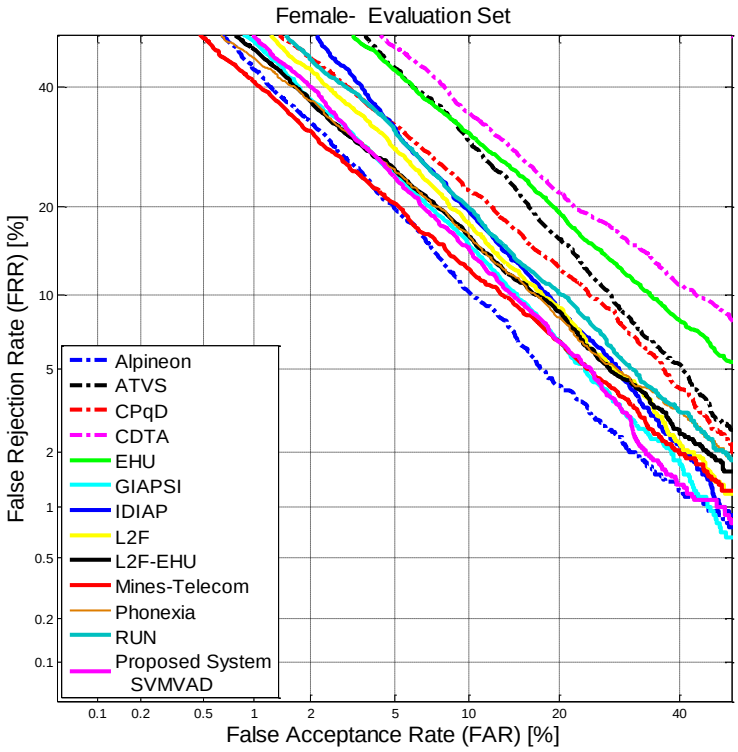

Figure 5.5. Résumé des performances: courbes DET des différents systèmes primaires (Khoury et al., 2013) sur les ensembles EVAL pour les essais Femmes.

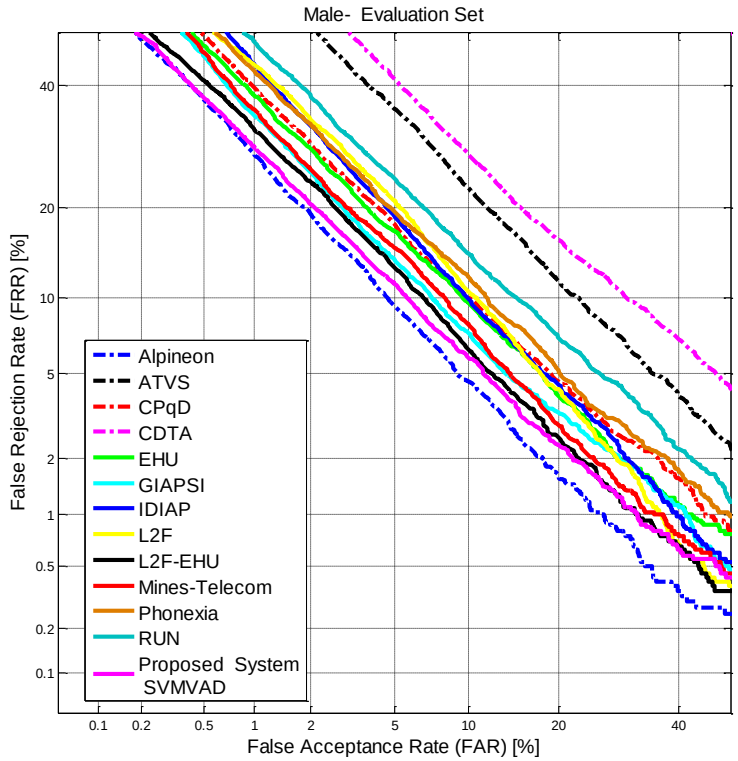

Figure 5.6. Résumé des performances: courbes DET des différents systèmes primaires (Khoury et al., 2013) sur les ensembles EVAL pour les essais Hommes.

