REPUBLIQUE ALGERIENNE DEMOCRATIQUE POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE YAHIA FARES DE MEDEA
FACULTE DES SCIENCES ET DE LA TECHNOLOGIE

Département du Génie Électrique



MEMOIRE

Présenté pour l'obtention du Diplôme
de Mastre en Génie Électrique

Spécialité : *système des télécommunications*

THEME

REHAUSSEMENT DE LA PAROLE POUR LE SYSTEME DE RECONNAISSANCE DU LOCUTEUR EN ENVIRONNEMENT MOBILE BASE SUR LA METHODE DES RESEAUX DE NEURONES PROFONDS (DNN).

Préparé et présenté par :

Mr.GACEM Zohir.

Encadreur :

Dr. KROBBA Ahmed.

Année universitaire 2019/2020

بسم الله الرحمن الرحيم

Dédicaces

*Au nom **Allah** le Miséricordieux le très miséricordieux.*

J'ai l'immense honneur de dédier ce mémoire :

A mes trés chers parents qui étaient présents pour moi durant toute ma vie.
A mes frères et mes sœurs et ma famille

✍ *Gacem zohir*

Je dédie ce modeste travail

A ma chère mère zohra

A mon cher père mohammed

Qui m'ont entouré avec leur amour tout au long de ces années d'étude.

A mes frères et sœurs

A tous mes amis

Et en fin à tous ceux qui m'ont aidé

Remerciement

Je remercie tout d'abord, Allah qui nous a donné la force et le courage afin de parvenir à élaborer ce modeste travail

Je tiens à remercier mon promoteur Monsieur Krobba Ahmed pour son aide et l'intérêt avec lequel il a suivi mon travail.

Je voudrai également exprimer mes remerciements à tous nos enseignants de l'université Dr. YAHIA Fares de Médéa, Département de Génie Electrique pour la qualité de l'enseignement qu'ils nous ont bien voulu prodiguer durant nos études et aux membres du jury pour l'honneur qu'ils nous ont fait en acceptant de juger notre travail.

Et enfin, que nos chers parents, nos familles et nos amis , et bien avant tout, trouvent ici l'expression de nos remerciements les plus sincères et les plus profonds en reconnaissance de leurs sacrifices, aides, soutien et encouragement afin de nous assurer cette formation de la License dans les meilleures conditions.

Résumé

Notre travail appartient au domaine du traitement de la parole, précisément la reconnaissance automatique du locuteur dans un environnement mobile bruité et le rehaussement de la parole. L'objectif de ce travail est d'améliorer la performance du système de reconnaissance du locuteur plus robuste en environnement mobile très bruité, en utilisant de méthode des réseaux de neurone profond pour réduire d'effet bruit s dans le signal de parole.

Mots-clés : Reconnaissance automatique du locuteur, rehaussement de la parole, domaine spectral, réseau de neurone profond, environnement mobile bruité.

Abstract

Our work belongs to the field of speech processing, specifically automatic speaker recognition in a noisy mobile environment and speech enhancement. The objective of this work is to improve the performance of the more robust speaker recognition system in a very noisy mobile environment, using deep neural networks method to reduce noise effects in the speech signal.

Key words: Automatic speaker recognition, speech enhancement, spectral domain, deep neuron network, noisy mobile environment.

ملخص

ينتمي عملنا إلى مجال معالجة الكلام ، وتحديدًا التعرف التلقائي على السماعات في بيئة الهاتف المحمول الصاخبة وتحسين الكلام. الهدف من هذا العمل هو تحسين أداء نظام التعرف على السماعات الأكثر قوة في بيئة متنقلة صاخبة للغاية ، باستخدام طريقة الشبكات العصبية العميقة لتقليل تأثيرات الضوضاء في إشارة الكلام.

الكلمات الرئيسية: التعرف التلقائي على السماعات ، تحسين الكلام ، المجال الطيفي ، شبكة الخلايا العصبية العميقة ، بيئة الهاتف المحمول الصاخبة

.

Tables des Matières

Table des Matières.

Listes des Figures

Liste des Tableaux

Liste des Abréviations

Introduction générale……1

CHAPITRE.1 : Signale de parole et reconnaissance du locuteur

1.1 Introduction ……4

1.2 Généralité sur le signale parole……4

1.2.1 Définition ……..4

1.2.2 Caractéristique du signale parole……..4

1.2.3 Mécanisme de production du signal de parole……..5

1.2.4 Les sources de variabilités du signal de parole……6

I.3 Reconnaissance du locuteur……6

I.3.1 Différentes tâches en RAL et ses applications……...6

I.3.1.1 Identification automatique du locuteur……..7

I.3.1.2 Vérification Automatique du Locuteur ……8

I.4 L'architecture générale d'un système RAL.....9

1.4.1 Extraction de paramètre du signale parole ……...9

1.4.2 Modélisation du locuteur……10

1.4.3 Décision et mesure des performances ……17

I.5 Conclusion ……18

CHAPITRE.2 : Reconnaissance du locuteur et environnement mobile

2.1 Introduction…….....….20

2.2 Généralité sur les réseaux mobile……..20

2.3 Les différentes normes de codeurs de la parole…….21

2.4 Les différents Architecture du système de reconnaissance en environnement mobile..22

2.5Conclusion…….....26

CHAPITRE.3 : Rehaussement de la parole basé sur la méthode des réseaux de neurones profonds (Deep Neural Network, DNN).

3.1 Introduction…….…...29

3.2 Technique de rehaussement de la parole……29

3.2.1 Rehaussement par soustraction spectrale……29

3.2.2 Rehaussement par estimateur spectral EQMM……...30

3.2.3 Rehaussement par soustraction spectrale dans le domaine des modulations du spectre…….32

3.2.4 Rehaussement par estimateur EQMM des modulations du Spectre……...34

3.2.5 Rehaussement avec la technique de Fusion…….35

3.2.6. Méthodes de rehaussement utilisant des réseaux de neurones……...37

3.3 Rehaussement de la parole et l'apprentissage profond …….39

3.4Conclusion…….....40

CHAPITRE.4 : Expériences et Résultats

4.1 Introduction……...……42

4.2 Logiciels et langages utilisés dans l'application……42

4.3 Description des bases de données utilisées……....42

4.4 Étapes de la réalisation des systèmes de reconnaissance du locuteur……43

4.5 La structure générale de Système de reconnaissance du locuteur …….43

4.6 Extraction des paramètres……..44

4.6.1 Paramétrisation des coefficients MFCCs …….…….....44

4.6.2 Phase d'apprentissage……..45

4.6.3 Phase de test……45

4.7 Expériences et résultats…….45

4.7.1 Expérience 1……45

4.7.2 Expérience 2……46

4.7.3 Expérience 3……46

4.8 Conclusion……..47

Conclusion générale……48

Bibliographie

LISTE DES FIGURES

Figure 1.1 : Le signal parole ……..4

Figure 1.2 : Vue schématique de l'appareil vocal, dans le plan sagittal médian….....5

Figure 1.3 : Différentes taches du traitement de la parole.....7

Figure 1.4 : Principe de base de l'identification du locuteur ……..…….....8

Figure 1.5 : Vérification Automatique du locuteur …….……9

Figure 1.6 : Schéma typique d'un système de RAL(identification et Vérification)…. ……9

Figure 1.7 : Principe des machines a support vecteur ……13

Figure 1.8 : classes non linéairement séparables …….……...13

Figure 1.9 : Graphe décrivant, dans le cadre du modèle PLDA, les deux hypothèses (Ho,HI) utilisées pour calculer le score de vérification en présence d'un i-vecteur d'enrôlement e et d'un autre 1-vecteur pour le test t. …….....16

Figure1.10 : Processus de décision à base de seuil sur les scores.…….……18

Figure 2.1 : Architecture déportée ……..……..23

Figure 2.2 : Architecture embarquée ……...…….….24

Figure 2.3 : Architecture répartie ……..…25

Figure 3.1 : Fonction de poids, $\Psi[\omega[l]]$..……..36

Figure 3.2 : Filtrage multicanal utilisant des masques temps-fréquence estimés par un réseau de neurones……36

Figure 3.3 : Signal propre, corrompu par un bruit et rehaussé...……...39

Figure 4.1 : Structure générale d'un système RAL……40

LISTE DES TABLEAUX

Tableau .1.1 : Chronologie des modèles d'approches utilisés en RAL……..……10

Tableau. 3.1 Score PESQ avec un bruit aditif blanc gaussien (BABG)…….. ……...……37

Tableau. 4.1 Performance des systèmes i-vecteur-GPLDA et DNN-i-vecteur en utilisant de données de test bruitées pour le bruit de foule……...46

Tableau. 4.2 Performance des systèmes i-vecteur-GPLDA et DNN-i-vecteur en utilisant de données de test bruitées pour le bruit de voiture ……6

Tableau. 4.3 Performance des systèmes i-vecteur-GPLDA et DNN-i-vecteur en utilisant de données de test bruitées pour le bruit d'usine……47

Liste des Abréviations

RAL Reconnaissance Automatique du Locuteur

GMM Gaussian Mixtures Model

UBM Universal Background Model

JFA Joint Factor Analysis

PLDA Probabilistic Linear Discriminant Analysis

SVM Support Vector Machine

IAL Identification Automatique du Locuteur

RAP Reconnaissance Automatique de la Parole

VAL Vérification Automatique du Locuteur

LPCC Linear Predictive Cepstral Coefficients

PLP Perceptual Linear Prediction

MFCC Mel Frequency Cepstral Coefficient

PPCA Analyse en composantes principales probabiliste (*Probabilistic Principal Component Analysis*).

