REPUBLIQUE ALGÉRIENNE DEMOCRATIQUEET POPULAIRE

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

Université YAHIA FARES DE MÉDÉA

Faculté de Technologie

Mémoire de fin d'études de Master

Spécialité : Système de Télécommunication

Étude de la performance du système de reconnaissance du locuteur basé sur des paramètres acoustiques multi-résolution en communication mobile

Encadré par :

- Dr. KROBBA Ahmed

Présenté par :

- ZERDALI Abdellah Okba
- HAMMA Salah Eddine

Année universitaire 2020/2021

Dédicaces

Je dédie cet humble travail à :

Tous ceux qui m'ont soutenu dans mon parcours, commençant par les personnes les plus précieuses que j'ai, mes chers parents qui ont tout sacrifié et n'ont rien retenu afin de m'aider à atteindre mon objectif, sans eux, tout cela n'aurait pas été possible.

Je veux aussi le dédier à mes frères **HOSSAM**, **ALI** et **AHMED** et leur demander pardon pour tous les ennuis que je leur ai suscités et les remercier de m'avoir soutenu en retour.

Et à tous les membres de la famille en particulier mes grand-pères et ma grand-mère **EL-ATRA** (Allah yarhmhom), ma grand-mère **MERIEM** qu'ALLAH la prête une longue vie, **HAMID**, **ABD EL-MADJID, ABD EL-FETAH, NISSOU, AMINE, SALAH, AMEUR et MONCEF.** Et toute la famille **ZERDALI** et mes oncles **SBAA.**

À tous mes amis proches avec qui j'ai passé les meilleurs moments de ma vie et qui étaient toujours là dans mes pires moments aussi, avec eux tout ira bien à la fin**, RIAD, HAMANI, ADEL, YAZID,** et **IDRISS**.

À tous mes collègues de classe avec qui j'ai passé mes années universitaires dans la joie comme dans la peine, et spécialement mon cher ami et mon frère **Sallah Eddine**.

À **AMINA** qui a laissé une marque spéciale éternelle avec son soutien et ses encouragements illimités et pour être ici tout au long à travers le meilleur et le pire.

Okba

Dédicaces

À cœur vaillant rien d'impossible quand il y a la soif d'apprendre, tous viennent à point à qui sait attendre les études sont avant tout, notre unique et seul à tout souhaitant que le fruit de nos efforts fournis jour et nuit, nous mène vers le bonheur fleuri.

Je dédie ce mémoire à :

À mes très chers Parents *autant de phrases et d'expressions aussi éloquentes soient-elles ne sauraient exprimer ma gratitude et ma reconnaissance. Vous avez su m'inculquer le sens de la responsabilité, l'optimisme et la confiance en soi face aux difficultés de la vie. Vos conseils ont toujours guidé mes pas vers la réussite. Vos patiences sans fin, vos compréhensions et vos encouragements sont pour moi le soutien dispensable que vous avez toujours supporter. Je vous dois ce que je suis aujourd'hui et ce que je serai demain et je ferai toujours de mon mieux pour rester votre fierté et ne jamais vous décevoir. Qu'Allah, le Tout-Puissant, vous préserve, vous accorde santé, bonheur, quiétude de l'esprit et vous protège de tout mal.*

À mes très chers frères : ***Mohamed, Abdel Aziz*** *et* ***Abdel Basset***
Mes chères sœurs : ***Fatima Zohra et Nabila,*** *merci d'être toujours à mes côtés, par votre présence, par votre amour dévoué et votre tendresse, pour donner du goût et du sens à ma vie. En témoignage de mon amour et de ma grande affection, je vous prie de trouver dans ce travail l'expression de mon estime et mon sincère attachement. Je prie Allah, le tout puissant, pour qu'il vous donne bonheur et prospérité.*

À ma chère grand-mère *que ce modeste travail, soit l'expression des vœux que vous n'avez cessé de formuler dans vos prières. Qu'Allah vous préserve santé et longue vie.*

À tous les membres des familles : ***Hamma*** *et* ***Maziz*** *Mes oncles, mes tantes ainsi que mes cousins et mes cousines, puisse ce travail témoigner de ma profonde affection et de ma sincère estime.*

À tous ***mes enseignants*** *et toutes* ***mes enseignantes***
Je saisis cette occasion pour vous exprimer mes respects et ma considération. À mon cher binôme ***Okba*** *je suis très content de t'avoir connu et de travailler avec toi durant cette année, tu étais plus qu'un frère pour moi, qu'Allah te réalise tes vœux et te protège…*

À mes meilleurs amis : Hommam, Sadam, Mohamed.
À toutes mes amies et collègues de promotion et d'études
À tous ceux dont l'oubli du nom n'est pas celui du cœur…

SALAH EDDINE

Remerciements

Avant tout, nous remercions ALLAH tout puissant de nous avoir donné la volonté, le courage et la patience pour terminer ce travail.

Notre profond respect et nos sincères remerciements à Monsieur ***A.Krooba****, enseignant à l'université de l'USTHB, pour nous avoir fait confiance, pour avoir accepté de diriger la réalisation de ce mémoire et pour son soutien inconditionnel.*

Nous adressons nos respectueux remerciements à tous les membres du jury qui nous ont fait l'honneur de participer à l'évaluation de notre travail.

Nos vifs remerciements s'adressent également à tous les enseignants qui ont contribué à notre formation tout au long de nos études.

Nos sentiments de reconnaissance et nos remerciements vont également à l'encontre de toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce travail.

Résumé

Ce travail appartient au domaine de la reconnaissance automatique du locuteur, dans lequel l'objectif est d'identifier une personne en analysant sa voix dans la communication mobile. Les systèmes de reconnaissance automatique des locuteurs se basent généralement sur les paramètres MFCC, GFCC et MRCG comme représentation acoustique du signal vocal. Les paramètres sont encore très sensibles aux changements de signal dans les environnements bruyants. Perceptuel Paramètres pour indiquer la variabilité pour motiver les chercheurs à l'utiliser à nouveau les paramètres rendent le modèle acoustique plus robuste. Dans ce travail, nous utilisons un nouveau concept d'énoncé de la parole, représente l'analyse multirésolution des paramètres acoustiques "MFCCs ", "GFCCs " et "MRCGs " sous le terme "i-vecteur". Nous comparons les performances de ces paramètres afin de trouver le paramètre le plus robuste dans les communications mobiles.

Mots clés :

Reconnaissance automatique du Locuteur (RAL), multirésolution, I-vecteur, MFCCs, GFCCs, MRCGs, Environnement mobile.

Abstract

This work belongs to the field of automatic speaker recognition, in which the objective is to identify a person by analysing his voice in mobile communication. The automatic speaker recognition systems are generally based on. The parameters MFCC, GFCC and MRCG as an acoustic representation of the voice signal. The settings are still very sensitive to signal changes in noisy environments. Perceptual Parameters to indicate variability to motivate researchers to use it again. The parameters make the acoustic model more robust. In this work, we use a new concept of utterance of speech, represents the multiresolution analysis of acoustic parameters " MFCCs ", "GFCCs " and "MRCGs " under the term "i-vector". We compare those parameter performance in order to find the most robust parameter in mobile communications.

Keywords :

Automatic Speaker Recognition (RAL), multiresolution, MFCCs, GFCCs, MRCGs, Mobile environnent.

ملخص

يندرج هذا العمل في ميدان التشخيص الأوتوماتيكي للمتكلم الذي يهدف إلى التعرف على الشخص من خلال تحليل صوته في الاتصالات المتنقلة.

أنظمة التعرف الأوتوماتيكي المتكلم تستند عادة على المعاملات MRCG و MFCC, GFCC شارة الصوتية.

حساسية هذه المعاملات لتغيرات الإشارة حفز الباحثين على استخدام عدة معايير و نماذج جديدة لجعل النماذج الصوتية أكثر قوة.

في هذا العمل قمنا باستعمال multirésolution لتحليل المعاملات الصوتية عن طريق معامل جديد لتمثيل بيان الصوت وهو "i-vecteur" والقيام بدراسة مقارنة بين الأنظمة " MFCCs", "GFCCs " و " MRCGs" واستنتاج التمثيل النموذجي الأمثل في الاتصالات المتنقلة.

كلمات مفتاحية:

التشخيص الأوتوماتيكي للمتكلم, بيئة المتلقي. multirésolution, "MFCCs ", "GFCCs ", "MRCGs".

Abréviations

CASA	Computational Auditory Scene Analysis
CDS	Centre de Développement des Satellites
DCT	Discret Cosinus Transform
DTW	Dynamic Time Warping
EER	Equal Error Rate
ERB	Equivalent Rectangular Bandwidth
FA	Fausses Acceptations
FAR	False Acceptation Rate
FR	Faux Rejets
FRR	False Rejection Rate
GFCC	Gammatone Frequency Cepstral Coefficients
GMM	Gaussian Mixture Models
GMM-UBM	Gaussian Mixture Model -Universal Background Model
IAL	Identification automatique du Locuteur
IBM	International Business Machines
IDFT	Inverse Discret Fourier Transform
JFA	Joint Factor Analysis
LDA	Linear Discriminant Analysis
MFCC	Mel Frequency Cepstral Coefficients
MRCG	Multi-Résolution CochléaGramme
NIST	National Institute of Standards and Technology

NAP	Nuisance Attribut Projection
PLDA	Probabilistic Linear Discriminant Analysis
RAL	Reconnaissance automatique du Locuteur
RAP	Reconnaissance automatique de la Parole
SNR	Rapport Signal sur Bruit
TIMIT	Texas Instrument, Massachusetts Institute of Technology
UBM	Universal Background Model
VAL	Vérification Automatique du Locuteur
WCCN	Within Class Covariance Normalization

Liste des figures

Figure I-1 : les différentes tâches du traitement de la parole. 3
Figure I-2 : Schéma de Principe de base de l'identification du locuteur. 4
Figure I-3: Vérification automatique du locuteur. 5
Figure I-4: Tâche de suivi de locuteurs 6
Figure I-5: Principe de base de la segmentation en locuteurs. 7
Figure I-6 : Structure générale d'un système RAL. 10
Figure I-7 : Schémas synoptiques de l'extraction de paramètres acoustiques. 11
Figure I-8: schéma typique d'un système de vérification de locuteur. 17
Figure II-1 : Organisation tonotopique cochléaire (A) et réponse de la membrane basilaire aux stimulations acoustiques (B, C, D). 24
Figure II-2 : Réponse en fréquence du banc de filtre Gammatone à 64 canaux. 25
Figure II-3 : Visualisation de la fonction MRCG. 29
Figure III-1: Structure générale d'un système de vérification du locuteur. 34
Figure III-2: Comparaison des taux d'erreurs (MFCC, GFCC et MRCG avec les différents niveaux bruit de foule. 37
Figure III-3: Comparaison des taux d'erreurs (MFCC, GFCC et MRCG) avec les différents niveaux bruit des voitures. 37
Figure III-4: Comparaison des taux d'erreurs (MFCC, GFCC et MRCG) avec les différents niveaux bruit d'usine. 38

Liste des tableaux

Tableau II-1: Différences en MFCC et GFCC Dérivation.....23
Tableau III-1 : Description des différents types de bruits du corpus Noisex-92. 33

Table des matières

Table des matières

Dédicaces **i**
Remerciements **iv**
Résumé **v**
Abréviations **vii**
Liste des figures **ix**
Liste des tableaux **x**
Table des matières **xii**
Introduction générale **1**
I.1 Introduction **2**
I.2 Principe de la reconnaissance automatique du locuteur **2**
I.2.1 Identification automatique du locuteur 3
I.2.2 Application IAL 4
I.2.3 Vérification automatique du locuteur 5
I.2.4 Application VAL 5
I.3 L'indexation par locuteur d'un flux multi-locuteurs **5**
I.3.1 Suivi de locuteur 6
I.3.2 Segmentation en locuteurs 6
I.3.3 Applications 7
I.4 Sources de variabilité du signal de la parole **7**
I.4.1 Variabilité intra-locuteur 8
I.4.2 Variabilité interlocuteur 8
I.4.3 Variabilité due au matériel 9
I.5 Mode dépendant et indépendant du texte **9**
I.6 L'architecture d'un système RAL **10**
I.6.1. La paramétrisation 10
I.6.2 Extraction des paramètres 11
I.6.3 Les Paramètres acoustiques 11
I.6.4 Modélisations des locuteurs 13
I.6.5 Quantification vectorielle 14
I.7 Décision et mesure des performances **15**
I.7.1 L'identification automatique du locuteur 15
I.7.2 Vérification automatique du locuteur 16
I.7.3 Taux d'Égal Erreur EER 16
I.8 les sources d'erreurs **17**
I .9 Conclusions **17**

II.1 Introduction **20**
II.2 Principe de multirésolution **20**
II.3 I-vecteur **21**
II.4 Analyse multirésolution des paramètres **22**
II.4.1 Différences entre MFCC et GFCC 22
II.5 Paramètres GFCC **23**
II.5.1 Filtre Gammatone 23
II.6 Cochléagramme multi-résolution **26**
II.6.1 Construction de MRCG 26
II.6.2 Analyse du MRCG 28
II.7 Conclusion **30**
III.1 Introduction **32**
III.2 Logiciels et langages utilisés dans l'application **32**
III.3 Description des bases de données utilisées **32**
III.3.1La base de données NIST 2008 32
III.3.2 la base de données NOISEX-92 33
III.4 Étapes de la réalisation des systèmes de vérification du locuteur **33**
III.5 la structure générale de Système de vérification du locuteur **34**
III.6 Extraction des paramètres **35**
III.7 Evaluation des résultats **35**
III.8 Expérience **35**
III.8.1 Effet du bruit de foule 36
III.8.2 Effet de bruit des voitures 37
III.8.3 Effet de bruit d'usine 38
III.9 Conclusion **39**
Conclusion générale **41**
Bibliographie **43**

Introduction générale

La révolution des machines et notamment des ordinateurs, chercheurs et industriels se sont efforcés d'entendre l'usage de la parole à la communication personne-machine. La parole est un moyen privilégie de communication de l'homme. Le traitement automatique de la parole a considérablement progresse ces dix dernières années, notamment dans les domaines de la Reconnaissance Automatique de la Parole (RAP), Reconnaissance Automatique du locuteur (RAL) et la synthèse de la parole.