5. Conclusion

Nous avons présenté dans ce chapitre un détecteur d'activité vocale facile à mettre en œuvre qui est destiné à la fois aux applications en temps réel et à la vérification robuste des locuteurs. Le LSVM-VAD proposé est basé sur l'utilisation des paramètres MFCC et un simple classificateur linéaire SVM formé en mode hors ligne. La sélection des trames pour l'entraînement est basée sur l'idée intelligente d'un classement d'énergie par trame après une soustraction spectrale. Les résultats obtenus confirment les performances du LSVM-VAD proposé dans la présence du bruit additif ainsi dans l'environnement mobile. De plus, notre nouvelle technique LSVM-VAD proposée présente de bonnes performances comparativement aux récents systèmes présentés dans le concours ICB2013 SRE (Speaker Recognition Evaluation) sur la base de données MOBIO en utilisant la technique I-vecteur.

Conclusions
&
Perspectives

Conclusions

Le principal objectif de cette thèse est d'améliorer la robustesse des systèmes de Reconnaissance Automatique du Locuteur RAL. Certes, la robustesse de ces systèmes est menacée par plusieurs facteurs une fois exploités dans les environnements réels. La première partie de cette thèse est consacrée au traitement de deux facteurs fondamentaux dégradant les performances des systèmes de vérification du locuteur, à savoir, le bruit d'environnement ainsi l'influence de codage, décodage et la compression de la parole dans les communications mobiles. Dans la deuxième partie de cette thèse, nous avons présenté une étude détaillée d'une nouvelle technique de détection d'activité vocale facile à mettre en œuvre, et qui est destinée à la fois aux applications en temps réel et à la vérification robuste des locuteurs.

Nous avons commencé par rappeler, dans la première partie de thèse, le principe de la Reconnaissance Automatique du Locuteur ainsi de présenter les différentes étapes du système de reconnaissance. Nous avons donné, par la suite, un état de l'art sur les travaux réalisés dans ce contexte et nous avons tiré un certain nombre de conclusions:

- La plupart des systèmes RAL utilisent les coefficients cepstraux à fréquence Mel (MFCC). En effet, ces paramètres 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.
- La majorité des systèmes actuels sont basées sur l'utilisation de modèles de mélange de Gaussiennes (GMM) appris par adaptation MAP d'un modèle du monde UBM (GMM-UBM).
- A ce jour, l'information de la phase n'a pas été pleinement exploitée dans le processus d'extraction des paramètres acoustique, cela est dû principalement à la difficulté de traiter sa nature. Néanmoins, récemment, des efforts ont été mis en place afin d'exploiter d'autres représentations de la phase comme l'utilisation de retard de groupe du spectre de phase et le produit spectral. Les résultats obtenus montrent une amélioration significative des performances en termes de taux de vérification.

- Le filtre Gammatone a attiré plusieurs chercheurs travaillant sur la reconnaissance de la parole et de locuteur. Récemment des nouveaux paramètres pour la RAL ont démontré une grande robustesse au bruit et performent mieux que les coefficients MFCC.

Dans ce contexte et afin de coupler la robustesse de filtres Gammatone aux bruits et la capacité de représentation de produit du spectre, nous avons proposé dans cette thèse des nouveaux paramètres appelés GPSCC (Gammatone Product Spectral Cepstral Coefficients). Ces paramètres, testés sur la base de données TIMIT en utilisant les modèles de mélange de Gaussiennes ont démontré une grande robustesse à des conditions de bruit sévères et de même pour un environnement mobile, avec une réduction d'erreur absolue de 12% par rapport à l'utilisation des paramètres de référence MFCC. Nous avons aussi proposé d'autres nouveaux paramètres appelés MGCC (Multitaper Gammatone Cepstral Coefficients) qui dépendent des filtres Gammatones et des fenêtres de pondération. Les résultats Les résultats trouvés ont montré que les paramètres proposés apportent des améliorations significatives en termes des performances.

Pour récapituler, les résultats obtenus, dans cette première partie de thèse, montre que les nouveaux paramètres proposés dérivés de filtres Gammatone sont plus performant que ceux dérivés de filtres Mel classiques.