PCA Principal Component Analysis

LLR Log-Likelihood Ratio

FR Full Rate

HR Half Rate

EER Taux d'Egal Erreur

AMR Adaptive Multirate

EFR Enhanced Full Rate

IMT International Mobile Telecommunications

ETSI European Telecommunications Standards Institute

CDMA Code Division Multiple Access

WCDMA wide band CDMA

UMTS Universal Mobile Telephone System

TDMA Time Division Multiple Accès

FDD Frequency Division Duplex

ITU-T International Télécommunications Union – Telecoms

LTE Long Term Evolution

CELP Code Excited Linear Prediction

ACELP Algébrique CELP

AMR-NB Adaptive Multirate narrow-band

AMR-WB Adaptive Multirate wide-band

ETSI European Telecommunications Standards Institute

DSP Densité Spectrale de Puissance

EQMM Erreur Quadratique Minimum Moyenne

RSB Rapport Signal à Bruit

TFD Transformée de Fourier Discrète

DAV Détection d'Activité Vocale

AMS Analyse-Modification-Synthèse

PESQ Perceptual Estimation of Speech Quality

Introduction générale

Introduction générale

De nos jours, des applications telles que le retrait d'argent aux guichets automatique nécessitent l'identification de l'utilisateur. Dans le cas général, cette identification est réalisée à l'aide d'une smart carte et un numéro d'identification personnel, mais cela peut également être accompli en utilisant des caractéristiques biométriques. Il peut s'agir de l'empreinte digitale, de la forme de la main, du visage ou de la voix de l'individu. Cependant, lorsqu'il s'agit d'identifier l'utilisateur à distance grâce à l'utilisation d'un téléphone, par exemple, la manière possible d'effectuer cette identification est par la voix du locuteur. Le signal de parole ne contient pas seulement des informations linguistiques, qui sont utilisées pour la communication humaine, il fournit également des informations sur l'identité du locuteur.

Reconnaissance du locuteur est interprétée comme une tâche particulaire de la reconnaissance de formes, qui consiste à reconnaitre l'identité d'une personne par analyse de sa voix. Ce demain se compose de deux l'identification du locuteur et la vérification du locuteur. Un système de reconnaissance du locuteur se compose de trois étapes principales: la première est l'extraction de paramètres du signal vocal et qui est représenté de manière compacte, la seconde est la modélisation du locuteur qui peut être définie comme un processus de description des propriétés des de paramètres du locuteur, et la dernière étape est la décision.

Ce travail du mémoire s'intéresse à la reconnaissance automatique du locuteur dans les réseaux de communication (mobiles). Plus précisément, nous nous intéressons la rehaussement de la parole dans un environnement mobile, ce sujet est très vaste, et nécessite plusieurs études et conceptions, c'est-à-dire le traitement de la parole (extraction des paramètres, codage…) , télécommunication (réseaux mobiles…),

les normes de codeurs de la parole (FR, HR, EFR, AMR…). Et les différentes architectures du système de reconnaissance du locuteur en environnement mobile. . Ensuit les méthodes de rehaussement de la parole pour la reconnaissance du locuteur, dont on a essayé de donner le mieux à travers ce travail.

Ce mémoire est structuré en quatre chapitres. Après une introduction générale, nous décrirons dans le premier chapitre, les différentes tâches sur le système de reconnaissance du locuteur. Les paramétrisations du signal, les différentes approches de modélisation du

locuteur et les méthodes d'évaluation des performances des systèmes de reconnaissance du locuteur.

Dans le deuxième chapitre, la reconnaissance du locuteur via le réseau de communication mobile. Nous présenterons le système général de reconnaissance vocale via les réseaux mobiles ainsi que de différentes techniques de codage de la parole. Nous donnerons aussi les facteurs de dégradation de la performance du système de reconnaissance du locuteur via le réseau mobile. Nous présentons les différentes architectures du système de reconnaissance du locuteur dans le réseau mobile.

Le troisième chapitre est consacré sur la technique de rehaussement de la parole. Nous présenterons le rehaussement par soustraction spectrale et par estimateur EQMM et le rehaussement avec la technique de fusion ainsi que la méthode de rehaussement utilisant des réseaux neurones. Nous présentons le rehaussement de la parole et l'apprentissage profond.

Le quatrième chapitre exprimée l'expérience et résultats. Nous présentons logiciels et langages utilisé, bases de données ainsi que les étapes de la réalisation de système de reconnaissance du locuteur.

Chapitre 1

Signale de parole et reconnaissance du locuteur

1.1 Introduction

Depuis plusieurs dizaines d'années, la reconnaissance automatique du locuteur, i.e., la reconnaissance par la voix, fait l'objet de travaux de recherche entrepris par de nombreuses équipes de recherche dans le monde. La reconnaissance automatique du Locuteur - RAL - s'inscrit dans le domaine du traitement de la parole. La reconnaissance automatique du Locuteur généralement décrite comme l'identification d'une personne par l'analyse de sa voix.

Dans ce premier chapitre on a parlé sur le signal parole et ces caractéristique et les différentes taches de traitement de la parole , on explore les outils nécessaires à la reconnaissance automatique du locuteur, ses outils comme le prétraitement de signal parole et techniques d'extraction de paramètres et un état de l'art des différents modèles utilisés dans la littérature (GMM-UBM, JFA, PLDA ,SVM).

1.2. Généralité sur le signal de parole

1.2.1 Définition

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire, c'est un signal réel, continu d'énergie finie non stationnaire. Sa structure est complexe et variable avec le temps [**1**].

1.2.2 Caractéristiques du signal de parole

La parole est représentée par un signal vocal qui met en évidence des traits acoustiques qui sont : sa fréquence fondamentale, spectre, énergie [2].

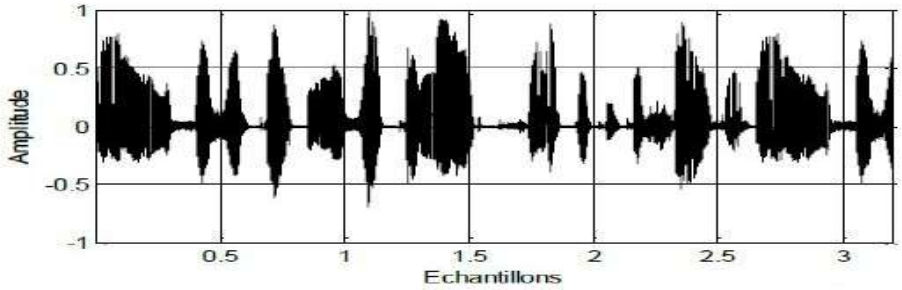


Figure.1.1 Le signal parole

1.2.3 Mécanisme de production du signal de parole

La reconnaissance automatique du locuteur basée sur la variabilité interlocuteurs due à la production de la parole pour séparer les individus les uns des autres.

a) L'appareil phonatoire

L'appareil phonatoire nous permet de produire des sons très variés dans un espace fréquentiel et énergétique pourtant limité. La production de la parole est un processus complexe dans lequel sont impliqués de nombreux organes dont les principaux composants responsables de la production de la parole sont illustrés dans le schéma de la figure1.2. L'appareil de phonatoire humain est divisé en trois composantes physiologiques : le système respiratoire (composé des poumons et de la trachée), qui fournit l'énergie et la quantité d'air nécessaire. La deuxième composante représente par le larynx qui permet d'obtenir un flux de nature vibratoire par l'action des cordes vocales. Le conduit vocal, compose des articulateurs, qui permet de moduler le signal glottique pour obtenir le signal de parole rayonne aux lèvres

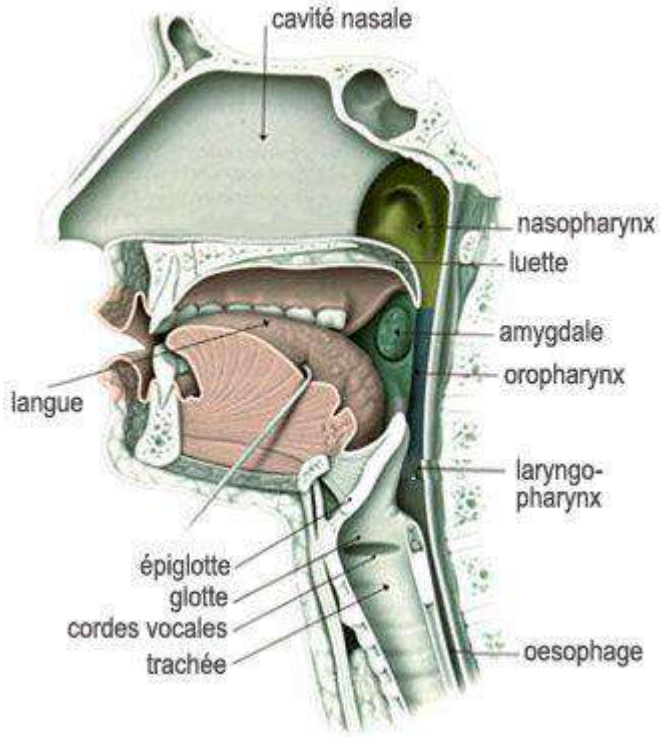


Figure.1.2 Vue schématique de l'appareil vocal, dans le plan sagittal médian.