La croissance de l'usage des téléphones mobiles cellulaires a connu un essor remarquable, au cours ces dernières années. L'intégration de technologie vocale dans la communication mobile a créé de nombreux nouveaux défis dans le domaine de reconnaissance du locuteur et la parole telle que : la distorsion de codage de la parole, le canal de transmission, bruit d'environnement mobile.

La RAL consiste à reconnaitre l'identité d'une personne par analyse de sa voix. Néanmoins, d'autres éléments que la voix peut être utilisée pour authentifier une personne tels que les empreintes digitales ou les empreintes génétiques. Contrairement à la voix, ces derniers éléments sont une composante du corps humain, ils ne varient pas ou très peu dans le temps et ne peuvent être modifiés sciemment par un individu. De par ces propriétés, le terme de biométrie est souvent employé pour les définir et souligne leur très grande fiabilité. Les gestes de parole ne sont, en aucune façon, un élément du corps humain et ne sont pas reproductibles à l'identique dans le temps. Dans ce sens, les dénominations de biométrie ou d'empreinte vocale ne sont pas appropriées pour caractériser la voix. Cependant, la voix reste pour certaines applications (services accessibles par réseau téléphonique) le seul élément disponible pour authentifier l'utilisateur.

Plusieurs progrès et méthodes en termes de performances et de robustesse des systèmes de la reconnaissance du locuteur dans une communication mobile. Le tout récent progrès, voire même le plus important, est la proposition de la représentation de la parole par des vecteurs à faible dimension, dénommés les i-vecteurs qui s'appuie

sur le paradigme de modélisation par mélange de gaussiennes et présente la particularité de se réduire numériquement à un vecteur de dimension faible, au regard des représentations précédentes, et pourtant très discriminantes vis-à-vis du locuteur. En 2014, Wang et al, ont utilisé un ensemble fixe de caractéristiques complémentaires qui incluent un spectrogramme de modulation d'amplitude, un coefficient de prédiction linéaire perceptuelle transformée spectrale, un Coefficient Cepstruel à Fréquence Mel (MFCC) et un caractéristique gamma à 64 canaux. De plus, Chen et al, ont proposé une nouvelle fonctionnalité appelée cochléagramme multi-résolution (MRCG) [1]. Cependant, la dimension MRCG est si grande que la complexité de l'algorithme est très élevée.

Dans le présent travail, nous nous sommes intéressées à la robustesse des systèmes de reconnaissance du locuteur dans une communication mobile basée sur l'espace des I vecteurs.

Le présent mémoire se subdivise en trois chapitres :

- Le premier chapitre est consacré aux principes de la reconnaissance automatique du locuteur (RAL) et des approches de modélisations utilisés dans ce domaine ainsi que l'exploration des différentes tâches de la reconnaissance du locuteur. On explore alors les outils nécessaires à la reconnaissance automatique du locuteur comme paramétrisation, les approches de modélisation du locuteur et la mesure des performances de chaque tâche.
- Le deuxième chapitre, définie le principe de multi-résolution et l'étude de la performance d'analyse multi-résolution des paramètres acoustiques et bruit d'environnements mobile.
- Le troisième chapitre, présente une étude expérimentale portera les résultats obtenus avec les trois systèmes mis au point, le premier système à base des coefficients MFCCs, le deuxième à base des coefficients GFCCs et le troisième à base des coefficients MRCGs, dans les trois environnements du bruit (foule, bruit de voiture et bruit d'usine).

Finalement, on termine par une conclusion générale retraçant les différentes étapes réalisées, résumant les résultats obtenus et soulignant les perspectives qui peuvent envisager dans la future clôture ce mémoire.

Chapitre I :

Généralités sur la reconnaissance automatique du locuteur

I.1 Introduction

La reconnaissance automatique du locuteur est un terme général qui catégorise les problèmes liés à la reconnaissance ou à la vérification du locuteur en fonction des informations contenues dans le signal sonore, il consiste à reconnaître une personne à partir de sa voix. Le système de reconnaissance du locuteur est divisé en trois étapes: l'analyse acoustique du signal vocal, la modélisation du locuteur et la dernière étape prisent la décision finale.

Dans ce chapitre, nous présentons le principe de la reconnaissance automatique du locuteur et nous introduisons les différentes étapes de ce système.

I.2 Principe de la reconnaissance automatique du locuteur

Comme montre, la figure I.1, la reconnaissance automatique du locuteur fait partie du domaine plus général du traitement de la parole, qui tire parti de la variabilité du locuteur et s'intéresse aux informations extralinguistiques du signal de parole. Il existe deux sources fondamentales de différences individuelles entre les locuteurs. Tout d'abord, les caractéristiques morphologiques de l'appareil vocal de chaque locuteur sont différentes, quelle que soit la phrase prononcée, alors deux locuteurs ne prononceront pas la même phrase de la même manière. Nous observerons la différence de vitesse de livraison, le degré de changement de ton ou la différence liée à son origine sociale et culturelle. Cette variabilité est l'essence même du la reconnaissance automatique du locuteur [2].

La reconnaissance automatique du locuteur peut être la méthode la plus ergonomique pour résoudre les problèmes d'accès, en particulier dans les transactions téléphoniques. Cependant, la parole ne peut pas être considérée comme une caractéristique biologique.

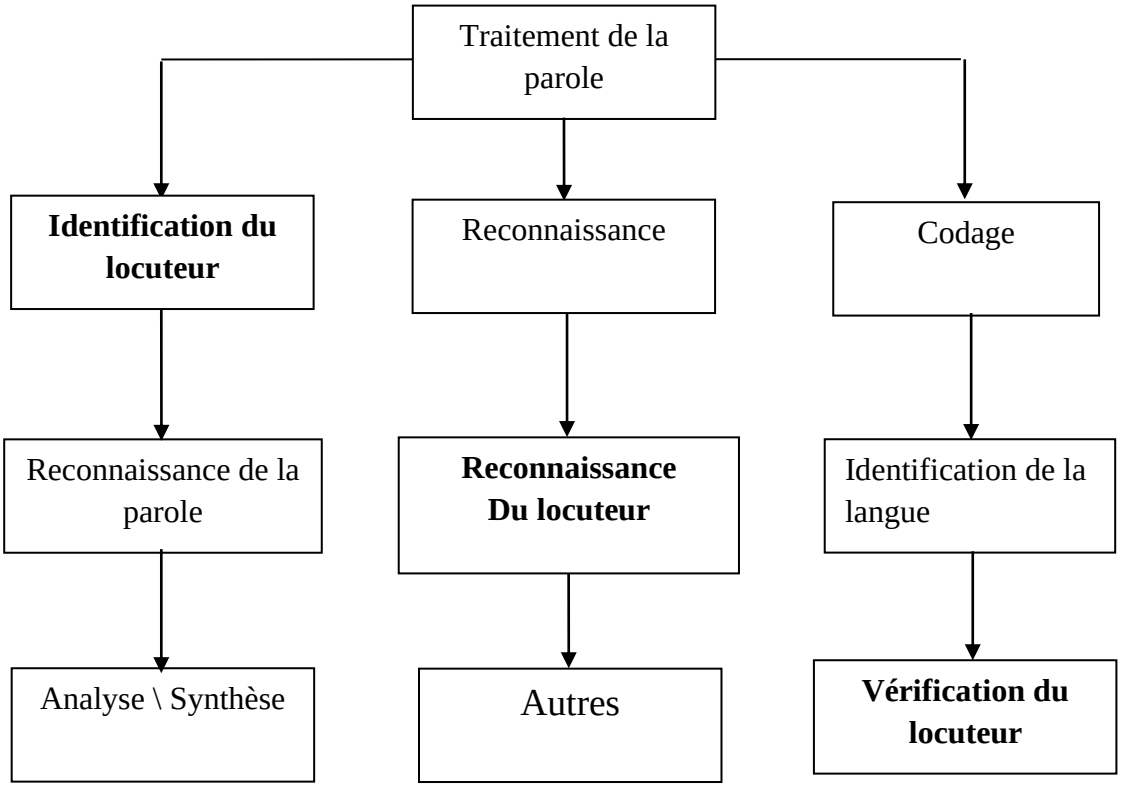


Figure I-1 : les différentes tâches du traitement de la parole [2].

Aucune entrée de table d'illustration n'a été trouvée. La figure d'une personne compte tenu de la variabilité interne du locuteur. Par conséquent, il est préférable de le limiter aux signatures vocales plutôt qu'à l'impression vocale.

Le discours a également impliqué un large éventail d'activités, telles que: le codage, Compression, synthèse, reconnaissance de la parole et reconnaissance du locuteur. Selon scénarios d'application envisagés. Les scénarios d'application sont divisés en trois catégories principales [3]:

- l'identification de locuteurs.
- La vérification du locuteur.
- l'indexation par locuteur d'un flux audio.

I.2.1 Identification automatique du locuteur

Le principe de l'identification automatique du locuteur consiste à reconnaitre une personne parmi un ensemble de locuteurs en comparant son expression vocale à des références connues. Du point de vue schématiquement, une séquence vocale est donnée à l'entrée du système d'IAL. Pour chaque locuteur connu du système, la

séquence vocale est "comparée" à la référence caractéristique du locuteur. L'identité du locuteur dont la référence est la plus "proche" de la séquence de parole est donnée en sortie du système d'IAL [4].

Deux modes sont disponibles dans l'IAL: l'identification de l'ensemble fermé, qui est présumée que la séquence de parole est effectivement prononcée par un locuteur connu et une identification ouverte pour laquelle le locuteur ne peut pas être connu. En mode ouverte «Open set», le système doit prendre une décision sur la crédibilité de son jugement en acceptant ou rejet de l'identité trouvée. Selon le principe - pour déterminer l'identité parmi identités potentielles - les performances de l'IAL se dégradent lorsque la population de locuteurs augmente [5] :

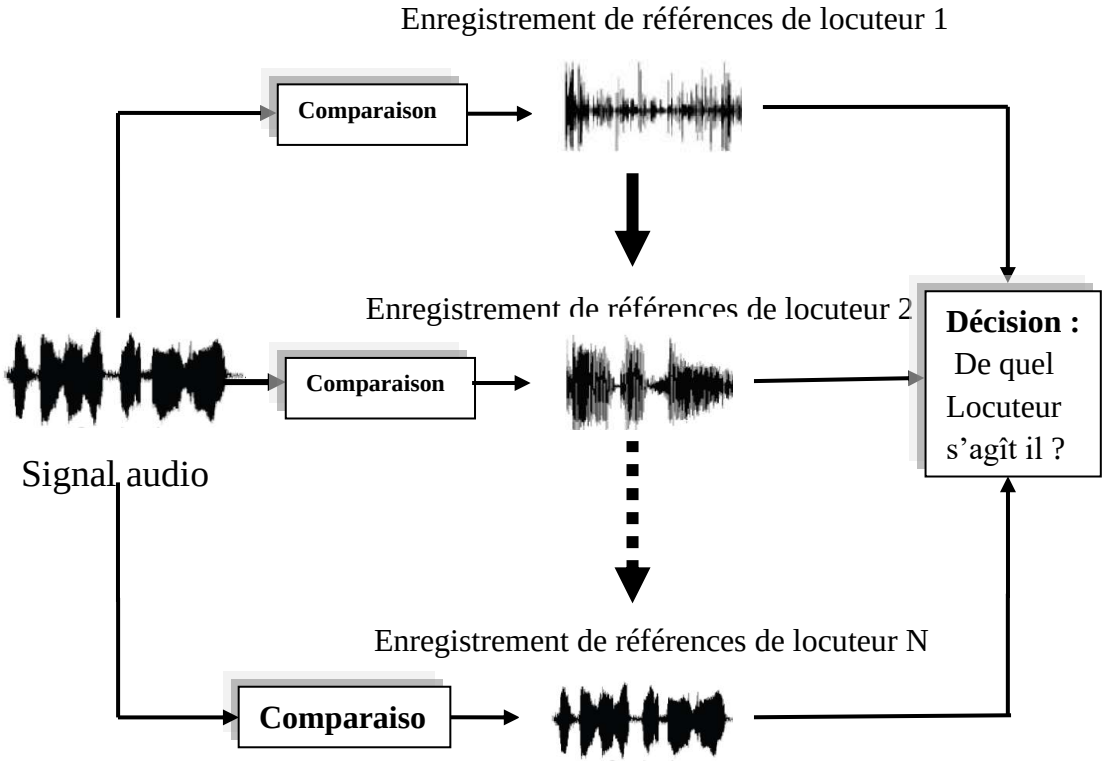


Figure I-2: Schéma de principe de base de l'identification du locuteur [6].