Dans la deuxième partie de cette thèse, nous avons présenté une étude détaillée d'une nouvelle technique de détection d'activité vocale facile à mettre en œuvre, et qui est destinée à la fois aux applications en temps réel et à la vérification robuste des locuteurs. Le VAD proposé est basé sur l'utilisation des paramètres MFCC et un simple classificateur linéaire SVM formé en mode hors ligne. La sélection des trames pour l'entraînement est basée sur l'idée intelligente d'un classement d'énergie par trame après une soustraction spectrale. Les résultats obtenus confirment les performances de la nouvelle technique LSVM-VAD proposée dans la présence du bruit additif ainsi dans l'environnement mobile, et elle se compare favorablement aux meilleurs systèmes présentés dans le concours ICB2013 SRE (Speaker Recognition Evaluation) sur la base de données MOBIO et avec l'application des méthodes des I-vecteur.

Perspectives

Bien que nos systèmes présentés dans cette thèse répondent à l'objectif que nous nous sommes fixés au départ, à savoir la réalisation d'un système de reconnaissance automatique du locuteur et l'amélioration de sa robustesse dans un environnement réel. Néanmoins, il reste d'autres axes qui peuvent faire l'objet de nos futurs travaux, parmi lesquels :

- Utilisation des méthodes de normalisation et calibration des scores afin d'accroitre davantage la robustesse des systèmes de la vérification du locuteur.
- tester notre système sur d'autres bases de données, comme celle de NIST ainsi que son ajustement pour une identification dans un espace ouvert pour plus de robustesse.

Bibliographie

Alam, M. J., Kenny, P., Ouellet, P., & O'Shaughnessy, D. (2011). *Multi-taper MFCC features for speaker verification using I-vectors.* Paper presented at the Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on.

Alam, M. J., Kinnunen, T., Kenny, P., Ouellet, P., & O'Shaughnessy, D. (2013). Multitaper MFCC and PLP features for speaker verification using i-vectors. *Speech communication, 55*(2), 237-251.

Aloui, K., Nait-Ali, A., & Naceur, M. S. (2018). Using brain prints as new biometric feature for human recognition. *Pattern Recognition Letters, 113*, 38-45.

Alsteris, L. D., & Paliwal, K. K. (2007). Short-time phase spectrum in speech processing: A review and some experimental results. *Digital Signal Processing, 17*(3), 578-616.

Amehraye, A. (2009). *Débruitage perceptuel de la parole.* Télécom Bretagne.

Asbai, N., Bengherabi, M., Amrouche, A., & Aklouf, Y. (2015). Improving the self-adaptive voice activity detector for speaker verification using map adaptation and asymmetric tapers. *International Journal of Speech Technology, 18*(2), 195-203.

Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America, 52*(6B), 1687-1697.

Bengio, S., & Mariéthoz, J. (2004). *The expected performance curve: a new assessment measure for person authentication.* Paper presented at the Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop.

Besacier, L., & Bonastre, J.-F. (1998). *Frame pruning for speaker recognition.* Paper presented at the Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute, 4*(510), 126.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., . . . Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP journal on applied signal processing, 2004*, 430-451.

Blouet, R. (2002). *Approche probabiliste par arbres de décision pour la vérification automatique du locuteur sur architectures embarquées.* Rennes 1.

Bonastre, J.-F., Bimbot, F., Boë, L.-J., Campbell, J. P., Reynolds, D. A., & Magrin-Chagnolleau, I. (2003). *Person authentication by voice: A need for caution.* Paper presented at the Eighth European Conference on Speech Communication and Technology.

Brümmer, N. (2005). Tools for fusion and calibration of automatic speaker detection systems. *URL http://niko. brummer. googlepages. com/focal*.

Brummer, N., Burget, L., Kenny, P., Matejka, P., De Villiers, E., Karafiat, M., . . . Baum, D. (2010). ABC system description for NIST SRE 2010. *Proc. NIST 2010 Speaker Recognition Evaluation*, 1-20.

Brummer, N., Cernocky, J., Karafiát, M., van Leeuwen, D. A., Matejka, P., Schwarz, P., & Strasheim, A. (2007). Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(7), 2072-2084.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery, 2*(2), 121-167.