La parole est une succession d'évènements sonores faisant alternativement apparaître des sons dits voisés caractérisés par la vibration des cordes vocales et des sons non voisés (qui ne font pas intervenir les cordes vocales.

Le signal voisé est un signal pseudo-périodique présentant des zones fréquentielles plus ou moins importantes. Ces zones fréquentielles d'enveloppe maximale sont appelées des formants.

1.2.4 Les sources de variabilités du signal de parole

La plupart des problèmes rencontrés par les systèmes de reconnaissance automatique du locuteur proviennent de la nature même de donnés manipulées. Les informations transmises par le signal parole sont multiples : (l'identité du locuteur, son genre, son état émotionnel, son état de santé, sa tranche d'âge, … etc.). Les conditions d'environnement influent sur le signal de parole ; les bruits additifs ambiants ou les bruits de canal de transmission. Ces composantes de variabilités interfèrent, rendant du même coup plus difficile l'extraction des seules informations caractéristiques du locuteur.

1.3. Reconnaissance du locuteur

1.3.1 Différentes tâches en RAL et ses applications

Le plus évident type d'application qui apparaît pour la reconnaissance automatique du locuteur est l'authentification de l'utilisateur au sein d'un système de sécurité comme le cas du contrôle d'accès à un bâtiment, un réseau ou toute autre ressource sensible. Ainsi des applications policières telles que l'automatisation d'écoute téléphonique. L'identification automatique du locuteur [3] et la vérification automatique du locuteur [4] sont les tâches essentielles de la RAL. Plus récemment, les besoins applicatifs ont fait naître de nouvelles tâches comme l'indexation par locuteur [5] de flux audio ou le suivi du locuteur (*Speaker Tracking*) [6] ou de nouvelles variantes telles que la détection de l'interaction d'un locuteur dans une conversation.

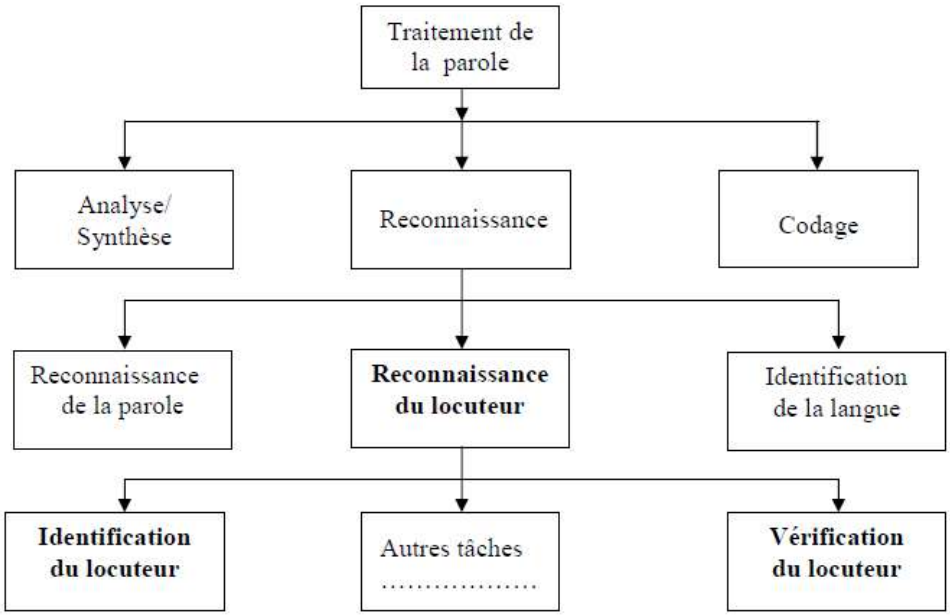


Figure1.3 Différentes taches du traitement de la parole

1.3.1.1 Identification automatique du locuteur

Le principe de l'identification automatique du locuteur 1'IAL, illustré par la figure 12. consiste à retrouver l'identité du locuteur associé parmi une population de locuteurs connus. D'un point de vue schématique, une séquence de parole est donnée en entrée du système d'IAL. Pour chaque locuteur connu du système, la séquence de parole est comparée " à une 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 [7]. Deux modes sont proposés en IAL : l'identification en ensemble fermé, dont on suppose que la séquence de parole est effectivement prononcée par un locuteur connu et l'identification en ensemble ouvert pour lequel le locuteur peut ne pas être connu. En mode "ensemble ouvert", le système doit décider de la fiabilité de son jugement en acceptant ou rejetant l'identité qu'il a trouvée [7]. De par son principe - déterminer une identité parmi des identités potentielles – les performances des systèmes d'IAL se dégradent à mesure que la population de locuteurs augmente [8.9].

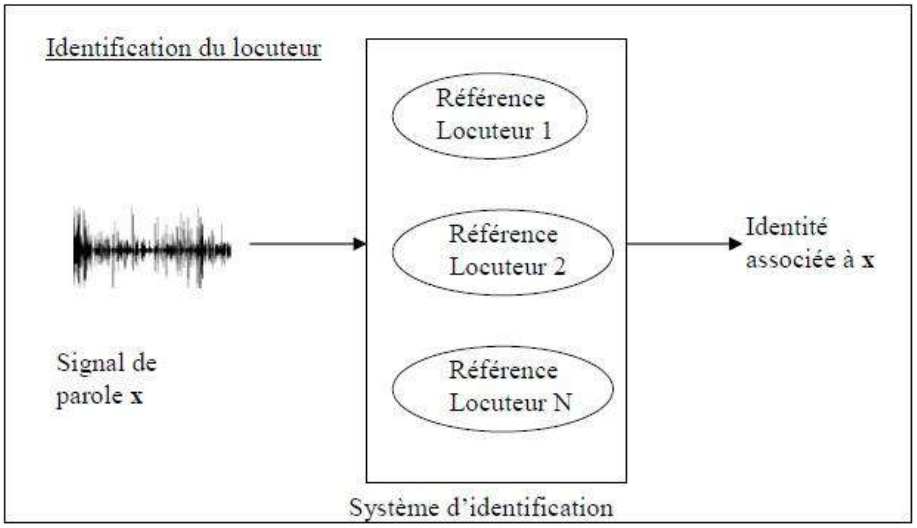


Figure 1.4 Principe de base de l'identification du locuteur

- Applications

En LAL. Les applications sont peu nombreuses. On peut retenir, par exemple, l'utilisation d'un système d'IAL en vue de faciliter l'adaptation au locuteur des systèmes de reconnaissance automatique de la parole (RAP). Par ailleurs, il peut être intéressant pour des applications commerciales d'associer un même mot de passe pour une petite population de locuteurs (membres d'une famille, d'une société). Dans une telle situation, un système d'IAL en ensemble ouvert et dépendant du texte peut être utilisé pour contrôler l'accès à des données sensibles [7].

1.3.1.2 Vérification Automatique du Locuteur

La Vérification Automatique du Locuteur (VAL) est le processus décisionnel permettant de déterminer, au moyen d'un message vocal, la véracité de l'identité revendiquée par un individu dont la figure 13 représente le principe de VAL [10, 11]. L'identité ainsi que le message vocal constitue les deux entrées du système de VAL. L'identité, nécessairement connue du système, désigne automatiquement la référence caractéristique d'un locuteur. Une mesure de similarité basée sur le rapport de vraisemblance est calculée entre cette référence et le message vocal puis comparée à un seuil de décision. Dans le cas où la mesure de similarité est supérieure au seuil, l'individu est accepté. Dans le cas contraire, l'individu est considéré comme un imposteur et sera rejeté [10, 11].

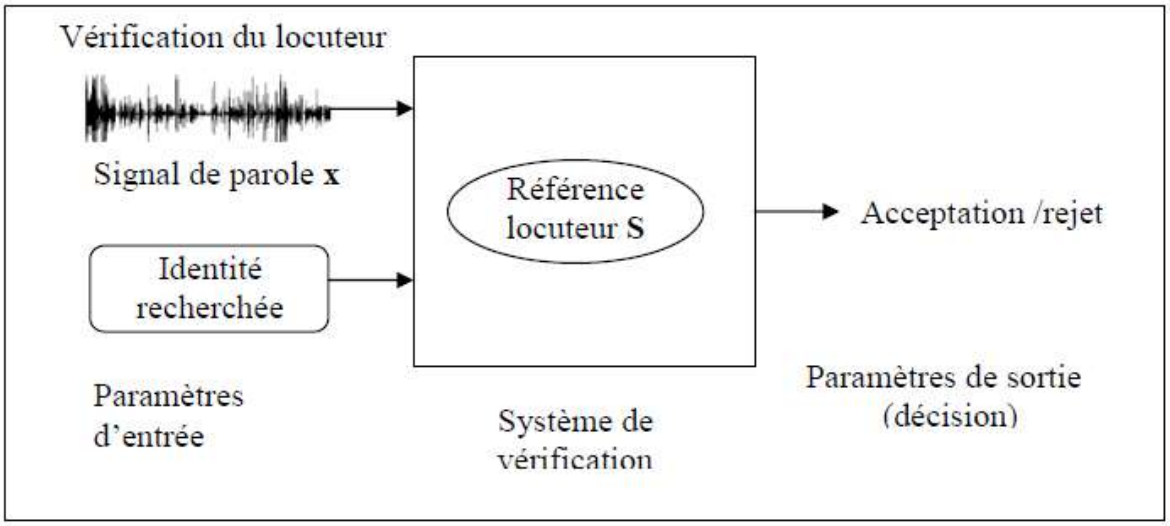


Figure1.5 Vérification Automatique du locuteur

Applications

Les applications de VAL sont multiples et principalement commerciales:

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.).