I.2.2 Application IAL

Il n'y a presque pas d'applications d'IAL. On peut continuer par exemple à utiliser système IAL pour s'adapter au locuteur du système de Reconnaissance automatique de la Parole (RAP). De plus, cela peut être intéressant pour les applications, les entreprises associent le même mot de passe à quelques intervenants

(Membres de la famille, membres de la société). Dans ces circonstances, les collections ouvertes et liées au texte peuvent être utilisées pour contrôler l'accès aux données sensible [4].

I.2.3 Vérification automatique du locuteur

La vérification automatique du locuteur (VAL) consiste à accepter ou refuser l'identité proclamée par un locuteur, en se basant sur un modèle qui lui est associé. Elle traite le scénario où le système prend en entrée un énoncé de test ainsi qu'une identité proclamée. La tâche consiste alors à prendre une décision binaire qui va confirmer ou infirmer le fait que l'enregistrement de test est effectivement prononcé par le locuteur proclamé [7] :

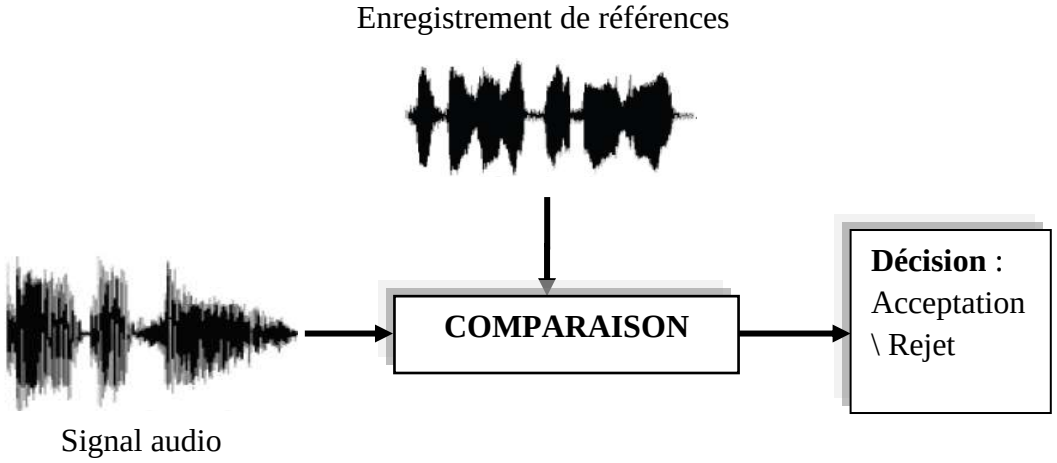


Figure I-3: Vérification automatique du locuteur [7].

I.2.4 Application VAL

VAL a de nombreuses applications, principalement commerciales par exemple [7]:

- Serrures vocales pour le contrôle d'accès à des locaux.
- Authentification pour l'accès à distance.
- Protection de matériel contre le vol (téléphones portables, voitures, etc.)

I.3 L'indexation par locuteur d'un flux multi-locuteurs

Le principe est de déterminer à partir de l'enregistrement de référence du locuteur si locuteur apparaît dans un enregistrement multi-locuteurs, comme une conversation [7]**.**

I.3.1 Suivi de locuteur

Le suivi de locuteur consiste à trouver les limites des interventions du locuteur qu'on a recherché au sein du document audio. Il s'agit donc de déterminer si ce locuteur intervient et si oui, quand. La figure I-4 donne une illustration de ce principe [6].

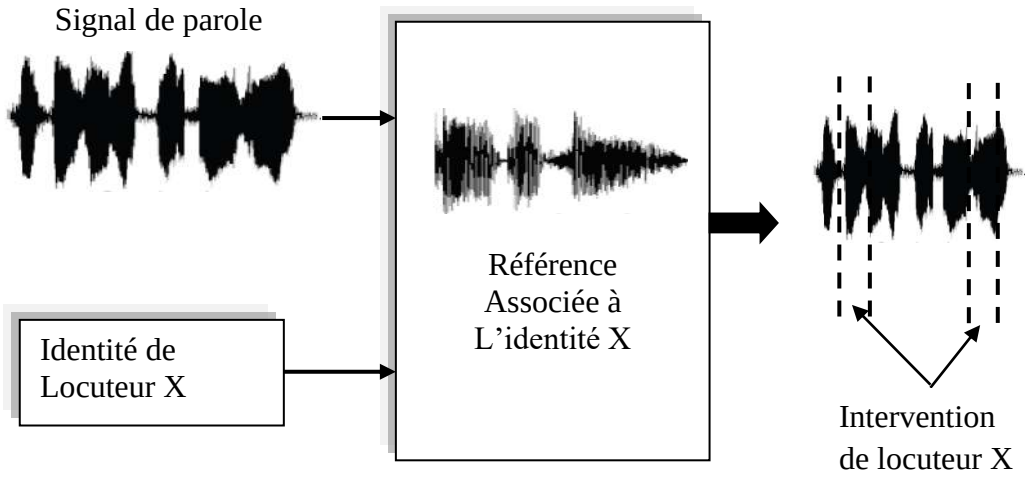


Figure I-4: Tâche de suivi de locuteurs [6].

I.3.2 Segmentation en locuteurs

C'est la détermination du nombre de locuteurs présents dans un document audio tout en délimitant leurs interventions, ça veut dire nous voudrions savoir qui parle et quand. La complication de cette tache résulte du traitement de documents pour lesquels peu ou pas d'informations sont connues a priori. Notamment, pas d'information n'est disponible à la primitive concernant les locuteurs participant dans le document : ni leur nombre, ni leur identité, ni aucun échantillon de leur voix permettant d'avoir une référence. Toutes ces informations doivent être extraites du document étudié [6] :

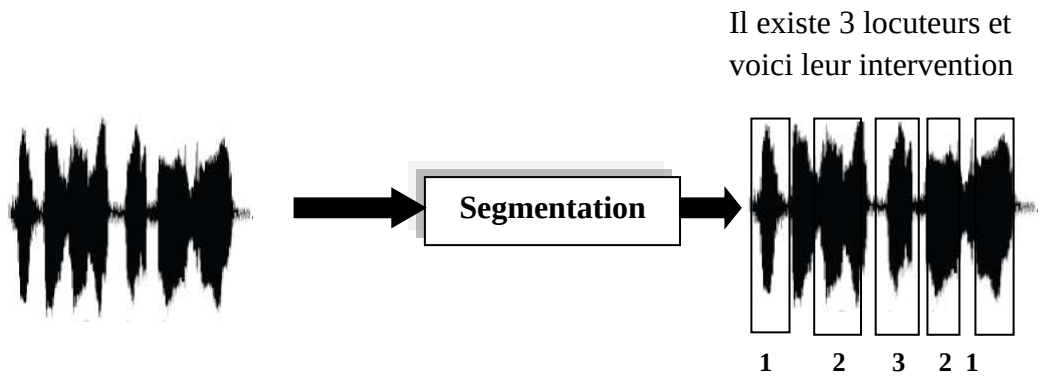


Figure I-5 : Principe de base de la segmentation en locuteurs [6].

I.3.3 Applications

Le domaine d'application est la segmentation automatique des échanges radio entre pilotes et contrôleurs aériens. Depuis lors, la portée de la segmentation des locuteurs s'est élargie et la tâche a été intégrée dans un cadre plus large d'indexation des locuteurs dans les bases de données de documents multimédias [6].

I.4 Sources de variabilité du signal de la parole

Le signal de parole est très complexe où se mêlent informations linguistiques, informations caractéristiques du locuteur, informations relatives au matériel utilisé pour la transmission ou l'enregistrement du signal, etc. En outre, le signal de parole est très redondant. Cette caractéristique est d'ailleurs reconnue pour faciliter la communication entre deux personnes dans un environnement très bruyant. Par ces différents aspects, le signal de parole présente une très grande variabilité.

La capacité des systèmes de RAL à différencier plusieurs individus repose essentiellement sur la variabilité interlocuteur la disposition du signal de parole à varier entre différents individus. Néanmoins, le signal de parole renferme d'autres types de variabilité qui rendent problématique la tâche de reconnaissance, telle que la variabilité intra-locuteur ou la variabilité due au matériel. Par ailleurs, les systèmes de RAL doivent faire face à d'autres difficultés liées davantage au domaine applicatif, comme l'utilisation des systèmes dans des conditions difficiles, les tentatives d'imposture, etc. [8].

I.4.1 Variabilité intra-locuteur

Si le signal de parole est variable entre deux individus, il varie également pour un même individu. Cette variabilité intra-locuteur est induite par l'évolution naturelle ou volontaire de la voix d'une personne. Cette évolution peut être :

- Ponctuelle ou à très court terme. L'état pathologique (fatigue, rhume, etc.) ou émotionnel (stress) d'une personne provoque des altérations momentanées dans sa voix. Dans ce sens, la voix d'une personne peut évoluer entre le début et la fin de la journée (fatigue, irritation due à la pollution) [9, 10]. D'autre part, il est impossible pour un individu de répéter consécutivement deux phrases identiques et de produire un même signal de parole pour ces deux phrases. Une légère variation est toujours observée. Finalement, une personne a la possibilité de modifier volontairement sa voix.
- À moyen terme. En RAL, le comportement d'un individu se modifie au fur et à mesure de son utilisation du système. L'individu devient de plus en plus confiant et sa voix évolue dans ce sens.
- À long terme. La voix change au fur et à mesure du vieillissement d'une personne.

I.4.2 Variabilité interlocuteur

Les signaux de parole véhiculent plusieurs types d'informations. Parmi eux, la signification du message prononcé est d'importance primordiale. Cependant, d'autres informations telles que le style d'élocution ou l'identité du locuteur jouent un rôle important dans la communication orale. Écouter un interlocuteur permet d'avoir des indications concernant son sexe, son état émotionnel et bien souvent de l'identifier si on l'a déjà entendu. Dans notre vie quotidienne, ces informations sont très utiles. Elles nous permettent, par exemple, de différencier les divers messages que nous entendons selon le locuteur et leur degré d'importance. Si toutes les voix étaient perçues de la même façon, il serait par exemple impossible de suivre une émission radio faisant participer des personnes différentes. La grande variabilité entre les locuteurs est due, d'une part, à l'héritage linguistique et au milieu socioculturel de l'individu, et d'autre part aux différences physiologiques des organes responsables de la production vocale. L'expression acoustique de ces différences peut être traduite par une variation de la fréquence fondamentale, dans l'échelle des formants (plus haute chez les femmes et les enfants que chez les hommes) et dans le timbre de la voix (richesse en

harmoniques due à la morphologie du locuteur et au mode de fermeture des cordes vocales) [11].

I.4.3 Variabilité due au matériel

Le signal de parole est porteur d'informations caractérisant le matériel utilisé lors de sa capture (ex. : microphone, combiné téléphonique), de sa transmission (ex. : lignes téléphoniques) et de son enregistrement (ex. : microphones, convertisseurs). Ces informations apparaissent le plus souvent sous la forme de déformations/dégradations du signal de parole. Ces déformations sont différentes selon le type de matériel utilisé [11].

I.5 Mode dépendant et indépendant du texte

Différentes applications basées sur la même tâche se distinguent par leur dépendance vis-à-vis du texte. Si la parole parlée n'est pas a priori, alors il s'agit du système RAL dit indépendant du texte. Au contraire, lorsque le texte prononcé par le locuteur est fixe et connu à l'avance, on parle de système dit dépendant du texte. En mode indépendant du texte, le locuteur peut prononcer n'importe quelle phrase à reconnaître. Par rapport aux systèmes indépendants du texte, les systèmes indépendants du texte offrent généralement de meilleures performances de reconnaissance, car la variabilité causée par le contenu de la langue parlée est ensuite éliminée. Cependant, dans un système basé sur une technologie de reconnaissance vocale à grand vocabulaire, les performances de reconnaissance indépendantes du mode texte ou indépendantes du mode texte sont en fait les mêmes [5]. Le niveau de dépendance du texte est classé en fonction de l'application [2]:

- Système à texte libre (ou free-text) : Les locuteurs peuvent prononcer librement ce qu'ils veulent. Dans ce mode, l'apprentissage et le test des phrases sont différents.
- Système à texte suggéré (ou text-prompted) : un texte, diffèrent à chaque session et pour chaque personne, est imposé au locuteur et déterminé par la machine. L'apprentissage et le test des phrases peuvent être différents.
- Système dépend de vocabulaire (ou vocabulary-dependent) : le locuteur prononce une séquence de mots issus d'un vocabulaire limité. Dans ce mode, l'apprentissage et le test sont réalisés sur des textes constitués à partir du même vocabulaire.

- Système personnalisé dépendants du texte (ou user-specific text dependent) : chaque locuteur a son propre mot de passe. Dans ce mode, l'apprentissage et le test sont réalisés sur le même texte.

D'évidence, la connaissance a priori de message vocal rend la tâche des systèmes RAL plus facile et les performances sont les meilleures. La reconnaissance en mode indépendant du texte nécessite plus de durée de parole que le mode dépendant du texte [2].