Chang, J.-H., Kim, N. S., & Mitra, S. K. (2006). Voice activity detection based on multiple statistical models. *IEEE Transactions on Signal Processing, 54*(6), 1965-1976.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing, 28*(4), 357-366.

Dehak, N., Dehak, R., Glass, J. R., Reynolds, D. A., & Kenny, P. (2010). *Cosine similarity scoring without score normalization techniques.* Paper presented at the Odyssey.

Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., & Dumouchel, P. (2009). *Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification.* Paper presented at the Tenth Annual conference of the international speech communication association.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(4), 788-798.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Dimitriadis, D., Maragos, P., & Potamianos, A. (2011). On the effects of filterbank design and energy computation on robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(6), 1504-1516.

Fedila, M., & Amrouche, A. (2012). *Automatic speaker recognition for mobile communications using AMR-WB speech coding.* Paper presented at the Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on.

Fedila, M., Bengherabi, M., & Amrouche, A. (2015). *Consolidating Product Spectrum and Gammatone filterbank for robust speaker verification under noisy conditions.* Paper presented at the Intelligent Systems Design and Applications (ISDA), 2015 15th International Conference on.

Fillon, T. (2004). *Traitement numérique du signal acoustique pour une aide aux malentendants.* Télécom ParisTech.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of human genetics, 7*(2), 179-188.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 29*(2), 254-272.

Garcia-Romero, D., & Espy-Wilson, C. Y. (2011). *Analysis of i-vector Length Normalization in Speaker Recognition Systems.* Paper presented at the Interspeech.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., & Zue, V. (1993). TIMIT acoustic-phonetic continuous speech corpus, linguistic data consortium, Philadelphia. *Fixed step size Steepest descent with optimal step size Conjugate gradient with optimal step size*.

Gelfand, S. A. (2016). *Hearing: An introduction to psychological and physiological acoustics*: CRC Press.

Gerkmann, T., Krawczyk-Becker, M., & Le Roux, J. (2015). Phase processing for single-channel speech enhancement: History and recent advances. *IEEE signal processing Magazine, 32*(2), 55-66.

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing research, 47*(1), 103-138.

Gold, B., Morgan, N., & Ellis, D. (2011). *Speech and audio signal processing: processing and perception of speech and music*: John Wiley & Sons.

Hansson-Sandsten, M., & Sandberg, J. (2009). *Optimal cepstrum estimation using multiple windows.* Paper presented at the Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on.

Hatch, A. O., Kajarekar, S. S., & Stolcke, A. (2006). *Within-class covariance normalization for SVM-based speaker recognition.* Paper presented at the Interspeech.

Hegde, R. M., Murthy, H. A., & Gadde, V. R. R. (2007). Significance of the modified group delay feature in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(1), 190-202.

Hermansky, H., & Morgan, N. (1994). RASTA processing of speech. *IEEE transactions on speech and audio processing, 2*(4), 578-589.

Hu, Y., & Loizou, P. C. (2004). Speech enhancement based on wavelet thresholding the multitaper spectrum. *IEEE transactions on Speech and Audio processing, 12*(1), 59-67.

Ireland, D., Knuepffer, C., & McBride, S. J. (2015). Adaptive multi-rate compression effects on vowel analysis. *Frontiers in bioengineering and biotechnology, 3*.

Jourani, R. (2012). *Reconnaissance automatique du locuteur par des GMM à grande marge.* Université de Toulouse, Université Toulouse III-Paul Sabatier; Université Mohammad V-Agdal de Rabat.

Kenny, P. (2005). Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM, Montreal,(Report) CRIM-06/08-13, 215*.

Kenny, P. (2010). *Bayesian speaker verification with heavy-tailed priors.* Paper presented at the Odyssey.

Kenny, P., Boulianne, G., & Dumouchel, P. (2005). Eigenvoice modeling with sparse training data. *IEEE transactions on speech and audio processing, 13*(3), 345-354.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2006). *Improvements in factor analysis based speaker verification.* Paper presented at the Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007a). Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(4), 1435-1447.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007b). Speaker and session variability in GMM-based speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(4), 1448-1460.