1.4 L'architecture générale d'un système RAL

Un système reconnaissance automatique du locuteur, quelle que soit la tache (IAL, VAL, indexation en locuteur) possède trois processus principaux, qui sont : la paramétrisation (extraction des paramètres), la modélisation du locuteur et la décision

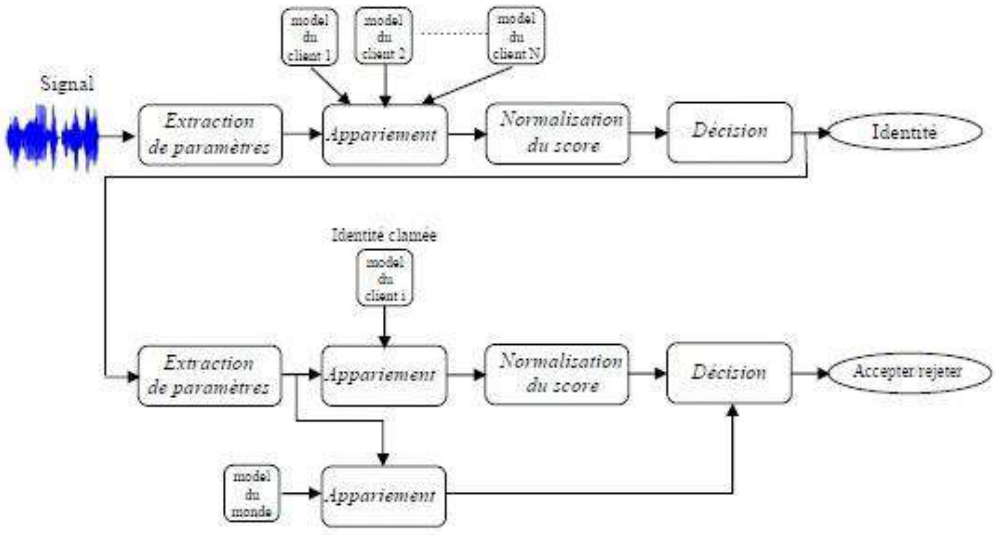


Figure1.6 Schéma typique d'un système de RAL(identification et Vérification)

1.4.1 Extraction de paramètres du signal parole

Il existe différentes méthodes de représentation du signal de parole. Certaines ont été spécifiquement développées par l'étude acoustique de la production de parole. Autres paramètres simulé les caractéristiques de l'oreille humaine. On peut classifier ces paramètres en deux grandes catégories :

- **Les paramètres prosodiques:** ces paramètres illustrent le débit, la durée et les fréquences de pauses, pitch et l'énergie.
- **Paramètres de l'analyse spectrale :** Ces paramètres sont utilisés en RAL. Ils représentent l'influence des caractéristiques physiques de l'appareil phonatoire de (forme du conduit vocal) chaque individu. Ces paramètres sont les LPCC, PLP et MFCC.

1.4. 2 Modélisation du locuteur

L'étape de modélisation exploite les données fournis dans l'étape de la paramétrisation afin de créer la représentation d'un individu qui servira, par la suite, à l'authentifier. Le modèle utilisé est généralement une représentation statistique des données acquises. Différentes approches ont été développées.

1987	VQ (Soong et al., 1987).
2000	GMM-UBM (Reynolds et al., 2000).
2003	Modèle Eigenchannel (Kenny et al., 2003).
2005	Modèle Eigenvoice (Kenny et al., 2005).
2006	GMM-SVM (Campbell et al., 2006a).
2007	Modèle JFA (Kenny et al., 2007).
2011	i-vecteur (Dehak et al., 2011).

Tableau 1.1 Chronologie des modèles d'approches utilisés en RAL.

a) GMM-UBM

La deuxième étape du processus d'extraction des i-vecteurs consiste à construire un modèle acoustique a priori. La distribution complexe des vecteurs acoustiques nécessite une modélisation multimodale. En fait, la modélisation par les mélanges

de gaussiennes **(GMM**) est devenue le paradigme le plus répandu dans le domaine de la reconnaissance du locuteur. Le modèle du monde (**UBM** pour Universal Background Model) est le terme utilisé pour référer au **GMM** de C composantes gaussiennes entrainé à partir d'une quantité importante (typiquement des centaines d'heures) de vecteurs acoustiques de la parole de plusieurs locuteurs. L'estimation des paramètres de ce **GMM, à** savoir $\mathbf{\Omega}=\{\mathbf{Wc}, \boldsymbol{\mu_c}, \sum\mathbf{c}\}$ **c=1..c** , est basée sur le critère du maximum de vraisemblance (ML). Dans le cas du modèle du monde, le choix de ce critère d'estimation est naturel en raison de la disponibilité d'une énorme quantité de données d'apprentissage.

De par sa nature, un extracteur des i-vecteurs ne fait aucune distinction entre les principales sources de la variabilité intrinsèque dans le signal vocal. En fait, un enregistrement d'un locuteur donné est considéré comme indépendant des autres enregistrements, même s'ils appartiennent à ce même locuteur. Par conséquent, un vecteur caché (c.-à-d un i-vecteur) **x** de dimension D x 1 est associé à chaque enregistrement afin de tenir compte de la variabilité inter-enregistrements uniquement. L'hypothèse principale est que la distribution des vecteurs acoustiques extraits d'un enregistrement donné est un mélange de C gaussiennes dont $\sum c$ est la matrice de covariance de la $c^{iéme}$ composante, son vecteur moyen est donné comme suit :

$$u_c = m_c + T_c \mathrm{X} \qquad (1.1)$$

où, m_c est le vecteur moyen de dimension F de la composante c du modèle a priori (c.-à-d. l'UBM), la matrice T_c de dimension F R D est la cième partie de la matrice T de dimension CF * D qui représente la matrice de projection d'espace des i-vecteurs. Elle est aussi souvent connue sous le nom de matrice de variabilité totale.

Donc, le logarithme de la vraisemblance d'un enregistrement vocal représenté par une séquence de vecteurs acoustique $\mathbf{A}=\{\boldsymbol{a_t}\}_{\boldsymbol{t=1.T}}$, dont l'alignement de chaque vecteur est préalablement donné, peut être écrit tout simplement de la manière suivante :

$$\sum\left[N_c \ln \frac{1}{(2\pi)^{f/2}\left|\sum_c\right|^{1/2}} - \frac{1}{2}\sum_t (a_t - T_c \mathrm{X} - m_c)\sum_c^{-1}(a_t - T_c \mathrm{X} - m_c)\right] \quad (1.2)$$

où le scalaire Ne représente le nombre des trames alignées avec la gaussienne c. Il est à noter que la formule du logarithme de vraisemblance (1.2) peut être évaluée en fonction des statistiques de premier et de second ordres calculés pour chaque gaussienne comme suit :

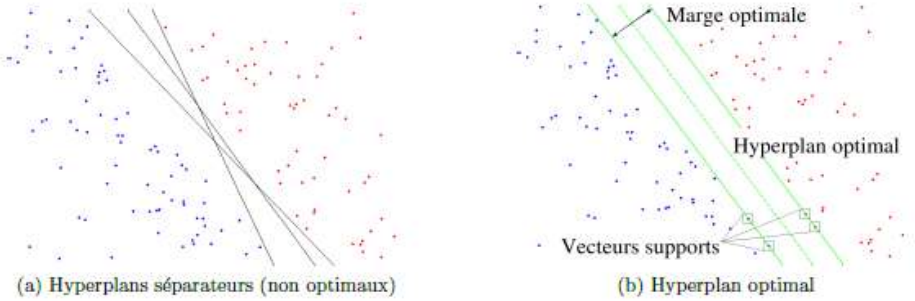
$$\boldsymbol{F_c} = \sum_{\boldsymbol{t}} \boldsymbol{a_t} \quad (1.3)$$

$$S_c = \sum_t a_t \alpha_t^{'} \quad (1.4)$$

b) SVM

La machine à vecteurs de support (Support Vector Machine) SVM est un classifieur discriminant qui sépare deux classes (ayant comme labels +1 et -1) par un hyperplan de séparation [12]. C'est une technique très utilisée en classification et régression, grâce à son bon pouvoir de généralisation.

Pour un ensemble de vecteurs paramétriques de deux classes (linéairement séparables). il existe une multitude d'hyperplans séparateurs qui séparent les données de ces deux classes. Mais seulement un seul de ces hyperplans maximise la marge entre les données et la frontière de séparation. La marge étant la distance entre l'hyperplan et les données les plus proches (appelées vecteurs supports). Cet hyperplan optimal donne les meilleures performances en généralisation. Disposant de données d'apprentissage {x} (de classe +1,-1) le problème dans la modélisation par SVM est de trouver l'hyperplan optimal. La figure 1.7. montre le principe de l'hyperplan optimal et de la marge optimale dans la modélisation par SVM



.