I.6 L'architecture d'un système RAL

Un système reconnaissance automatique du locuteur tel qu'il illustré par la figure I.6, se compose de trois processus principaux, qui sont : la paramétrisation (analyse acoustique), la modélisation et la décision. En plus, un système de RAL possède deux modes [12] :

- Apprentissage, où un modèle est estimé pour chaque locuteur « client » du système puis servira de référence pour les tâches de reconnaissance à venir.
- Test, où l'étape de reconnaissance (vérification, identification...) est effectuée. En sortie de cette phase, le système émet une réponse : une identité pour la tâche d'identification ou une décision accès/rejet pour la vérification.

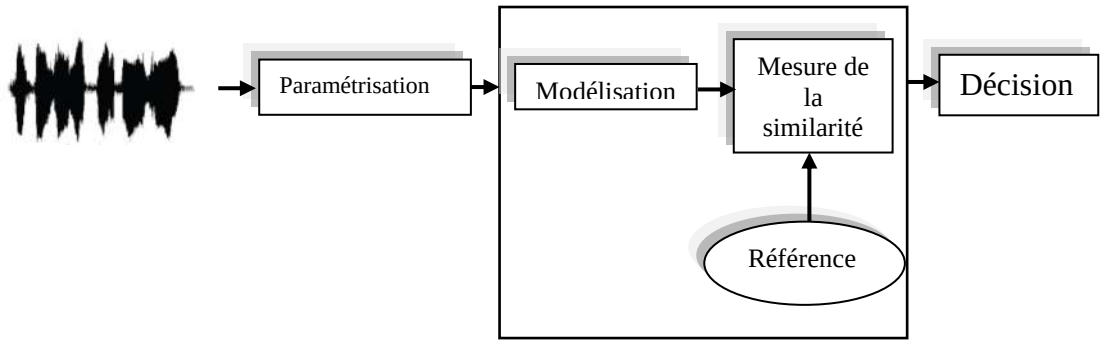


Figure I-6 : Structure générale d'un système RAL [12].

I.6.1. La paramétrisation

Dans cette étape en décompose le signal parole en trames de longueur variantes de (10 à 30ms) car le signal vocal est supposé stationnaire dans cette durée (invariance par changement d'observation). Pour chaque fenêtre on fait l'extraction des paramètres acoustiques qui peuvent fournir une représentation plus simple du signal parole. Il existe trois grandes catégories de paramètres [13]:

- **Les paramètres prosodiques:** Ces paramètres illustrent le débit, la durée et les fréquences de pauses, pitch et l'énergie [13].
- **Paramètres acoustiques:** Ces paramètres sont utilisés en RAL. Ils représentent l'influence des caractéristiques physiques de l'appareil phonatoire de chaque individu. Ces paramètres sont les GFCCs et MFCCs [14].
- **Les paramètres dynamiques:** Le vecteur des paramètres issus des paramétrisations précédentes peut être complété par le vecteur correspondant aux dérivées du premier et second ordre de ces paramètres qui sont calculées à partir de plusieurs trames adjacentes. Elles permettent d'introduire une information concernant le contexte temporel d'une trame courante [14].

I.6.2 Extraction des paramètres

Le signal de parole présente de la redondance et il est porteur de plusieurs types d'informations, ce qui justifie la recherche d'une représentation spécifiquement pertinente pour la reconnaissance. L'extraction des paramètres consiste à associer au signal de parole une série de vecteurs de paramètres acoustiques. Cette extraction débute par une mise en forme du signal de parole suivie par le calcul de coefficients acoustiques comme indiqué sur la figure I.7 suivante [18]:

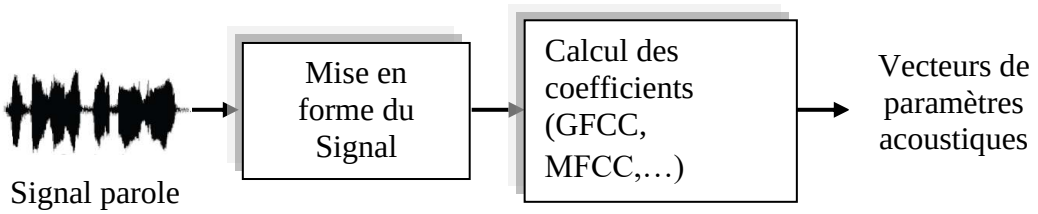


Figure I-7 : Schémas synoptiques de l'extraction de paramètres acoustiques [18].

I.6.3 Les Paramètres acoustiques

L'analyse spectrale de la parole présente des avantages au niveau de la perception, car l'oreille humaine effectue ce genre d'analyse. On introduit en traitement du signal parole les outils suivants [14]:

a) Coefficients MFCC : Les paramètres les plus généralement utilisés dans la reconnaissance de la parole est MFCC. Le calcul de cepstre est semblable à ce lui

décrit dans la section précédente. Mais, le banc de filtre mel (ou tout autre) est appliqué dans le domaine fréquentiel avant le logarithme et IDFT. Le but du banc mel est de simuler les filtres des bandes critiques du mécanisme d'audition. Les filtres sont également espacés sur une échelle mel, et habituellement ils ont une forme triangulaire.

Les sorties du filtre triangulaire sont comprimées en utilisant le logarithme et la Transformée Cosinus Discret (DCT) est appliquée :

$$c[n] = \sum_{i=1}^{M} logY(i) cos\left[\frac{\pi n}{M}(i - \frac{1}{2})\right] \quad \text{(I-1)}$$

Typiquement *c[0]* est exclu. Une des propriétés importantes des coefficients cepstrals est qu'ils sont particulièrement non corrélés. Cette propriété a quelques conséquences pratiques importantes. Par exemple, si une distance de Mahalanobis est employée en tant que métrique dans le classificateur, on aura un gain faible en employant une matrice de covariance pleine. Par contre une matrice de covariance diagonale présente un gain plus considérable [15].

b) Coefficients GFCC : Les coefficients cepstraux en fréquences gammatone, ou GFCC pour Gammatone (Frequency Cepstral Coefficients), sont substantiellement équivalents aux MFCC, mais utilisent une banque de filtre gammatone calqués sur une échelle de largeurs de bandes rectangulaires. Un filtre gammatone est un filtre linéaire décrit comme une réponse impulsionnelle étant le produit d'une distribution gamma et d'une tonalité sinusoïdale. Tout comme le MFCC, GFCC est utilisé dans les travaux de reconnaissance et transcription vocales [1].

c) Coefficients MRCG : De bonnes caractéristiques sont cruciales pour la performance de la parole renforcement.une fonctionnalité appelée MRCG [1] est proposée par Chen et al, qui est extrait de quatre cochléagrammes de différentes résolutions pour capturer à la fois local information et contexte spectrotemporel. Comme nous le savons tous, la non-linéarité auditive humaine étend les petits sons et compresse les gros sons. Le MRCG simule l'audition non-linéarité par une fonction log. La fonction de journal peut bien compresser intensité de puissance élevée, mais sur expansion très faibles signaux à l'infini. Considérant que le bruit se produit souvent dans très petite énergie, la fonction log accentue souvent le bruit et entraîne une mauvaise robustesse au bruit ; pour éviter sur expansion à très faible bruit, nous

utilisons une fonction de puissance en quatre cochléagrammes, ce qui permet de mieux simuler non-linéarité auditive humaine [1].

I.6.4 Modélisations des locuteurs

De manière à modéliser des caractéristiques qui dépendent du locuteur, nous utilisons des algorithmes capables de capturer les points communs entre différentes représentations de motifs spectraux issus du même locuteur (constituant ainsi un modèle du locuteur), tout en ayant la possibilité de s'adapter aux variations d'échelles fréquentielles et temporelles liées au signal de parole. Ces motifs peuvent être soit des segments de parole déterminés (mots, phonèmes) si nous travaillons en mode dépendant du texte, soit des segments de parole dont on ne connaît pas le contenu phonétique si l'application fonctionne en mode indépendant du texte. Ces algorithmes doivent être couplés avec une mesure qui permettra de donner une valeur de distance (ou de similitude) entre le modèle du locuteur et un motif inconnu dont on cherche à déterminer la provenance. Nous proposons ici un aperçu des méthodes déterministes (comparaison dynamique et quantification vectorielle) et statistiques (modèles à mélange de distributions gaussiennes et modèles de Markov cachés) qui fournissent les meilleurs résultats en reconnaissance vocale.

Le problème de reconnaissance du locuteur peut se formuler selon un problème de classification. Différentes approches ont été développées, néanmoins on peut les classer en trois grandes familles [2]:

- L'approche vectorielle : le locuteur est représenté par un ensemble de vecteurs de paramètres dans l'espace acoustique. Ses principales techniques sont la reconnaissance à base de DTW (Dynamic Time Warping) et par quantification vectorielle.
- L'approche statistique : consiste à représenter chaque locuteur par une densité de probabilités dans l'espace des paramètres acoustiques. Elle couvre les techniques de modélisation par les modèles de Markov cachés, par les mélanges de gaussiennes et par des mesures statistiques du second ordre.
- L'approche connexionniste : consiste principalement à modéliser les locuteurs par des réseaux de neurones.

a) L'approche connexionniste: Les réseaux de neurones ont été assez largement utilisés en reconnaissance du locuteur. Ils offrent en effet une bonne alternative au problème de la discrimination entre les locuteurs. Ces outils de classification

permettent de séparer des classes, dans un espace de représentation donné, de façon non linéaire. L'inconvénient important de l'application de cette technique en identification du locuteur est le coût important lié à l'ajout d'un nouveau locuteur dans la base de référence (ce n'est pas le cas en vérification du locuteur). On peut aussi utiliser les réseaux de neurones en les couplant à d'autres techniques, par exemple les modèles de Markov cachés, on parle alors de méthodes hybrides [17].

b) L'approche vectorielle (Reconnaissance du locuteur a base DTW) : La reconnaissance par l'alignement temporel par programmation dynamique DTW repose sur le principe que chaque mot est représenté par une prononciation de référence. Compte tenu des décalages temporels entre les différentes prononciations d'un même mot, l'algorithme met en correspondance des séquences de paramètres par distorsion temporelle (Time warping). La programmation dynamique permet d'aligner temporellement une phrase de test avec une phrase d'apprentissage ce qui signifie que c'est une technique exclusivement utilisée en mode dépendant du texte [2].

Ce type de technique a été peu à peu abandonné au profit des modèles séquentiels statistiques (comme les modèles de Markov cachés) qui sont moins "rigides" et donc plus robustes vis-à-vis de la variabilité inhérente au signal de parole.

I.6.5 Quantification vectorielle

La quantification vectorielle est une méthode non paramétrique qui permet de décrire un ensemble de données par un faible nombre de vecteurs formant un dictionnaire associé aux données. Le dictionnaire est en général calculé de telle façon que la distance moyenne entre un vecteur issu des données et son plus proche voisin dans le dictionnaire soit la plus petite possible. La quantification vectorielle est une technique de groupage qui est d'autant plus adapté que les données présentent naturellement des "points d'accumulation" autour duquel la densité de vecteurs issus des données est importante.

Pour la reconnaissance du locuteur, la mesure de similarité entre deux ensembles de vecteurs acoustiques consiste à évaluer la distance moyenne d'un des deux ensembles de vecteurs acoustiques en utilisant le dictionnaire optimisé pour l'autre ensemble de vecteurs acoustiques par quantification vectorielle.

La caractérisation de la distribution des données obtenue par la quantification vectorielle est en fait voisine de celle fournie par un modèle de mélange de distributions gaussiennes. Les performances des deux types de systèmes sont donc assez proches. Lorsque les données disponibles pour l'apprentissage sont suffisantes, il semble que le modèle de mélange de densités gaussiennes soit plus robuste.

I.7 Décision et mesure des performances

Un module de reconnaissance fournit en sortie un score. La nature de ce score varie selon les modules de reconnaissance utilisés. Il s'agit la plupart du temps d'une distance, d'une probabilité ou d'une vraisemblance. Un module de décision doit, à partir de ce score, fournir une décision qui constituera la réponse finale du système de reconnaissance. Le module de décision décrit dans le paragraphe précédent reçoit, en entrée et pour chaque test, un score. Celui-ci résulte de la comparaison entre les caractéristiques biométriques de l'utilisateur testé et la référence apprise lors de la phase d'enrôlement. Un score élevé signifiera que la probabilité pour que l'utilisateur testé corresponde à l'identité qu'il annonce est élevée et un score faible signifiera que cette probabilité est faible. La décision binaire qui constitue la sortie du module résulte de la comparaison de ce score avec un seuil défini à l'avance. Si le score est supérieur au seuil, l'utilisateur est accepté et s'il est inférieur au seuil, l'utilisateur est rejeté [12].

I.7.1 L'identification automatique du locuteur

Consiste à reconnaître un locuteur parmi un ensemble de locuteurs en comparant son identité vocale à des références connues**.** Les performances du système d'identification sont données en termes de taux d'identification correcte *Ic*ou incorrecte *Ii*, soit :

$$I_c = \frac{Nombre\ de\ tests\ correctement\ identifiés}{Nombre\ total\ de\ tentatives} \times 100$$

$$I_i = \frac{Nombre\ de\ test\ sma\ lidentifiés}{Nombre\ total\ de\ tentatives} \times 100$$

Avec$I_i + I_c = 100\%$.