Khoury, E., Vesnicer, B., Franco-Pedroso, J., Violato, R., Boulkcnafet, Z., Fernández, L. M., . . . Cipr, T. (2013). *The 2013 speaker recognition evaluation in mobile environment.* Paper presented at the Biometrics (ICB), 2013 International Conference on.

Kim, C., & Stern, R. M. (2012). *Power-normalized cepstral coefficients (PNCC) for robust speech recognition.* Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.

Kinnunen, T., Chernenko, E., Tuononen, M., Fränti, P., & Li, H. (2007). *Voice activity detection using MFCC features and support vector machine.* Paper presented at the Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication, 52*(1), 12-40.

Kinnunen, T., & Rajan, P. (2013). *A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data.* Paper presented at the ICASSP.

Kinnunen, T., Saeidi, R., Sandberg, J., & Hansson-Sandsten, M. (2010). *What else is new than the Hamming window? Robust MFCCs for speaker recognition via multitapering.* Paper presented at the Eleventh Annual Conference of the International Speech Communication Association.

Kinnunen, T., Saeidi, R., Sedlák, F., Lee, K. A., Sandberg, J., Hansson-Sandsten, M., & Li, H. (2012). Low-variance multitaper MFCC features: a case study in robust speaker verification. *IEEE transactions on audio, speech, and language processing, 20*(7), 1990-2001.

Li, K.-P., & Porter, J. E. (1988). *Normalizations and selection of speech segments for speaker recognition scoring.* Paper presented at the Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on.

Li, Q., & Huang, Y. (2011). An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. *IEEE transactions on audio, speech, and language processing, 19*(6), 1791-1801.

Li, Z., & Gao, Y. (2016). Acoustic feature extraction method for robust speaker identification. *Multimedia Tools and Applications, 75*(12), 7391-7406.

Linde, Y., Buzo, A., & Gray, R. (1980). An algorithm for vector quantizer design. *IEEE Transactions on communications, 28*(1), 84-95.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech communication, 22*(1), 1-15.

Lyon, R. F., Katsiamis, A. G., & Drakakis, E. M. (2010). *History and future of auditory filter models.* Paper presented at the Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on.

Madikeri, S. R., Talambedu, A., & Murthy, H. A. (2015). Modified group delay feature based total variability space modelling for speaker recognition. *International Journal of Speech Technology, 18*(1), 17-23.

Mami, Y. (2003). *Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence.* Télécom ParisTech.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance: National Inst of Standards and Technology Gaithersburg MD.

Martin, A. F., & Greenberg, C. S. (2010). *The NIST 2010 speaker recognition evaluation.* Paper presented at the Eleventh Annual Conference of the International Speech Communication Association.

McCool, C., Marcel, S., Hadid, A., Pietikäinen, M., Matejka, P., Cernocký, J., . . . Levy, C. (2012). *Bi-modal person recognition on a mobile phone: using mobile phone data.* Paper presented at the Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics, 5*(4), 115-133.

Meriem, F., Farid, H., Messaoud, B., & Abderrahmene, A. (2014). *Robust speaker verification using a new front end based on multitaper and gammatone filters.* Paper presented at the Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on.

Meriem, F., Farid, H., Messaoud, B., & Abderrahmene, A. (2017). New Front End Based on Multitaper and Gammatone Filters for Robust Speaker Verification *Recent Advances in Electrical Engineering and Control Applications* (pp. 344-354): Springer.

Morris, A., Wu, D., & Koreman, J. (2004). GMM based clustering and speaker separability in the Timit speech database. *IEICE Transactions on Fundamentals of Communications, Electronics, Informatics and Systems, 85*.

Mowlaee, P., Saeidi, R., & Stylianou, Y. (2016). Advances in phase-aware signal processing in speech communication. *Speech Communication, 81*, 1-29.

Nolan, F. (1980). *The phonetic bases of speaker recognition.* University of Cambridge.

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

Paliwal, K. K., & Atal, B. S. (2003). *Frequency-related representation of speech.* Paper presented at the Eighth European Conference on Speech Communication and Technology.