Figure 1.7 Principe des machines a support vecteur

La plupart des applications pratiques correspondent en réalité à des classes non linéairement séparables, où il n'existe aucune frontière de séparation linéaire capable de séparer les données des deux classes. Le problème est dit non linéairement séparable. La figure Fig. 1.8 donne un exemple de ce type de problèmes. Afin de pouvoir traiter le cas de données non linéairement séparables, le principe des SVM est de transposer le problème dans un espace de dimension supérieure, où il est probable de trouver une frontière de séparation linéaire (un hyperplan de séparation). Ceci est réalisé grâce à l'utilisation d'une fonction noyau (kernel function) respectant les conditions de Mercer. La fonction noyau s'exprime sous forme **de** $\mathbf{K(x,y) = \phi(x)^T \phi(y)}$, où $\boldsymbol{\phi(x)}$ est une expansion de l'espace de données

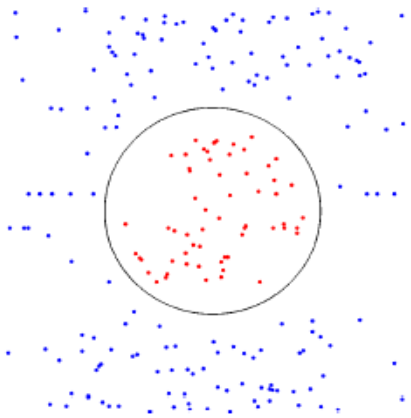


Figure 1.8 classes non linéairement séparables

(l'espace d'entrée) à l'espace des caractéristiques de plus haute dimension (l'espace de redescription). La fonction discriminante de la **SVM** est donnée par :

$$f(x) = \sum_{i=1}^{L} a_i y_i k(x, x_i) + b$$
$$= \left(\sum_{i=1}^{L} a_i y_i \theta(x_i) \right)^T \theta(\mathrm{X}) + b = w^T \theta(\mathrm{X}) + b \qquad (1.5)$$

où y_i ***E {+1, -1}*** sont les sorties (les labels) idéales,$\sum_{i=1}^{L} a_i y_i$**= 0 et** y_i**> 0**. Notons que l'équation de l'hyperplan de séparation *H(W,b) est* ***f(x) = 0***. Le modèle SVM, i.e., les **L** vecteurs supports x_i, leur poids correspondants a_i; et le paramètre **b** (un décalage), est obtenu en résolvant un problème

d'optimisation quadratique. Durant la phase de reconnaissance, le score du modèle SVM f_x sur les données y est :

$$f_x(y) = w_x^T \theta(y) + b_x \qquad (1.6)$$

Étant de base des classifieurs binaires, deux principales approches ont été proposées pour utiliser les **SVM** dans la séparation multi-classes : l'approche une contre toutes (one against all) et l'approche par paires (one against one, appelée aussi pairwise) (Hsu and Lin, 2002). On apprend dans la première, **C SVM** (C étant le nombre de classes) séparant chacune entre une classe donnée et l'ensemble des autres classes. Alors qu'on apprend dans la deuxième, **C(C-1)/2 SVM** séparant chacune les deux classes de toute paire de classes possible, et construire ensuite un arbre de décision à partir de ces **SVM**.

c) JFA

Le modèle d'analyse conjointe de facteurs proposé par [13], (Joint Factor Analysis, JFA), est un modèle génératif des variables cachées qui tente principalement de s'attaquer au problème de la disparité entre les données d'enrôlement et celles de test dans l'espace des supervecteurs des GMM. L'hypothèse de base de ce modèle stipule qu'un supervecteur S de dimension CF dépendant du locuteur et du canal (c.-à-d un supervecteur estimé d'un segment donné d'un locuteur donné) peut se diviser en deux supervecteurs statistiquement indépendants comme suit :

S=S+C (1.7)

où s est le supervecteur qui dépend uniquement du locuteur et c est celui qui dépend du canal. Le supervecteurs dépendant du locuteur est donné par la formule suivante :

$$\mathbf{S=m+Vy+Dz} \qquad (1.8)$$

où les supervecteurs moyens m représentés souvent par le supervecteur du modèle du monde UBM, V connue par la matrice des voix propres, c'est une matrice rectangulaire de dimension CF * R; (avec R <CF), elle peut aussi être vue comme une matrice de projection en dimension réduite, elle modélise principalement la variabilité interlocuteur. D est une matrice diagonale de dimension CF X CF qui joue un rôle similaire à celui de l'information a priori dans l'adaptation MAP du modèle GMM-UBM. Les vecteurs cachés y et z sont respectivement les facteurs du locuteur (speaker factors) et les facteurs communs (common factors). Les vecteurs cachés suivent une distribution normale standard N.,0,1) où 0 est un vecteur moyen dont toutes ses composantes sont égales à zéro et I est une matrice-identité. Donc, le supervecteur s suit une distribution gaussienne d'un vecteur moyen m et d'une matrice de covariance qui est égale à VV'+D.

Le supervecteur c dépendant du canal est donné par la formule suivante :

$$\mathbf{c=Ux} \qquad (1.9)$$

où la matrice U, dite matrice des canaux propres (eigenchannels), est une matrice rectangulaire de dimension CF * Rc (avec R<CF), ses colonnes constituent les axes de l'espace du canal de dimension réduite. Le vecteur cachés de dimension Rc distribué selon une loi normale standard est appelé le vecteur des facteurs du canal (channel factors). Ainsi, le supervecteur c suit une loi normale d'un vecteur moyen nul et d'une matrice de covariance égale à UU'.

De plus amples détails sur l'implémentation du modèle JFA se trouvent dans [13].

d) PLDA

Contrairement à ce qu'indique son appellation, le PLDA est un modèle purement génératif, il n'est en réalité qu'une version miniature du modèle JFA opérant

Cette appellation est une analogie avec celle de PPCA qui est une version probabiliste de PCA ordinaire [14].

dans le sous-espace des i-vecteurs au lieu de celui des supervecteurs. Le modèle de base de PLDA est apparu pour la première fois dans le domaine de l'imagerie où il est utilisé pour la reconnaissance faciale [15]. Il est ensuite introduit dans la vérification du locuteur avec deux variantes et quelques modifications dans l'algorithme d'apprentissage [16].

En présence d'un essai de vérification x={e t}composé de deux enregistrements vocaux représentés par des i-vecteurs respectivement pour enrôlement et pour test, la tâche de la vérification du locuteur via le modèle PLDA consiste à évaluer un score mesurant le degré de la ressemblance entre ces deux i-vecteurs. le rapport de vraisemblance LLR est le score calculé à travers un test d'hypothèses (voir Figure). A vrai dire, le LLR est le logarithme du rapport entre la probabilité que les deux i-vecteurs appartiennent au même locuteur (hypothèse H.) et la probabilité qu'ils appartiennent à deux différents locuteurs (hypothèse H1).

Les scores produits par le modèle PLDA de la manière précédemment décrite ne nécessitent aucune normalisation subsequente et dans le cas de la parole téléphonique [16] et dans le cas de la parole microphonique [17]

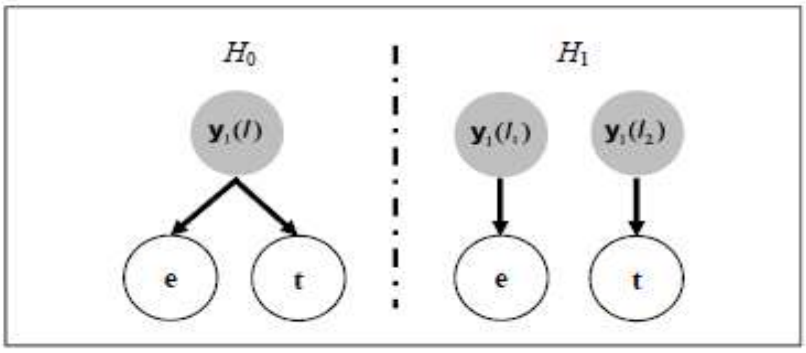


Figure 1.9 Graphe décrivant, dans le cadre du modèle PLDA, les deux hypothèses (Ho,HI) utilisées pour calculer le score de vérification en présence d'un i-vecteur d'enrôlement e et d'un autre 1-vecteur pour le test t.

1.4.3 Décision et mesure des performances

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.

a) Taux identification

Pour l'identification du locuteur on a la relation suivante:

$$Ic = \frac{\text{Nombre de tests correctement identifiés}}{\text{Nombre total de tentatives}}$$

$$Ii = \frac{\text{Nombre de tests mal identifiés}}{\text{Nombre total de tentatives}}$$

Les performances du système d'identification sont données en termes de taux d'identification correcte *Ic* ou incorrecte *Ii*.

b) Taux d'erreur (EER)

Le taux d'égale erreur (EER; Equal Error Rate) est l'une des mesures les plus populaires en vérification de locuteurs vu qu'elle permet de comparer deux systèmes en se basant sur une seule mesure. Elle est définie comme le point opératoire où les valeurs FAR et FRR deviennent presque égales. Cette configuration est atteinte en faisant varier le seuil de décision t jusqu'à ce que les deux zones correspondant aux fausses acceptations et aux faux rejets (voir la figure) deviennent égales.

Il convient de noter que l'EER ne constitue pas nécessairement un point opératoire optimal dans les applications réelles qui peuvent nécessiter des niveaux de sécurité élevés.