Alors qu'en ensemble ouvert ce sont les faux rejets et les fausses acceptations qui sont mesurées [2].

I.7.2 Vérification automatique du locuteur

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 est un processus de comparaison entre la voix d'une personne déterminée et l'échantillon de propre voix cette méthode vérifie si la voix de la personne proclamant 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 [18]:

a) 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 ça on parle de FR la deuxième erreur survient lorsqu'un imposteur est accepté alors que l'identité annoncée n'est pas la sienne Ce cas est appelé FA 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 [2] :

$$FAR = \frac{nombre\ d'imposteur\ acceptés}{nombre\ d'imposteur\ totale} \times 100$$

$$FRR = \frac{nombre\ de\ client\ rejetées}{nombre\ totale\ du\ client} \times 100$$

I.7.3 Taux d'Égal Erreur EER

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

I.8 les sources d'erreurs

Le signal acoustique de la parole présente des caractéristiques qui rendent complexe son interprétation. L'information portée par se signale peut être analysée de bien des façons et à plusieurs niveaux (acoustique, phonologique, syntaxique, sémantique et pragmatique). Ce qui rend la tâche de traitement de la parole complexe. Plus particulièrement, on a u que la variabilité interlocuteur est l'essence même de la reconnaissance. Il existe, cependant, plusieurs facteurs qui peuvent augmenter la variabilité intra-locuteur par exemple :

- L'état pathologique de locuteur (maladie, émotions…).
- Vieillissement (la voix d'une personne change au fur et à mesure de son vieillissement).
- Facteurs socioculturels (le locuteur peut changer d'accent).
- Locuteurs non coopératifs (notamment dans des applications judiciaires).
- Conditions de prise de son (bruit, ambiant …)

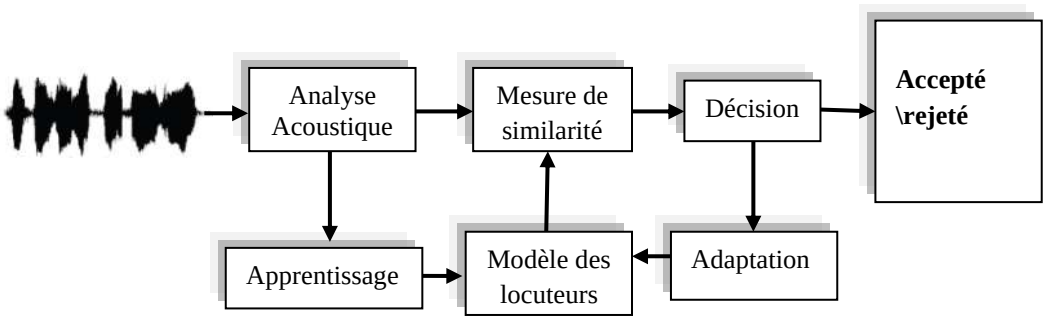


Figure I-8: schéma typique d'un système de vérification de locuteur.

I .9 Conclusions

Dans ce chapitre, nous avons introduit le principe de la reconnaissance automatique du locuteur ainsi que les différentes étapes d'un système de RAL. Le système de reconnaissance de locuteur procède généralement en trois étapes : l'analyse acoustique du signal parole, la modélisation du locuteur et l'étape de décision. Dans notre travail nous avons utilisé pour l'étape analyse acoustique les coefficients MFCCs, GFCCs et MRCGs. Pour la modélisation de locuteur, l'approche

statistique GMM, utilisée en mode indépendant de texte et enfin l'étape de décision est basée sur le processus d'identification de locuteur.

Chapitre II :

Extraction des paramètres multirésolution MRCG

II.1 Introduction

La parole en tant que moyen efficace de communication entre humains a donné lieu à plusieurs projets de recherche dans le domaine de traitement de la parole. Le signal de la parole, de par sa nature pseudo-aléatoire et très variable. Par conséquent, ce signal nécessite un prétraitement appelé l'étape du paramétrage ou bien l'extraction des vecteurs de caractéristiques. L'extraction des paramètres est une étape importante pour le processus de la reconnaissance du locuteur. Dans ce chapitre, nous présentons une nouvelle technique de paramétrisation appelée cochléagramme multi-résolution (MRCG), cette technique est fondée sur le banc de filtre gammatone.

II.2 Principe de multirésolution

La technique de multirésolution a été spécialement conçue pour la parole séparation, qui a été évaluée avec une série de caractéristiques acoustiques existantes systématiquement et a finalement obtenu les meilleures performances. Les fonctionnalités cepstrales de sous-bande multi-résolution s'efforcent indices discriminants dans des régions localisées du domaine spectral en complétant les fonctionnalités cepstrales à pleine bande passante avec une sous-bande traits cepstraux dérivés de plusieurs niveaux de sous-bande décomposition. Des vecteurs de caractéristiques multi-irrésolutions, formées par concaténation des caractéristiques cepstrales de sous-bande dans un vecteur de caractéristiques, sont montrés pour donner de meilleures performances que MFCC conventionnel pour la reconnaissance de phonèmes. Stratégies possibles pour la recombinaison de scores de reconnaissance de la sous-bande multi-résolution indépendante des modèles sont explorées. En exploitant les variations de sous-bandes dans rapport signal sur bruit pour la recombinaison pondérée linéairement du log probabilités de vraisemblance, nous avons obtenu un phonème amélioré performances de reconnaissance dans le bruit à large bande par rapport au MFCC fonctionnalités. C'est un avantage par rapport à une approche purement sous-bande utilisant une recombinaison non linéaire qui n'est robuste que pour rétrécir bruit de bande [19].

La multi-résolution les contextes d'une unité T-F peuvent potentiellement aider à classer le T-F l'unité comme cible-dominante ou interférence-dominante. La

conception des caractéristiques MRCG est destinée à incarner le spectrotemporel contexte systématiquement, afin de faciliter l'estimation de l'IBM [20].

II.3 I-vecteur

Inspiré de la modélisation classique de l'Analyse Factorielle Conjointe (JFA) basée sur l'analyse factorielle et quelques travaux suivants, Dehak et al. Ont récemment proposé l'état de technologie de reconnaissance de haut-parleur basée sur I-vecteur, dérivée de la technologie GMM-UBM [21]. Cette méthode de modélisation est basée sur la définition d'un espace unique, au lieu de deux espaces séparés dans JFA. Ce nouvel espace, qui porte le nom espace de variabilité total, contient simultanément les variabilités du haut-parleur et du canal. Par conséquent, lors de l'extraction d'I-vecteur, il n'y a pas de distinction entre les effets de haut-parleur et les effets de canal dans l'espace du superviseur GMM, car les expériences ont démontré les facteurs de canal du JFA, qui ne modélisent normalement que le canal effets, contiennent également des informations sur le haut-parleur [22].

Dans le modèle i-vecteur, le nouveau superviseur GMM dépendant du haut-parleur et du canal défini peut s'écrire comme suit: $M = m + Tw$ où m est un composant indépendant à la fois du haut-parleur et du canal (utilisant généralement UBM superviseur), T est une matrice rectangulaire de bas rang représentant les directions primaires de variabilité et w est un vecteur aléatoire ayant une distribution normale standard $N(0, I)$. Les composants w sont appelés facteurs totaux, également appelés I-vecteur pour court. On suppose que M soit normalement distribué avec le vecteur moyen m et la covariance matrice TTt. Le processus de formation de la matrice T est exactement le même que celui de la formation du Eigenvoice V matrix dans JFA, sauf qu'il n'est pas nécessaire de considérer l'étiquette des haut-parleurs. Pour cette raison, la modélisation i-vecteur peut être considérée comme un simple facteur analyse, qui nous permet de projeter un superviseur GMM de grande dimension dérivé de l'énoncé de la parole au vecteur i de faible dimension dans l'espace de variabilité totale. Comme l'espace T contient à la fois la session de l'orateur et la variabilité du canal, les vecteurs i extraits nécessitent des technologies de compensation supplémentaires pour atténuer les effets. Comparaison d'un certain nombre d'approches existantes telles que la normalisation de la covariance intercalasse (WCCN), analyse discriminante linéaire (LDA) et attribut de nuisance Projection (NAP), les expériences

de montrent que les meilleurs résultats sont obtenus avec la combinaison LDA et WCCN suivie d'un score de distance cosinus (CDS) classificateur ou autre classificateur comme PLDA. Pour comparer assez au résultat, nous utilisons la même configuration que dans. La méthode que nous avons utilisée sera décrite dans le reste de cette section [23].

II.4 Analyse multirésolution des paramètres

La reconnaissance automatique du locuteur peut atteindre un niveau élevé de performances dans des conditions d'entraînement et de test adaptées. Cependant, de telles performances chutent de manière significative dans des conditions de bruit incompatibles. Des recherches récentes indiquent qu'une nouvelle fonction de haut-parleur, les coefficients GFCC, présente une robustesse au bruit. À acquérir une compréhension approfondie de la robustesse intrinsèque de GFCC par rapport au MFCC sur tout dans l'identification automatique du locuteur (IAL).

II.4.1 Différences entre MFCC et GFCC

D'une manière générale, il existe deux différences majeures. La plus évidente est l'échelle de fréquence GFCC, basé sur l'équivalente échelle de bande passante rectangulaire (ERB), a une résolution plus fine à basse fréquence que MFCC. L'autre est l'étape de rectification non linéaire avant le DCT. MFCC utilise un journal tandis que GFCC utilise une racine cubique. Les deux ont été utilisés dans la littérature. De plus, l'opération log transforme la convolution entre la source d'excitation et le conduit vocal (filtre) en plus dans le domaine spectral. En Outre ces deux différences majeures, il existe d'autres différences notables qui sont résumées dans le Tableau II.1 suivant :

catégories	MFCC	GFCC
Préaccentuation	oui	non
Nombre de bandes de fréquences	26	64
Levage cepstral	oui	non
Échelle de fréquence	Mel	ERB
Rectification non linéaire	logarithmique	Racine cubique
Invariant à l'échelle (w/o 0th coefficient)	oui	non
Représentation intermédiaire T-F	Spectre Mel	Variante de Cochléagramme

Tableau II-1: Différences en MFCC et GFCC Dérivation.

II.5 Paramètres GFCC

Il informe sur le taux de changement dans différents spectres bandes, utilisées pour déterminer la fréquence fondamentale de l'homme parole et d'analyser le retour du signal radar. Comme la MFCC, les vecteurs de caractéristiques dans cette technique sont calculés à partir des spectres de séries de trames. Après avoir pris le logarithme à chaque sortie de filtre et transformation en cosinus discrète appliquée à sortie du filtre, algorithme GFCC utilisé pour l'extraction des caractéristiques de signaux vocaux, par ex : des émotions telles que la tristesse, la joie, la peur, etc. ainsi pour améliore la validité de la reconnaissance vocale dans le bruit.

II.5.1 Filtre Gammatone

Le signal vocal est un signal typique variant dans le temps. Sa décomposition temps-fréquence centrée sur la variation dans le temps caractéristique spectrale des composantes du signal vocal, qui décompose un signal vocal unidimensionnel en un signal bidimensionnel afin de révéler la relation entre ces composantes fréquentielles obtenues et temps.

Le banc de filtre gammatone est une représentation physiologique de la partie interne du système auditif, c'est-à-dire la cochlée. La perception du son commence par la conversion du son acoustique en impulsions électriques grâce aux cellules ciliées disposées le long de la membrane basilaire. Chacune de ces cellules correspond à une fréquence. Lorsque la cochlée était stimulée par une onde sonore, la membrane basilaire était déplacée suivant une onde qui se propageait à partir de la base de la cochlée. Figure II-1 représente la réponse de la membrane basilaire aux stimulations acoustiques. L'amplitude de cette onde augmente au fur et à mesure qu'elle se propage, passe par un maximum et décroit ensuite très rapidement.

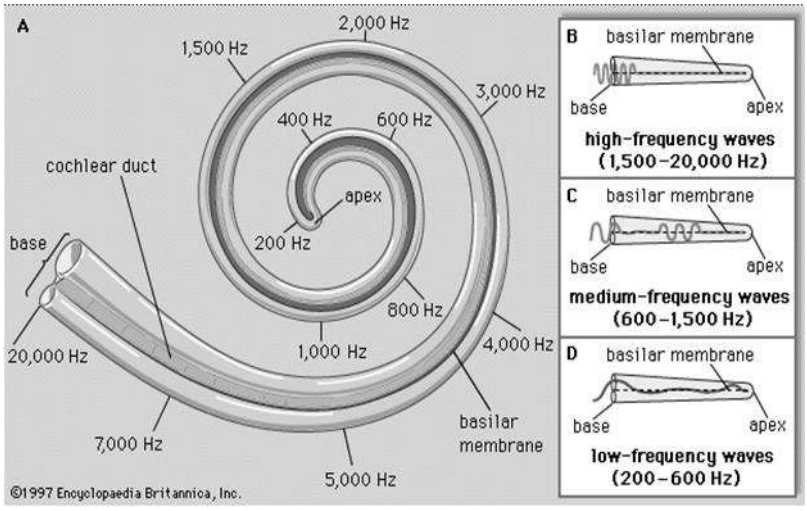


Figure II-1 : Organisation tonotopique cochléaire (A) et réponse de la membrane basilaire aux stimulations acoustiques (B, C, D).