Patterson, R., Nimmo-Smith, I., Holdsworth, J., & Rice, P. (1987). *An efficient auditory filterbank based on the gammatone function.* Paper presented at the a meeting of the IOC Speech Group on Auditory Modelling at RSRE.

Patterson, R. D., Allerhand, M. H., & Giguere, C. (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *The Journal of the Acoustical Society of America, 98*(4), 1890-1894.

Pelecanos, J., & Sridharan, S. (2001). Feature warping for robust speaker verification.

Pépiot, E. (2014). *Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers.* Paper presented at the Speech Prosody 7.

Pickles, J. O. (2008). An introduction to the physiology of hearing.

Pinto, S., & Sato, M. (2016). *Traité de neurolinguistique: Du cerveau au langage*: De Boeck Superieur.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257-286.

Reynolds, D. A. (2003). *Channel robust speaker verification via feature mapping.* Paper presented at the Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing, 10*(1-3), 19-41.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE transactions on Speech and Audio Processing, 3*(1), 72-83.

Romani, G. L., Williamson, S. J., & Kaufman, L. (1982). Tonotopic organization of the human auditory cortex. *Science, 216*(4552), 1339-1340.

Rose, P., & Clermont, F. (2001). A comparison of two acoustic methods for forensic speaker discrimination. *Acoustics Australia, 29*(1), 31-36.

Sadjadi, S. O., Slaney, M., & Heck, L. (2013). MSR identity toolbox v1. 0: A MATLAB toolbox for speaker-recognition research. *Speech and Language Processing Technical Committee Newsletter, 1*(4).

Senoussaoui, M., Kenny, P., Dumouchel, P., & Dehak, N. (2013). *New cosine similarity scorings to implement gender-independent speaker verification.* Paper presented at the Interspeech.

Shao, Y., & Wang, D. (2008). *Robust speaker identification using auditory features and computational auditory scene analysis.* Paper presented at the Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on.

Smith, J. O., & Abel, J. S. (1999). Bark and ERB bilinear transforms. *IEEE Transactions on speech and Audio Processing, 7*(6), 697-708.

Sohn, J., Kim, N. S., & Sung, W. (1999). A statistical model-based voice activity detection. *IEEE signal processing letters, 6*(1), 1-3.

Sönmez, M. K., Heck, L., Weintraub, M., & Shriberg, E. (1997). *A lognormal tied mixture model of pitch for prosody based speaker recognition.* Paper presented at the Fifth European Conference on Speech Communication and Technology.

Soong, F. K., Rosenberg, A. E., Juang, B. H., & Rabiner, L. R. (1987). Report: A vector quantization approach to speaker recognition. *Bell Labs Technical Journal, 66*(2), 14-26.

Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE, 70*(9), 1055-1096.

Van Vuuren, S. (1996). *Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch.* Paper presented at the Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on.

Varga, A., & Steeneken, H. J. (1993). Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech communication, 12*(3), 247-251.

Venkitaraman, A., Adiga, A., & Seelamantula, C. S. (2014). Auditory-motivated Gammatone wavelet transform. *Signal Processing, 94*, 608-619.

Viikki, O., Bye, D., & Laurila, K. (1998). *A recursive feature vector normalization approach for robust speech recognition in noise.* Paper presented at the Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on.

Vijayan, K., Reddy, P. R., & Murty, K. S. R. (2016). Significance of analytic phase of speech signals in speaker verification. *Speech Communication, 81*, 54-71.

Wilhelm, S. (2012). Prosodie et correction phonétique: Dijon: Université de Dijon.

Ying, L. (2006). Phase unwrapping. *Wiley Encyclopedia of Biomedical Engineering*.

Zhao, X., & Wang, D. (2013). *Analyzing noise robustness of MFCC and GFCC features in speaker identification.* Paper presented at the Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.

Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C., & Shamma, S. (2011). *Linear versus mel frequency cepstral coefficients for speaker recognition.* Paper presented at the Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on.

Zhu, D., & Paliwal, K. K. (2004). *Product of power spectrum and group delay function for speech recognition.* Paper presented at the Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on.

Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *The Journal of the Acoustical Society of America, 68*(5), 1523-1525.