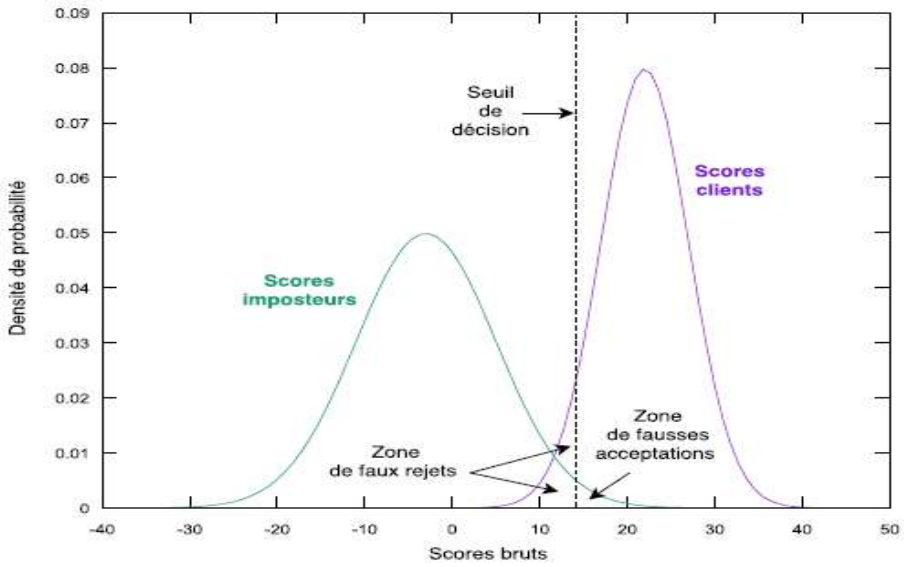


Figure 1.10 Processus de décision à base de seuil sur les scores.

1.5 Conclusion

Ce chapitre a un rôle introductif au domaine de la reconnaissance automatique du locuteur. Dans ce chapitre nous avons décri les différentes classes de paramètres de l'analyse acoustique du signal parole. Le signal de parole est un processus aléatoire non stationnaire à long terme. Un système de reconnaissance automatique du locuteur, quelle que soit la tâche considérée, se résume en trois étapes principales : l'analyse acoustique du signal parole, la modélisation du locuteur et la décision soit une décision et vérification. Également, tout système de RAL dépend de la technique d'extraction de paramètres utilisé, modélisation, décision et ainsi la phase de prétraitement.

Chapitre 2

Reconnaissance du locuteur et environnement mobile

2.1 Introduction

Dans ce chapitre, nous nous intéressons aux bases de la reconnaissance du locuteur dans la communication mobile, au débit nous présentons d'une façon générale les différentes générations de réseaux mobiles. Nous décrivons aussi un aperçu sur les normes de codeurs de la parole: (FR, HR, EFR, AMR-NB et AMR-WB). Enfin, nous citons les différentes architectures du système de reconnaissance du locuteur en environnement mobile.

2.2 Généralité sur les réseaux mobile

La téléphonie mobile numérique a été principalement développé et déployé durant les années quatre-vingt et quatre-vingt-dix, poussés par la forte demande pour cette technique et la nécessité d'un système unifiés pour elle afin de réduire les coûts et d'offrir une qualité maximale de service (QoS) (Quality of Services). En Europe, le mobile GSM (Global System of Mobile communication) a été créé en 1982 par la CEPT (Conférence européenne des postes et télécommunications) pour la standardisation d'un système unifié de radio télécommunication de deuxième génération (2G) et dans la bande de 900 MHz . Versions de GSM ont été également créés dans les bandes de 1800 et 1900 MHz (systèmes DCS1800 et DCS1900). Le travail du groupe GSM a pris une dizaine d'années, période pendant laquelle le GSM acquit le système global de sens pour mobile, et ses résultats ont été publiés comme un ensemble de normes de l'ETSI (European Telecommunications Standards Institute). Actuellement, cette norme est mise en oeuvre dans presque le monde entier .

La troisième génération (3G) des communications mobiles est un concept identifié avec I'IMT-2000 (International Mobile Telecommunications, 2000). Il s'agit d'une série de recommandations qui, en fait, a permis d'obtenir plusieurs systèmes 3G comme le CDMA -2000 (CDMA-Code Division Multiple Access) et la téléphonie mobile universelle UMTS (Universal Mobile Telephone System). UMTS a été lancé en 1998 par une association de plusieurs organisations appelées 3GPP.

La vitesse de transmission (de 2G vers 3G) des données s'est augmentée de 144 kbit/s à 2 Mbit/s. L'interface radio UTRA (UMTS Terrestrial Radio Access) utilise deux modes différents pour la liaison radio: TDD (Time Division Duplex- duplex à division de temps pour les bandes 1900 – 1920 MHz et 2010 - 2025 MHz ainsi pour la liaison radio FDD (Frequency Division Duplex -division de fréquence duplex) pour les bandes 19201980 MHz et 2110-2170 MHz. La technique d'accès est CDMA avec deux variantes: WCDMA (wide band CDMA) et une combinaison TDMA (Time Division Multiple Accès)/CDMA pour TDD

Il y a en conséquence d'autres générations de réseaux mobiles comme la 4e génération (4 G) et actuellement les recherches tendent à réaliser le 5G. En télécommunications, les 4 G, permet le « très haut débit», c'est-à-dire des transmissions de données à des débits théoriques supérieurs à 100 Mb/s, voire supérieurs à 1 Gb/s (débit minimum défini par l'ITU-T - International Télécommunications Union - Telecoms - Union internationale des télécommunications). En pratique, les débits sont de l'ordre de quelques dizaines de Mb/s selon le nombre d'utilisateurs, puisque la bande passante est partagée entre les terminaux actifs des utilisateurs présentes dans une même cellule radio. Une des particularités de la 4G est d'avoir un « cœur de réseau » basé sur IP et de ne plus offrir de mode commuté (établissement d'un circuit pour transmettre un appel "voix"), ce qui signifie que les communications téléphoniques utiliseront la voix sur IP (en mode paquet).

Les sociétés actuellement ont publié des spécifications qui imposent des téléphones pour pouvoir exploiter son réseau 4G LTE (Long Term Evolution). La $5^{ième}$ génération de standards pour la téléphonie mobile, faisant suite à la 4G. La technique 5G pourrait permettre des débits de télécommunication mobile, de plusieurs gigabytes de données par seconde, soit jusqu'à 1 000 fois plus rapides que les réseaux mobiles en 2010.

2.3 Les différentes normes de codeurs de la parole

Le but de codage du signal parole est de représenter le signal par le moins de bits possible, mais maintenir un certain niveau de qualité du signal parole décodé. En général, les techniques de codage sont basées sur la réduction de la redondance du signal, afin que le signal résiduel (après réduction de redondance) puisse être encodé

avec moins de bits que l'originale, en distingue plusieurs norme de codeur de la parole :

Codeur EFR

Le codeur EFR (Enhanced Full rate) Un exemple important d'un codeur ACELP, il apporte une certaine amélioration de la qualité par rapport aux codeurs FR et HR, notamment dans les communications mobiles. Le codeur EFR utilise 12.2 kb/s pour le codage de la parole et 10.6 kb/s pour la gestion des erreurs.

Codeur AMR

Le codeur AMR (Adaptive Multirate-AMR) est une famille de codeurs ACELP avec la même structure que l'EFR, mais travaillant à différents débits. À un instant donné. Les codeurs AMR utilisent une longueur de trame de 20 ms et un jeu de coefficient LSP est calculé une fois par image (sauf à 12,2 kbit/s, EFR) en appliquant une fenêtre asymétrique qui met l'accent sur les échantillons plus récents. La fiche d'AMR a également été adoptée par l'UMTS (spécification 3GPP TS 26.090).

- **Codeur AMR-NB :** (Adaptive Multirate narrow-band) est un système de compression de données audio optimisée pour le codage de la parole de faible bande passante normalisé par l'ETSI.
- **Codeur AMR-WB :** (Adaptive Multirate wide-band) est un système de compression de données audio optimisée pour le codage de la parole de bande passante plus grande que celle des téléphones classique normalisé par l'ETSI.

2.4 Les différents Architecture du système de reconnaissance en environnement mobile

Nous pouvons dénombrer trois grandes architectures pour la reconnaissance de la parole dans les systèmes embarqués

a) La reconnaissance déportée

Le principe de la RAP déportée consiste à se servir du terminal uniquement comme un système d'acquisition du son. Une fois le signal enregistré, il est transmis via un réseau, généralement sans fil, à un serveur qui effectue le travail de reconnaissance.

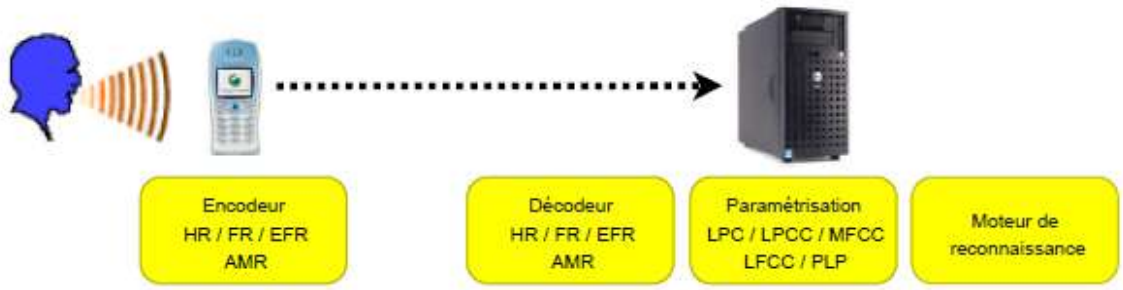


Figure.2.1 Architecture déportée

Cette approche permet d'envisager l'utilisation de serveurs beaucoup plus puissants et donc de fournir des services plus divers et généralement de meilleure qualité. La présence d'un serveur performant permet d'adjoindre d'autres modules à la reconnaissance vocale tels qu'un système de dialogue ou encore des vocabulaires dédiés à chaque application. Pour résumer, l'avantage majeur de cette architecture est la levée des contraintes liées aux ressources limitées des systèmes embarqués.