Les sons de haute fréquence font vibrer uniquement la partie apicale de la cochlée (B) alors que les sons de basse fréquence induisent une vibration jusqu'à l'apex (D)

La réponse impulsionnelle du le filtre gammatone est le suivant :

$$g(t, f_c) = t^{l-1} e^{-2\pi B f_c t} \cos(2\pi f_c t + \emptyset)\, t \geq 0 \qquad \text{(II-1)}$$

Où *t* représente l'indice de l'échantillon, *fc* est la fréquence centrale pour le canal *cth* qui varie de 50 à 8000 Hz et ∅ la phase initiale du filtre gammatone. Donc pour simplifier le modèle, est mis à 0. l est le filtre ordre. Un grand nombre d'expériences montrent que, lorsque l = 4, le filtre peut bien simuler le filtre cochléaire caractéristique.

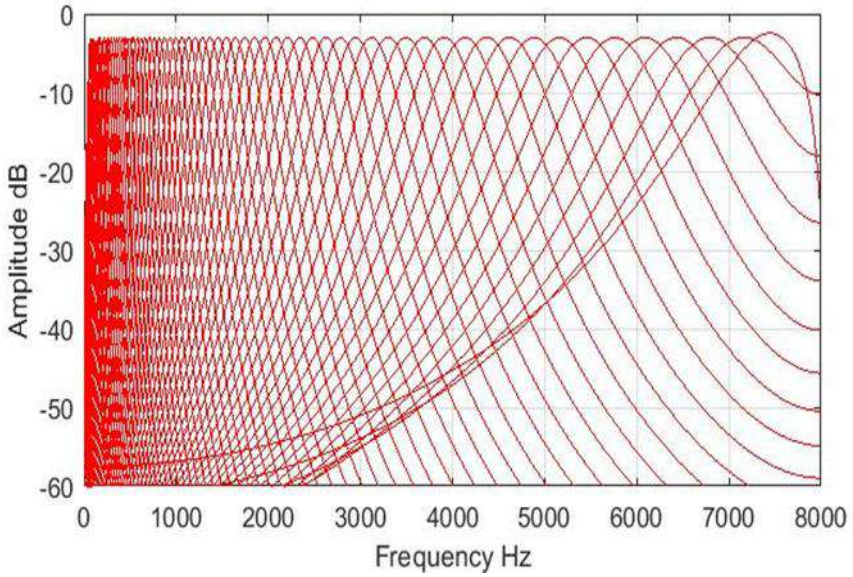


Figure II-1 : Réponse en fréquence du banc de filtre Gammatone à 64 canaux.

Les fréquences centrales sont uniformément espacées sur l'échelle ERB dans [200 3400] Hz gamme. Un taux d'échantillonnage de 8000 Hz.

Donc l'est égal à 4 dans cet article. L'échantillonnage le taux de données expérimentales est fixé à 16 kHz. Afin de mieux refléter les caractéristiques harmoniques de la parole dans chaque signal de sous-bande, le nombre de filtres est déterminé à 64 dans l'algorithme proposé. *B(fc)* est la bande passante de chaque canal de fréquence, qui est défini comme:

$$B(f_c) = b.ERB(f_c) \tag{II-2}$$

Où *b* est un facteur d'atténuation, la meilleure performance de filtre peut être obtenu lorsque *b* est égal à 1,019, donc *b* est défini à 1,019. *ERB(fc)* représente l'équivalent bande passante rectangulaire (bande passante rectangulaire équivalente (ERB)), et la relation entre l'équivalent largeur de bande rectangulaire et la fréquence centrale *fc* peut être décrit par :

$$ERB(f_c) = 24.7(\frac{4.37f_c}{1000} + 1) \tag{II-3}$$

Où les coefficients 24,7 et 4,37 sont les valeurs empiriques valeurs obtenues dans l'expérience. L'expression du signal d'entrée peut être exprimée par :

$$x(t) = s(t) + n(t) \tag{II-4}$$

Où *x(t)* représente un signal vocal bruyant, *s(t)* représente signal propre, et *n(t)* représente le signal de bruit. *x(t)* est décomposé en 64 signaux de sous-bande *G(t, fc)* par Filtres gamma à 64 canaux, comme indiqué dans la formule (II-5) :

$$G(t, f_c) = g(t, f_c).U(t).x(t) \quad \text{(II-5)}$$

Où *U(t)* est la fonction échelon unitaire. Ensuite, chaque signal de sous-bande est divisé en temps-fréquence unités (T-F) avec une trame de 20 ms avec une trame de 10 ms décalage de trame. Une unité T-F correspond à une petite auditive unité de la parole bruyante. Il est défini comme :

$$y_i(t, f_c) = w(t).G(\big((i-1).inc + t\big), f_c) \quad \text{(II-6)}$$

Où w(t) est une fonction fenêtre. Comparé à la fenêtre rectangulaire, la fenêtre de Hamming peut mieux refléter la caractéristique de fréquence du signal vocal, donc la fenêtre de Hamming est choisie dans cet article. *yi(t, fc)* est l'unité T-F de sous-bande pour le canal cth à la période je; inc est un décalage de cadre. La puissance du filtre auditif (cochléagramme) de chacun L'unité T-F *CG (i, fc)* est calculée par :

$$CG(i, f_c) = \sum_{t=0}^{L-1} y_i^2(t, f_c) \quad \text{(II-7)}$$

II.6 Cochléagramme multi-résolution

En plus des fonctionnalités existantes, nous proposons une nouvelle fonctionnalité acoustique, qui encode les distributions de puissance multirésolution dans la représentation temps-fréquence d'un signal. Nous combinons quatre cochléagrammesà différentes résolutions pour construire la fonction MRCG. Un cochléagramme haute résolutioncapture les informations locales tandis que trois cochléagrammes basse résolution capturent contextes spectrotemporel à différentes échelles.

II.6.1 Construction de MRCG

La construction de MRCG est basée sur la représentation du cochléagramme, qui est largement utilisé dans la littérature CASA [24]. Pour calculer le cochléagramme, nous passons d'abord un signal d'entrée à une banque de filtres gamma, où la réponse impulsionnelle d'un filtre gamma particulier est [25],

$$gfc(t) = t^{N-1} exp[-2\pi tb(fc)]cos(2\pi fct)u(t) \quad \text{(II-8)}$$

Où *fc* désigne la fréquence centrale, *N* l'ordre du filtre et *u(t)* la fonction échelon. La fonction *b(fc)* décide de la bande passante donnée *fc*. Imiter l'audition humaine filtres, les fréquences centrales *fc* sont uniformément espacées sur la rectangle équivalente échelle de bande passante (ERBN). La relation entre *b(fc)* et *fc* est montrée dans l'équation :

$$b(fc) = 1.019 * ERBN(fc) = 1:019 * 24:7 * \left(4:37 * \frac{fc}{1000} + 1\right) \text{(II-9)}$$

La bande passante *b(fc)* augmente à mesure que *fc* augmente, conduisant à des résolutions plus élevées à faibles fréquences et des résolutions inférieures aux hautes fréquences sur l'échelle de fréquence linéaire.

Après avoir obtenu les signaux de réponse de la banque de filtres gammatone, nous divisons chaque réponse signal en trames de 20 ms avec un décalage de trame de 10 ms. On dérive le cochléagramme par calcul de la puissance de chaque trame à chaque canal [24].

Chaque unité T-F du cochléagramme ne contient que des informations locales, qui ne peuvent pas être suffisantes pour estimer le masque. Pour compenser cela, la fonction MRCG fournit des informations contextuelles en incluant la distribution électrique dans le quartier de chaque unité T-F [26]. La fonction MRCG est similaire à la fonction GFB dans le sens que les deux sont conçus pour coder le contexte spectrotemporel systématiquement (voir aussi les étapes de calcul du MRCG sont décrites comme suit) [1]:

1. Étant donné un mélange d'entrées, calculez le premier cochléagramme à 64 canaux, CG1. Un journal l'opération est appliqué à chaque unité T-F.

2. De même, calculez CG2 avec une longueur de trame de 200 ms et un décalage de trame de 10m / s.

3. CG3 est dérivé en faisant la moyenne de CG1 sur une fenêtre carrée de 11 canaux de fréquence et 11 trames temporelles centrées sur une unité T-F donnée. Si la fenêtre va au-delà du cochléagramme donné, les unités extérieures prennent la valeur zéro (c'est-à-dire zéro rembourrage).

4. CG4 est calculé de la même manière que CG3, sauf qu'une fenêtre carrée de 23 est utilisée.

5. Concaténez CG1-4 pour obtenir la fonction MRCG, qui a 64 4 dimensions pour chaque tranche de temps. Notez que, bien que soit défini à l'aide d'un cochléagramme à 32 canaux, les fonctionnalités peuvent être extraites d'un cochléagramme de taille différente avons trouvé que les fonctionnalités 64 canaux extraites à l'étape 1 fonctionnent un peu mieux que 32 canaux fonctionnalités. De plus, l'utilisation d'un remplissage nul à l'étape 3 pour les unités T-F extérieures entraîne un léger meilleur résultat que la simple moyenne des unités à l'intérieur d'une fenêtre.

II.6.2 Analyse du MRCG

Dans les paramètres MRCG, CG1 contient les informations locales intégrées dans un cochléagramme tandis que CG2-4 fournit des contextes à grain ne et à grain grossier. Les paramètres utilisés dans la construction de MRCG sont décidés expérimentalement comme suit.

1. D'abord, la longueur de trame de CG1 est choisie pour correspondre à la longueur de trame de l'IBM.
2. Ensuite, nous fixons CG1 et déterminons CG2 en l'étendant à différentes longueurs de trame pour sélectionner la meilleure longueur. De même, nous décidons de la taille de la fenêtre de moyennage pour CG3, puis pour CG4.
3. Après avoir obtenu CG1-4, nous constatons que l'ajout de plus de cochléagrammes ne fournit pas d'autres améliorations des performances.

La figure II-3 illustre les T-F contextes sur les résultats de la séparation, l'ajout de CG2-4 de manière cohérente améliore les résultats pour le bruit de babillage à 5 dB SNR. Des tendances similaires sont observées pour les autres bruits. Une visualisation de la caractéristique MRCG, où les tracés de gauche caractéristiques extraites d'un mélange de babillage à 5 dB SNR et la droite du discours propre correspondant. Comme le montre la figure II-3, CG1 est le cochléagramme

régulier, CG2 capture le contexte temporel, CG3 code un contexte spectro temporel relativement petite CG4 code un contexte spectro temporel relativement grand. La justification large derrière MRCG est qu'une unité T-F est plus susceptible d'être dominante si elle réside dans un groupe de nombreuses unités T-F à dominante vocale. En d'autres termes, un discours dominant l'unité T-F n'apparaît probablement pas seule dans un cochléagramme [15].

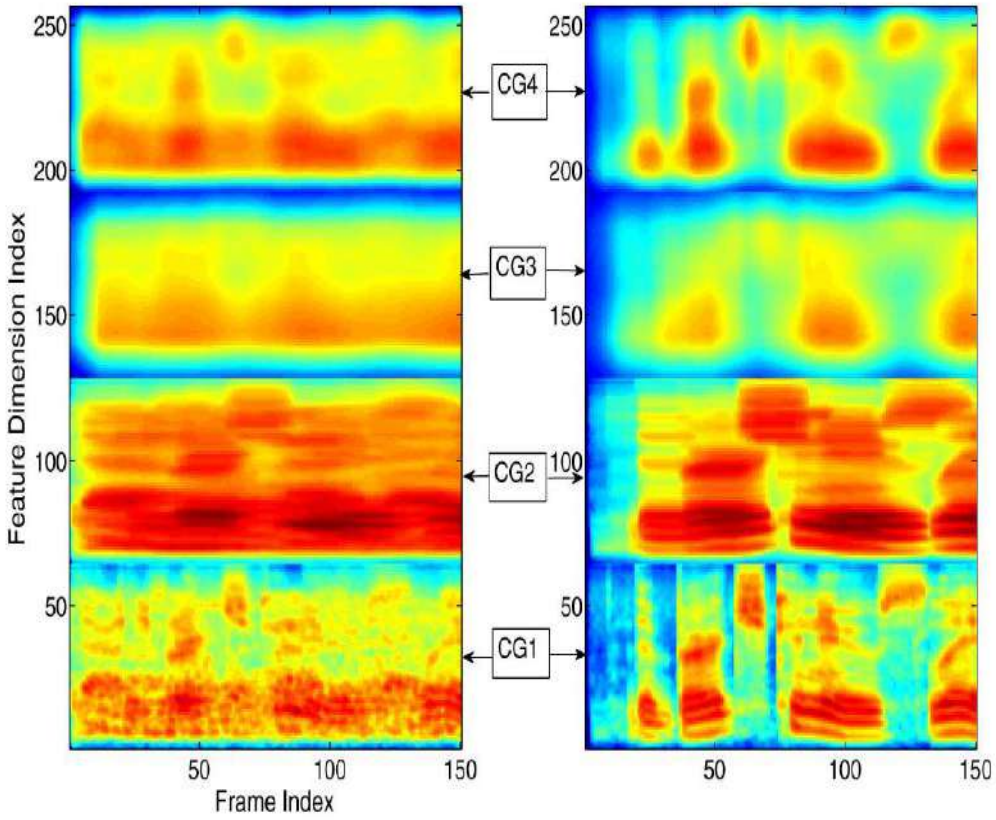


Figure II-3 : Visualisation de la fonction MRCG [15].