Production scientifique

Articles:

1. Fedila, M., M. Bengherabi, and A. Amrouche, Gammatone filterbank and symbiotic combination of amplitude and phase-based spectra for robust speaker verification under noisy conditions and compression artifacts. Multimedia Tools and Applications, 77**(13)**, pp 16721-16739, 2017. **DOI :** 10.1007/s11042-017-5237-1

   Journal homepage: https://doi.org/10.1007/s11042-017-5237-1

   Impact Factor: 1,53

   Indexation: Science Citation Index Expanded (SciSearch), Journal Citation Reports/Science Edition, SCOPUS, INSPEC, Zentralblatt Math, Google Scholar, ProQuest, Academic OneFile, ACM Digital Library, Current Abstracts, Current Contents/Engineering, Computing and Technology, DBLP, EBSCO Applied Science & Technology Source, EBSCO Associates Programs Source, EBSCO Business Source, EBSCO Computer Science Index, EBSCO Computers & Applied Sciences Complete, EBSCO Engineering Source, EBSCO Ergonomics Abstracts, EBSCO Military Transition Support Center, EBSCO Science & Technology Collection, EBSCO STM Source, EBSCO TOC Premier, EBSCO Vocational Studies, EI-Compendex, Gale, OCLC, PASCAL, SCImago, Summon by ProQuest

   Date de soumission : 15/12/2016
   Date de la 1ère réponse : 12/04/2017
   Date de la 2ème réponse : 11/07/2017
   Date d'acceptation : 12/09/2017
   Date de publication : 30/09/2017

2. Fedila, M. Bengherabi, M. and Amrouche, A. simple yet effective SVM-Based Voice Activity Detector for Speaker Authentication in Mismatched Noise Environment and Mobile Phone. Acoustics Journal, 2018. **Sous Révision**.

Chapitre du livre :

1. Meriem, F., Farid, H., Messaoud, B., and Abderrahmane, A. New Front End Based on Multitaper and Gammatone Filters for Robust Speaker Verification, in Recent Advances in Electrical Engineering and Control Applications2017, Springer. p. 344-354. **DOI:** 10.1007/978-3-319-48929-2_27.
   https://link.springer.com/chapter/10.1007/978-3-319-48929-2_27

Communications internationales:

1. Fedila, M., M. Bengherabi, and A. Amrouche. Consolidating Product Spectrum and Gammatone filterbank for robust speaker verification under noisy conditions. in Intelligent Systems Design and Applications (ISDA), 2015 15th International Conference on. 2015. IEEE. **DOI:** 10.1109/ISDA.2015.7489252 http://ieeexplore.ieee.org/document/7489252/

2. Meriem, F., Farid, H., Messaoud, B., and Abderrahmane, A. Robust speaker verification using a new front end based on multitaper and gammatone filters. in Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on. 2014. IEEE.

3. Zergat, K.Y., Amrouche, A., Fedila, M., and Debyeche, M. Robust Arabic speaker verification system using LSF extracted from the G. 729 bitstream. in Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on. 2013. IEEE. **DOI:** 10.1109/MLSP.2013.6661939 http://ieeexplore.ieee.org/document/6661939/

4. Fedila, M. and A. Amrouche. Influence of G722. 2 speech coding on text-independent speaker verification. in Microelectronics (ICM), 2012 24th International Conference on. 2012. IEEE. **DOI:** 10.1109/ICM.2012.6471385 http://ieeexplore.ieee.org/document/6471385/

5. Fedila, M. and A. Amrouche. Automatic speaker recognition for mobile communications using AMR-WB speech coding. in Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on. 2012. IEEE. **DOI:** 10.1109/ISSPA.2012.6310441 http://ieeexplore.ieee.org/document/6310441/

Communications nationales

1. M. Fedila and A. Amrouche, “SVM-RFE Based Feature Selection for Network Speaker Recogntion Task,” International Conference on Nanoelectronics, Communications and Renewable Energy, Jijel, Algeria, September 2013.

2. M. Fedila**,** A. Amrouche and M. Bengherabi, “A novel voice activity detection approach for automatic speaker verification,” Journées Scientifiques du laboratoire Communication parlée et Traitement des Signaux "LCPTS", USTHB, Janvier 2015.