Un des inconvénients de ce type d'architecture est la perte d'informations due à la transmission du signal. En effet, avant d'être envoyé au(x) serveur(s) de reconnaissance, le signal doit être encodé suivant un des codecs de l'opérateur (HR ,FR , EFR , AMR. . .). Ces codages (avec pertes) sont nécessaires car l'émission du signal sur le réseau a un cout non négligeable. De plus, au cours d'une transmission sur un réseau (sans fil ou non), différents phénomènes, comme la perte de paquets d'informations ou l'ajout de bruits viennent perturber le signal.

Parmi les inconvénients, nous pouvons encore citer le cout (tant sur le plan financier que sur l'aspect du trafic réseau) de transmission du signal au serveur et de la réponse du serveur au terminal.

Enfin, l'absence d'autonomie représente certainement l'inconvénient majeur de cette architecture. En effet, elle suppose que le téléphone soit connecté au réseau pour pouvoir fournir les services de reconnaissance, interdisant l'emploi de ceux-ci en local pour des applications comme les jeux, l'agenda. . .

Comme exemple d'application, le lecteur pourra se tourner vers le projet

”MultiMedia Terminal Mobile” [20] (projet MTM IST 1999-11100) auquel a participé le LIA. Ce projet avait pour objectif de concevoir et de réaliser un système qui regrouperait les fonctionnalités d’un PDA, d’un téléphone portable de troisième génération (UMTS), d’un ordinateur portable et d’une caméra.

Dans le cadre de ce projet, le LIA a développé un serveur permettant de faire de la reconnaissance déportée de mots isolés, ainsi que de la reconnaissance de locuteurs (un mode de reconnaissance répartie a également été présenté).

b) **La reconnaissance embarquée**

Avec cette approche, toutes les étapes de la reconnaissance sont effectuées dans le système embarqué. Aucun serveur `a distance est nécessaire, le système est donc entièrement autonome : il réalise lui-ême l’acquisition du son puis le décodage.



Figure.2.2 Architecture embarquée

L’intérêt principal de cette approche est l’absence de contrainte : en effet, il n’est pas nécessaire de se limiter `a l’utilisation des codecs de parole définis par l’ETSI1, car aucune compatibilité avec un serveur n’est requise. De plus, les inconvénients de la reconnaissance déportée sont résolus par l’absence de transmission.

Cette solution présente toutefois un inconvénient majeur : les ressources nécessaires. En effet, toutes les phases (paramétrisation du signal et reconnaissance) sont effectuées au sein du téléphone, les différentes phases de la reconnaissance nécessitent beaucoup de ressources comparées aux capacités d’un téléphone portable. De plus, l’augmentation des ressources d’un téléphone (mémoire et puissance de calcul), bien que possible, est trés couteuse.

Dans la littérature plusieurs implémentations dans des systèmes embarqués sont proposées. Les ressources disponibles dans ces systèmes sont variables suivant les projets. Trois grandes classes peuvent être distinguées :

- **très peu de ressources** : DSPFactory [21] propose un système nécessitant un DSP2/CPU à 4MHz. Un système présenté par Wang permet d'embarquer un système de reconnaissance de mots isolés sur un MCU3 à 8MHz [22] ;
- **peu de ressources** : le système d'IBM [23] recommande un DSP/CPU `a 50 MHz et 1Mo de mémoire, celui développé par l'université de Graz[24] un DSP/CPU à 30 MHz et ceux de Siemens [25] et [26] et Texas Instrument [27] - 128Ko de mémoire) exigent un DSP/CPU à 100 MHz;
- **"système de demain"** : NEC [28] propose un système ayant besoin d'un processeur triple cœur (3 systèmes à 200MHz) et 6Mo de mémoire.

Ce dernier système est beaucoup plus gourmand que les autres. Il semble plutôt destiné aux futures applications et paraît difficilement intégrable aujourd'hui au regard des ressources des systèmes embarqués disponibles actuellement.

c) La reconnaissance répartie

Ce mode de reconnaissance se situe entre la RAP déportée et la RAP embarquée. En effet, une partie du travail est effectuée sur le système embarqué et une autre partie sur un serveur distant. Généralement, la phase d'extraction des paramètres est réalisée sur le terminal (cf.[29] , [30],[31]. . .) et la partie reconnaissance - à proprement parler - se trouve sur un serveur.

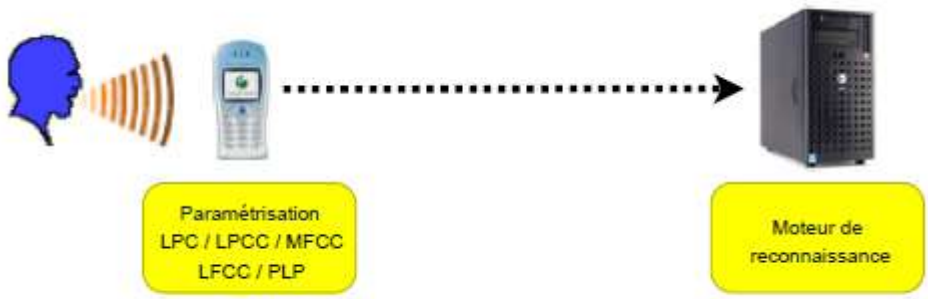


Figure.2.3 Architecture répartie

Cette approche permet de contourner les problèmes liés à la perte d'informations due au transport. Avec cette architecture, une seule phase "d'extraction de l'information" est nécessaire : les deux premières étapes (encodage et décodage) de l'architecture deviennent inutiles. Les pertes engendrées par l'encodage effectué sur le

téléphone sont alors évitées (HR, FR, EFR et AMR sont tous des encodeurs avec perte).

Cette approche implique, par ailleurs, que le téléphone (et donc son fabricant) s'accorde avec l'opérateur désirant fournir le service de reconnaissance vocal, étant donné que l'extraction des paramètres est effectuée directement dans le téléphone. Ces paramètres sont ensuite utilisés par le système de reconnaissance durant la phase de décodage. Les codecs4 **(**Logiciel permettant de convertir le signal analogique en signal numérique. Généralement, il permet aussi la compression du signal numérisé.**)** classiques, définis par l'ETSI, sont plutôt bas´es sur les coefficients LPC alors que les systèmes actuels de RAP sont essentiellement bas´es sur les coefficients MFCC ou PLP ou encore sur des dérivés de l'une de ces deux paramétrisations, il est donc nécessaire d'ajouter des fonctions spécifiques de paramétrisation au sein des téléphones.

Enfin, cette approche (comme la RAP déportée) impose une connexion constante au réseau GSM pour bénéficier des services de RAP.

2.5 Conclusion

Ce chapitre a présenté une vue d'ensemble sur la reconnaissance du locuteur sur les réseaux mobile, sans fil et internet. On a donné un aperçu sur le signal parole dans les réseaux mobiles .Le codage ou la compression du signal parole est considéré comme un facteur important dans la dégradation d'information transmise à travers le canal. Bien que le type de codage ou compression soit un facteur important, dans ce chapitre on explore les différents standards définissant les types de codage du signal parole. Après avoir présenté les trois architectures envisageables (reconnaissance déportée, reconnaissance embarquée et reconnaissance répartie) pour la RAP dans les réseaux mobile, le choix d'une reconnaissance embarquée a été fait car cette architecture est la seule qui permette une utilisation entièrement autonome des fonctions de reconnaissance vocale. En effet, la reconnaissance répartie et la reconnaissance déportée font toutes deux appel à des serveurs de reconnaissance externes.

Chapitre 3
Rehaussement de la parole par la méthode DNN

3.1 Introduction

Dans ce chapitre, nous nous intéressons aux bases de Rehaussement de la parole et ses techniques. Le rehaussement de la parole est un domaine de recherche très actif qui s'est largement développé durant ces trois dernières décennies. Sa nécessité s'est manifesté en particulier dans de nombreuses situations, dans lesquelles le signal vocal provient soit d'une source bruitée, soit il est affecté directement par le bruit du canal des transmissions. En effet, le rehaussement peut être utilisé comme un prétraitement au niveau des codeurs de la parole, mis en place dans les systèmes de radiocommunications de types GSM, 3G, 4G .Ce rehaussement consiste à retirer l'énergie du bruit pour atteindre l'intelligibilité du signal de parole propre. Pour ce faire, plusieurs méthodes de rehaussement de la parole ont va développées dans ce chapitre, les méthodes de rehaussement de la parole peuvent être répartie en trois principales catégories à savoir : Méthodes de soustraction spectrale, Méthodes basés sur des modèles statistiques, Méthodes de décompositions en sous-espace.

3.2 Technique de rehaussement de la parole

Dans cette section, les techniques de rehaussement suivantes sont présentées : rehaussement par soustraction spectrale, par estimeur EQMM spectral, par soustraction spectrale dans le domaine des modulations du spectre, par estimateur EQMM des modulations du spectre et avec la technique de Fusion.