Le côté gauche montre les caractéristiques MRCG extraites d'un mélange, tandis que le côté droit montre les caractéristiques MRCG extraites de discours propre prémélangé. Dans CG2-4, les motifs de caractéristiques du mélange ressemblent à ceux de discours propre dans une certaine mesure, indiquant que la fonction MRCG pourrait partiellement retenir modèles spectrotemporel de la parole en présence de bruit.

II.7 Conclusion

Dans ce chapitre nous avons présenté une nouvelle paramètres de fréquence temporelle multi-résolution (abrégée MRTF) méthode d'extraction, obtenue en effectuant une opération DCT 2D à plusieurs échelles sur la matrice de blocs de la matrice du spectrogramme temporel et en combinant et en sélectionnant éléments transformés multi-échelles. Par rapport à paramètres MFCC traditionnelle, la méthode proposée peut faire un meilleur usage des informations temps-fréquence multi-résolution. Ces paramètres exploiteront dans chapitre suivant pour évaluation des performances du système de vérification du locuteur en environnement mobile réel.

Chapitre III :

Simulations et Résultats

III.1 Introduction

Dans ce chapitre, nous allons présenter les résultats obtenus après l'évaluation de notre système de vérification du locuteur. Notre objectif étant d'étudier la performance du système basé sur l'analyse multi-résolution des paramètres acoustiques en communication mobile dans un environnement bruité. Les protocoles de développement et d'évaluation de cette technique de para-métrisation sont décrits dans cette section. Ils mettent en jeu des modules d'extraction des paramètres acoustiques à savoir MFCCs, GFCCs et MRCGs.

III.2 Logiciels et langages utilisés dans l'application

Nous avons utilisé un logiciel MATLAB (matrix laboratory) sous Windows 7, Matlab est un langage de programmation émulé par un environnement de développement, il est utilisé à des fins de calcul numérique et commercialisé par la société américaine the Math Works. Ce langage est basé essentiellement sur le calcul matriciel à syntaxe simple, avec des fonctionnalités mathématiques spécialisées et graphiques étendues.

III.3 Description des bases de données utilisées

La simulation du système de reconnaissance a été effectuée en utilisant les bases de données NIST-2008 et la base de bruit standard NOISEX-92.

III.3.1La base de données NIST 2008

L'évaluation de système de reconnaissance du locuteur fait une partie des séries continues d'évaluations annuelles menées par le NIST (National Institute of Standards and Technology).En effet, depuis 1996 une série d'évaluations SRE (Speaker Recognition Evaluation), dans le but de promouvoir la recherche en reconnaissance automatique du locuteur. NIST-SRE 2000, qui était d´éditée à la vérification du locuteur. D'autres travaux se sont attaqués àl'identification et l'indexation des locuteurs. La base de données contient des fichiers audio (hommes et femmes) sous format (WAV.file), avec une fréquence d'échantillonnage égale à 8 KHz. La durée des fichiers varie de 10 s À trois minutes. Le découpage est le suivant: 3 enregistrements servent à l'apprentissage des modèles et 1 à la reconnaissance

(tests). Le répertoire « WAV » est divisé en 2 sous répertoire « Apprentissages » pour modèles et « test» pour la reconnaissance (tests).Nous avons pris 200 locuteurs à partir de cette base. La base de données mobile a été obtenue par passer la base de données test propre à travers le codec standard AMRWB.

III.3.2 la base de données NOISEX-92

Le but de ce corpus est de fournir un ensemble de bruits artificiels pour évaluer les performances des différents systèmes de reconnaissance de la parole et locuteur dans un l'environnement bruité. Le corpus Noisex-92 a été développé en 1992 pour l'étude de la perception et par l'équipe de recherche sur la parole. Le corpus Noisex-92 contient de 8 bruits, ces fichiers de bruits de sont enregistrés sous format 'wav' avec une fréquence d'échantillonnage de 20 kHz et quantifié à 16 bits. La durée de chaque fichier est 255 secondes. Le Tableau III.1 contient une description des différents types de bruit :

Type de bruit	**Description**
Babble	Bruit de la foule avec de 100 personnes dans un restaurant
Buccaneer1	Bruit de buccaneer à 450 noeuds à 300 pieds
Buccaneer2	Bruit de boucaner à 190 noeuds à 1000 pieds
Car	Bruit de voiture à 120km/h
Deroyerops	Bruit de destroyer
Deroyerengine	Bruit de destroyer
F16	Bruit d'un chasseur F16 biplace à 500 noeuds et 300-600 pieds en place copilote
Factory1	Bruit d'une usine : bruits de soudures électriques lors de l'assemblage du bas de caisse

Tableau III-1 : Description des différents types de bruits du corpus Noisex-92.

III.4 Étapes de la réalisation des systèmes de vérification du locuteur

Différentes étapes sont nécessaires à la construction et à la validation du système:

- Créer l'ensemble d'apprentissages: la base de données que nous avons utilisée est constituée de 1200 fichiers, prononcés par 400 locuteurs de différents sexes qui répètent la même phrase 3 fois.
- Effectuer l'analyse acoustique: l'ensemble des fichiers de la base de données ont été convertis en une série de vecteurs (MFCC, GFCC et MRCG).
- Créer un modèle PLDA Gaussien avec des I vecteurs de développement.
- Calculer le score entre les modèles d'I vecteurs et le test.
- Faire la reconnaissance et évaluer la performance des systèmes de vérificationdu locuteur par un taux EER.

III.5 la structure générale de Système de vérification du locuteur

La structure générale de notre système est représentée par la figure III-1.

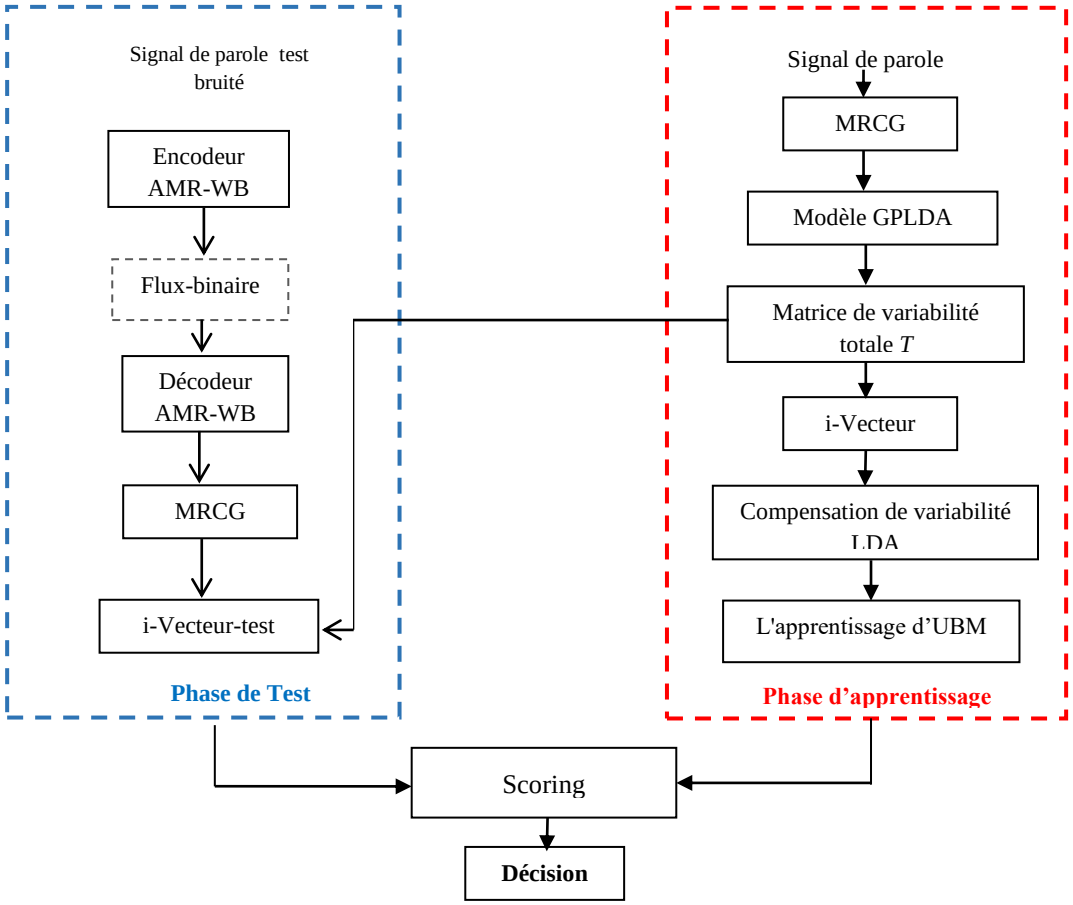


Figure III-1: Structure générale d'un système de vérification du locuteur.

III.6 Extraction des paramètres

Nous présentons en bref les différents algorithmes implémentés pour extraction les paramètres robustes pour la tâche de vérification du locuteur. Comme mentionnée précédemment, trois paramètres ont été considérés : MFCC comme notre référence, GFCC et MRCG. Ces paramètres sont extraits toutes les 10ms, sur des fenêtres de 25 ms. La fenêtre de pondération utilisée est la fenêtre de Hamming, une analyse de banc de filtres auditifs a été appliquée sur les spectres obtenus. En particulier, dans le cas du MFCC, un ensemble de 24 filtres triangulaires à l'échelle de Mel a été utilisé, tandis que pour les GFCC et les MRCG, un banc de 64 filtres Gammatones dont les fréquences centrales sont linéairement espacées sur l'échelle ERB a été appliqué. Les coefficients Cepstraux C1 à C20 ont été retenus pour les MFCC, le GFCC et le MRCG.

III.7 Evaluation des résultats

Pour évaluer notre système de reconnaissance, nous avons structuré la base La base d'apprentissage et La base de test comme suivant:

- **La base d'apprentissage:** elle est composée de 1200 fichiers vocaux prononcés par 400 locuteurs comprenant les deux sexes. Chaque locuteur répète la phrase 3 fois.
- **La base de test:** elle est composée de 1200 fichiers vocaux prononcés par 600 locuteurs, comprenant les deux sexes chaque locuteur répète la phrase 2 fois.

.

III.8 Expérience

Pour évaluer la robustesse des systèmes de reconnaissance du locuteur à base de coefficients MFCC, GFCC et MRCG. Nous avons réalisé des expériences de reconnaissance du locuteur pour trois types de bruits (environnement) qui sont respectivement: bruit des usines, foule et bruit de voiture pour l'entrainement des modèles acoustiques a été réalisé avec la base de données NIST-2008 non bruitée, le bruit a été seulement ajouté pour la base de test :(Babble : le bruit de la foule, car : le bruit de la voiture, factory : le bruit des usines) issus de la base de données de bruits standards NOISE, en utilisant les niveaux SNR suivants : 0dB, 5dB, 10dB, -5dB.

Nous avons validé notre système de reconnaissance du locuteur en utilisant comme entrée les coefficients MFCC, GFCC et MRCG, cette expérience concerne l'environnement mobile bruité avec trois types de bruit tel que le bruit de la foule , bruit de voiture et bruit des usines, nous avons varié le niveau SNR : -5dB, 0dB, 5dB, 10dB, l'entrainement des modèles acoustiques a été réalisé avec la base de données NIST non bruitée, le bruit a été ajouté pour le corpus de test.

Pour les bruits d'environnement, la qualité d'un signal est souvent représentée par le rapport de l'énergie du signal divisé par l'énergie du bruit, appelée "Rapport signal/bruit" (SNR en anglais).

$$s(t) = x(t) + n(t) \quad \text{(III-1)}$$

Le rapport signal sur bruit est défini par :

$$RSB(dB) = \frac{E_S}{E_n} \quad \text{(III-2)}$$

E_S : représente l'énergie moyenne du signal *x(t)*.
E_n *:* représente l'énergie moyenne du bruit *n(t)*.

III.8.1 Effet du bruit de foule

Le but dans cette partie nous avons utiliséle niveau de SNR : -5dB, 0dB, 5dB et 10dB appliqué sur les trois techniques des para-métrisation tels que : MFCC, GFCC et MRCG dans un système bruité, le type de bruit d'environnement dans cette expérience est bruit de foule :

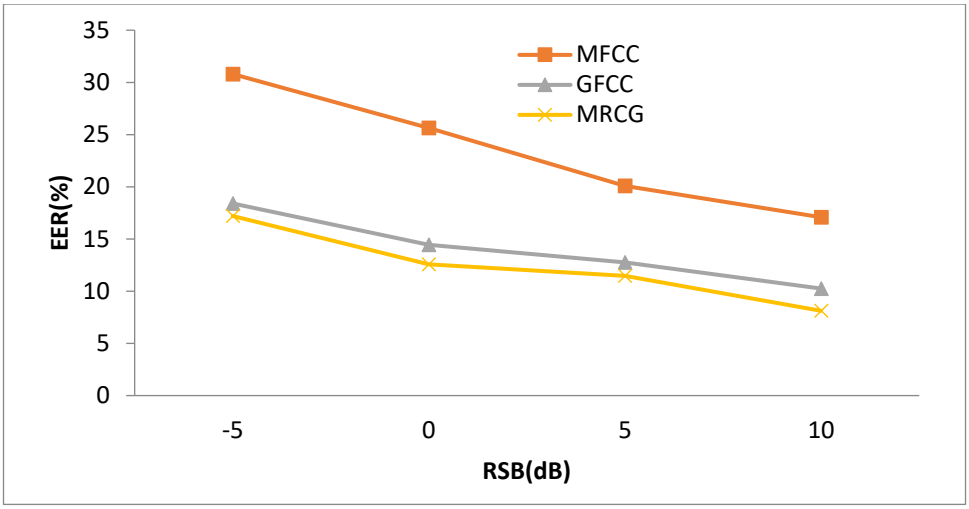


Figure III-2: Comparaison des taux d'erreurs (MFCC, GFCC et MRCG avec les différents niveaux bruit de foule.

III.8.2 Effet de bruit des voitures

Cette expérience a été consacrée à l'étude de performance des trois paramètres MFCC, GFCC et MRCG dans l'environnement de bruit de voiture en utilisant le niveau de SNR : -5dB, 0dB, 5dB et 10dB.

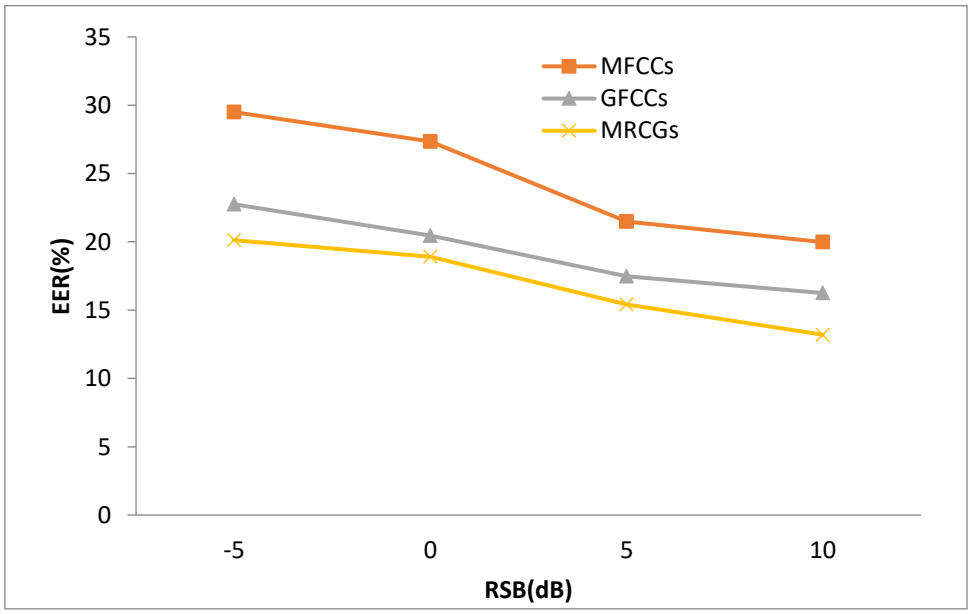


Figure III-3: Comparaison des taux d'erreurs (MFCC, GFCC et MRCG) avec les différents niveaux bruit des voitures.

III.8.3 Effet de bruit d'usine

Cette dernière expérience le bruit d'environnement utilisé c'était le bruit d'usine, et comme les deux expériences précédentes le niveau de SNR : -5dB, 0dB, 5dB et 10dB pour voir la performance des paramètres MFCC, GFCC et MRCG.

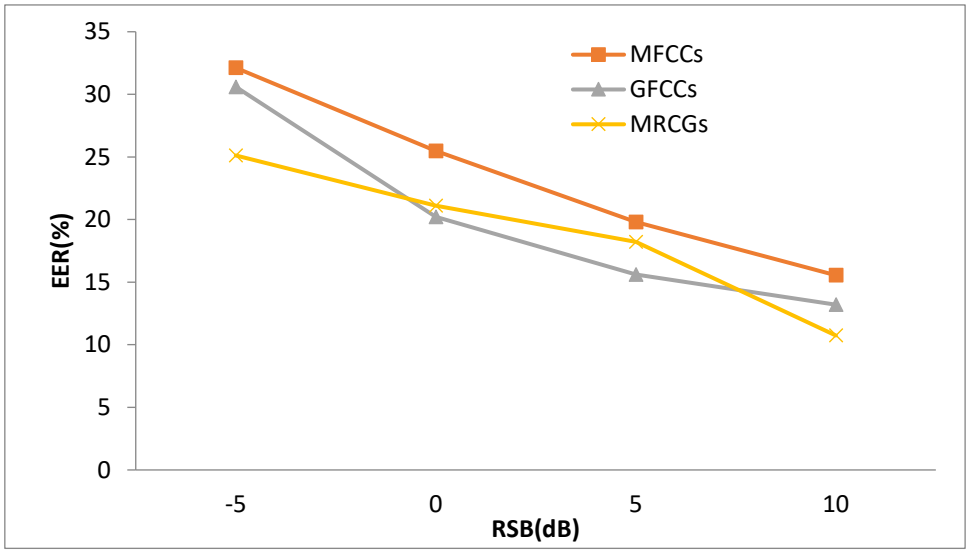


Figure III-4: Comparaison des taux d'erreurs (MFCC, GFCC et MRCG) avec les différents niveaux bruit d'usine.

À partir de ces représentations nous constatons que le taux d'erreurs obtenu par le système de RAL à base de MRCG est plus élevé que celui obtenu à base de MFCC et GFCC dans les trois types du bruit, ce qui signifie que le système de RAL à base de coefficients MRCG donne un meilleur résultat que celui à base des deux autres coefficients.

Les résultats obtenus à partir de cette expérience montrent que les valeurs des taux d'erreurs obtenus en bruit d'environnement (la voiture) sont sensiblement équivalentes que ceux obtenus en bruit d'environnement.

Nous constatons aussi que les coefficients MRCGsdonnent de meilleurs résultats des taux d'erreurs par rapport aux coefficients MFCC et GFCC.

III.9 Conclusion

Dans ce chapitre, nous avons présenté les résultats obtenus pour les différents bruits d'environnement mobile, nous avons réalisé deux systèmes de reconnaissance automatique du locuteur basé sur la méthode de para-métrisation MFCC, GFCC et MRCG. D'après les résultats obtenus, le taux d'erreur en utilisant les coefficients MRCG est nettement inférieur aux taux d'erreurs basés sur les coefficients MFCC et GFCC dans tous les cas. Au regard des résultats obtenus, nous avons conclu que le système de RAL à base de MRCG est le plus robuste.

Conclusion générale

Conclusion générale

L'objectif de notre travail était d'étudier la performance de système reconnaissance automatique du locuteur dans un environnement mobile à base d'analyse multi-résolution des paramètres acoustiques, la performance du RAL est étroitement liée à la qualité du signal parole, les signaux transmis par le réseau mobile montrent une dégradation plus importante à cause de l'effet du codage sur le signale parole, canal de transmission et les bruits d'environnement restent encore un grand problème dans une communication mobile, puisque bruits viennent perturber le signal parole.

Pour réaliser le travail dans un environnement mobile, nous avons utilisé une base de données NIST 2008 issue du codeur AMRWB. Nous avons étudié l'influence du bruit sur la performance du système de reconnaissance du locuteur basé sur l'espace i-vecteurs. Nous avons utilisé comme vecteurs (d'entrée) pour le système de reconnaissance les paramètres : MFCC, GFCC et MRCG. Nous avons simulé l'environnement mobile bruité par une base de données de bruits standards. Les expériences que nous avons effectuée sur les différents types de bruits (foule, bruit de voiture, bruit d'usine). Les résultats obtenus en utilisant les coefficients MRCG montrent que le système de reconnaissance du locuteur est plus performant que le système de reconnaissance avec des paramètres MFCCs et GFCC, il offre donc une meilleure robustesse au bruit d'environnement mobile comparé aux autres paramètres.

Références bibliographique

Bibliographie

[1] J. Chen, Y. Wang, D.L. Wang, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). A feature study for classification-based speech separation at very low signal-to noise ratio (2014).

[2] Yassine Mami, « Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence ».Thèse de doctorat, spécialité d'images et de signal, École nationale supérieure des Télécommunications Paris Tech, 2003.

[3] Alexandre Preti. «Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur ». Thèse de doctorat, spécialité d'informatique, Université d'Avignon et des pays de Vaucluse 2008.

[4] SayoudHalim, « Reconnaissance automatique du Locuteur-Approche connexionniste », Thèse de Doctorat spécialité d'électronique et Électrotechnique, Faculté d'Électronique et d'Informatique USTHB, 2003.

[5] AjgouRiadh, « Reconnaissance automatique du Locuteur à travers les canaux digitaux », Thèse de Doctorat, spécialité d'électronique, Université Mohamed Khider Biskra, 2016.

[6] Vincent Jousse, « Identification nommée du locuteur: exploitation conjointe du signal sonore et de sa transcription». Thèse de doctorat, spécialité d'informatique, Université du Maine, 2011. Français.

[7] Teva Merlin. «Amiral, une plateforme générique pour la reconnaissance automatique du locuteur de l'authentification à l'indexation». Thèse académie d'Aix-Marseille, spécialité d'informatique, université d'Avignon et des Pays de Vaucluse, 2004.

[8] I. Karlsson, T. Banziger, J. Dankovicov, T. Johnstone, J. Lindberg, H. Melin, F. Nolan, K. Scherer, 1998. *Speaker verification with elicited speaking-styles in theVerivox project*.Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pp. 207–210, Avignon (France).

[9] T. Banziger, G. Klasmeyer, T. Johnstone, T. Kamceva, K. R. Scherer. *Améliorer les systèmes de vérification automatique du locuteur en intégrant lavariabilité émotionnelle: Méthodes et premières données*. XXIIIème Journées d'Études sur la Parole (JEP), pp. 341–344, Aussois (France).

[10] A. Setlur et T. Jacobs. *Results of a speaker verification service trials using HMM models.* Workshop on Automatic Speaker Recognition, Identification,Verification, pp. 639–642, Martigny (Suisse).

[11] Bouzidi Adnane ,Guendouz ouard, «Vérification automatique du locuteur sur IP», Thèse de master, Spécialité d'électronique, Université Mohammed Boudiaf Msila, 2018.

[12] Ben Mahria Bilal, «L'approche hybride GMM-SVM pour l'identification des locuteurs en mode indépendant du texte en milieu fermé», thèse de master, spécialité de système intelligent et réseau, Université Sidi Mohamed Ben Abdallah, 2014-2015.

[13] Fredouille, Corinne, « Approche statistique pour la reconnaissance automatique du locuteur : informations dynamiques et normalisation bayésienne des vraisemblances». Thèse de doctorat, spécialité d'informatique, Université d'Avignon et des Pays de Vaucluse Donneur de grade 2000.

[14] Antonio M. Peinado, José C. Segura, «Reconnaissance de la parole sur les canaux numériques: Robustesse et normes », 2006. www.onlinelibrary.wiley.com/book/10.1002/0470024720

[15] Houda Kadi. « La reconnaissance automatique du locuteur par la voix IP», Thèse de master, spécialité de système intelligent et réseau Université sidi Mohamed ben Abdallah, 2014.

[16] Ruwei Li, Xiaoyue Sun, Yanan Liu, Dengcai Yang and Liang Dong, EURASIP Journal, Advances in Signal Processing, Multi-resolution auditory cepstral coefficients and adaptive mask for speech enhancement with deep neural network,2019.

[17] J. Oglesby et J. S. Mason, 1990. *Optimization of neural models for speaker identification*. International Conference on Acoustics, Speech, and SignalProcessing (ICASSP), pp. 261–264.

[18] KROBBA Ahmed, «Reconnaissanceautomatique du Locuteur via une Ligne de Communication GSM», thèse de Magister, Spécialité de Communication parlée, Université des Sciences et de la Technologie Houari Boumediene, Alger,2010.

[19] H. Dillon, Harvey, Author. Hearing aids, Hearing impaired , Rehabilitation, Sydney : Boomerang Press ; New York : Thieme, 2012. 9780957816817. 17 Vernon St. Turramurra NSW 2074

[20] A. S. Bregman. Auditory scene analysis: The perceptual organization of sound. Cambridge MA: MIT Press, 1990.

[21] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models, Digit. Signal Process. 10(1–3) (2000) 19–41.

[22] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. Audio Speech Lang. Process. 19(4) (2011) 788–798.

[23] Chen, J., Wang, Y., and Wang, D. "A bilevel framework for joint optimization of session compensation and classification for speaker identification.", Digital Signal Processing , Prn:22/03/2019; 9:06.

[24] D. L. Wang and G. J. Brown, editors. Computational auditory scene analysis: principles, algorithms and applications. Hoboken NJ: Wiley-IEEE Press, 2006.

[25] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice.An ecient auditory lterbank based on the gammatone function.Applied Psychology Unit Report 2341, 1988.

[26] S. K. Nemala, K. Patil, and M. Elhilali. A multistream feature framework based on bandpass modulation ltering for robust speech recognition. IEEE Trans. Audio, Speech, Lang. Process., 21:416{426, 2013.