3.2.1 Rehaussement par soustraction spectrale

Une des premières techniques de rehaussement dans le domaine spectral ayant été proposée dans la littérature est le rehaussement par soustraction spectrale. Le but est d'estimer le module du spectre du signal de parole débruité en soustrayant le module du spectre du bruit au module du spectre du signal de parole bruité [32]. Le spectre du bruit pourra, par exemple, être approximé durant les périodes de silence. La définition de l'approche par soustraction spectrale est donnée par

$$\left|\hat{s}[n,l]\right| = \left|X[n,l]\right| - D_u \tag{3.1}$$

Où $D_u$correspond à l'espérance du bruit $D[n,l]$ lors des périodes de silence, soit

$$D_u = E[D[n,l]] \tag{3.2}$$

Un des gros désavantages de la soustraction spectrale telle que présentée est que l'estimation du bruit induit un nouveau bruit qui contient une certaine musicalité, nommé "bruit musical" [33]. Deux améliorations sont proposées dans [33] pour tenter de réduire ce

bruit : premièrement, l'estimation du bruit soustrait est amplifiée par un facteur **α** plus grand que l'unité ; deuxièmement, un seuil minimal est fixé pour éviter que la soustraction cause un résultat plus bas qu'un certain niveau. Ce seuil correspond à une fraction de l'estimé du bruit, représenté par **β**. Ainsi, l'équation (3.1) devient :

$$\left|s[n,l]\right|^{\lambda} = \begin{cases} \left|X[n,l]\right|^{\lambda} - \alpha[l]\hat{D}[n,l]^{\lambda} si \left|X[n,l]\right|^{\lambda} > (\alpha[l]+\beta)\hat{D}[n,l]^{\lambda} \\ \left|\hat{D}[n,l]\right|^{\lambda} cinon \end{cases} \quad (3.3)$$

où **λ** correspond à la puissance et sera expliqué plus bas, tandis que $\boldsymbol{\alpha}[\boldsymbol{l}]$est donné par l'équation linéaire suivante

$$\alpha[l] = \alpha_0 - (RSB_l / S), \; pour \; -5 \le RSB \le 20 \quad (3.4)$$

$\alpha_0[l]$ correspond à la valeur de $\alpha[l]$ désirée pour un RSB de 0 dB, RSB_l correspond au RSB estimé de la trame *l* et 1/s correspond à la pente permettant de réduire la valeur de **α** plus le RSB est grand. Selon les auteurs, cette réduction permet de réduire la distorsion causée par la soustraction.

De manière expérimentale, les auteurs proposent d'avoir un facteur α entre 3 et 6 pour un RSB de 0 dB et un facteur de seuil **β** entre 0.005 et 1. Pour la pente de l'équation de **α**, une valeur recommandée est de 20/3, ce qui permet d'avoir un seuil de 1 pour **α** avec 4:75 comme plafond. La valeur de s est donc dépendante des seuils et plafonds imposés pour **α**.

En plus des deux améliorations proposées ci-haut, une modification supplémentaire a été faite à l'équation (3.1) qui consiste à pouvoir appliquer la soustraction à des valeurs différentes de puissance du spectre, qui est représentée par **λ**. Les auteurs recommandent d'utiliser la puissance du spectre, soit λ= 2. Cette valeur fait aussi consensus dans les autres types de rehaussement sur le spectre du signal, comme il sera vu plus bas. Les valeurs pour **α** sont bonnes uniquement lorsque la puissance du spectre est utilisée pour la soustraction. En cas d'utilisation du module du spectre (donc λ = 1), il est suggéré d'utiliser une valeur de **α** comprise entre 2 et 2.2.

3.2.2 Rehaussement par estimateur spectral EQMM

Il est possible d'obtenir une représentation spectrale en y appliquant une TFD, on obtient alors sous forme polaire :

$$R[n,l] * e^{-\angle X[n,l]} = A[n,l] * e^{-j\alpha[n,l]} + D[nl] \quad (3.5)$$

Où $\boldsymbol{A[n,l]} = |\boldsymbol{S[n,l]}|$, $\boldsymbol{R[n,l]} = |\boldsymbol{X[n,l]}|$ **et** $\boldsymbol{\alpha[n,l]} = \angle \boldsymbol{S[n,l]}$ correspondent au module de la composante spectrale *n* de S et X respectivement de même qu'à la phase de S. Il est

posé que les coefficients A puissent être modélisés statistiquement comme une variable aléatoire gaussienne [34], le problème d'estimation de l'EQMM peut être réduit vers une estimation de A à partir d'un ensemble infini d'observations discrètes. Dans notre cas, ces observations sont les signaux bruités. De plus, comme les composantes spectrales sont considérées statistiquement indépendantes l'une de l'autre, il est nécessaire de disposer seulement de $X[n,l]$ pour obtenir l'estimateur EQMM du module. Ainsi, l'estimateur EQMM $\hat{A}$ de A est

$$\begin{aligned}\widehat{A}[n,l] &= E\left\{A[n,l]\middle|X[n,l]\right\}\\ &= \frac{\int_0^{\infty}\int_0^{2\pi}\alpha[n,l]\,p(|X[n,l]\alpha[n,l]|)\,p(a_k\alpha[n,l])\,d\alpha[n,l]da[n,l]}{\int_0^{\infty}\int_0^{2\pi}p(|X[n,l]\alpha[n,l]|)\,p(a_k\alpha[n,l])\,d\alpha[n,l]da[n,l]}\end{aligned} \tag{3.6}$$

Ou $a[n,l]=A[n,l]$, E{.} correspond à l'espérance statistique et p(.) à la fonction de densité de probabilité (FDP). Sous le modèle statistique gaussien, $p(X|a[n,l],\alpha[n,l])$et $p(a[n,l],\alpha[n,l])$sont respectivement donnés par

$$p(X|a[n,l],\alpha[n,l]) = \frac{1}{\pi\lambda_d[n,l]}\exp\left\{-\frac{1}{\lambda_d[n,l]}\left|X[n,l]-a[n,l]\exp^{j\alpha[nl]}\right|^2\right\} \tag{3.7}$$

$$p(a[n,l],\alpha[n,l]) = \frac{a[n,l]}{\pi\lambda_s[n,l]}\exp\left\{-\frac{a[n,l]^2}{\lambda_s[n,l]}\right\} \tag{3.8}$$

Ou $\lambda_s[n,l]=E\{|S[n,l]|^2\}\ et\ \lambda_d[n,l]=E\{|D[n,l]|^2\}$ sont les variances de la $n^{\text{ième}}$ composante spectrale de la parole et du bruit respectivement. Ensuite en substituant (3.7) et (3.8) dans (3.6), on obtient (voir Appendix A de [34]) :

$$\widehat{A}[n,l] = \Gamma(1.5)\frac{\sqrt{v[n,l]}}{\gamma[n,l]}\exp\left\{-\frac{v[n,l]}{2}\right\}\left[1+v[n,l]\,I_0(\frac{v[n,l]}{2})+v[n,l]\,I_1(\frac{v[n,l]}{2})\right]R[n,l] \tag{3.9}$$

.

Où $\Gamma(.)$ est la fonction gamma avec $\Gamma(.)=\sqrt{\pi}/2$. I0(.) et I1(.) sont respectivement les fonctions modifiées de Bessel d'ordre zéro et un, $v[n,l]$ est défini par

$$v[n,l] = \frac{\varepsilon[n,l]}{1+\varepsilon[n,l]}\gamma[n,l] \tag{3.10}$$

Où $\varepsilon[n,l]$et $\gamma[n,l]$sont respectivement définis par

$$\varepsilon[n,l] = \frac{\lambda_s[n,l]}{\lambda_d[n,l]} \tag{3.11}$$

$$\lambda[n,l] = \frac{R[n,l]^2}{\lambda_d[n,l]} \qquad (3.12)$$

$\mathcal{E}[n,l]$ et $\lambda[n,l]$ sont interprétés comme étant les RSB a priori et a posteriori respectivement. On obtient ensuite le signal rehaussé.

3.2.3 Rehaussement par soustraction spectrale dans le domaine des modulations du spectre

La majeure partie des techniques de rehaussement ayant été développées opère dans le domaine spectral. De plus en plus de recherches physiologiques et psychoacoustiques semblent indiquer que le système auditif humain représente la parole sous forme spectrale, mais également en considérant la modulation de ce spectre. Ces modulations sont obtenues en prenant les variations temporelles des composantes spectrales d'un signal [35]. Les modulations du spectre peuvent aussi être vues comme la variable indépendante de la TFD d'une TFD [36].

Puisqu'il est relativement facile pour le système auditif de déceler un signal de parole dans un environnement bruité, il est tentant d'utiliser une représentation semblable à celle du système auditif afin de faire le rehaussement. Des approches de rehaussement opérant dans le domaine des modulations du spectre ont donc été proposées. Dans une de ces approches, un algorithme de rehaussement de la parole basé sur la soustraction des modulations du spectre est proposé [37].

En prenant le module du spectre du signal de parole$|X[n,l]|$, il est possible d'obtenir les modulations du spectre à partir d'une seconde TFD, ce qui donne

$$\chi[n,l,m] = \sum_{k=0}^{N-1} |X[n,k+lz]| v[k] e^{-j\frac{2\pi mk}{N}} \qquad (3.13)$$

où m, N, Z et v[.] sont respectivement l'index de fréquence des modulations, la taille des trames de cette seconde TFD, l'index de décalage des trames modulatoires et la fonction de fenêtrage.

La soustraction des modulations du spectre est semblable à celle présentée pour la soustraction spectrale à la section 3.3.1. Le spectre modulatoire $\chi[n,l,m]$est converti en forme polaire permettant ainsi d'obtenir le module des modulations du spectre tel que

$$\chi[n,l,m] = |\chi[n,l,m]| e^{j\angle\chi[n,l,m]} \qquad (3.14)$$

Pour obtenir le signal de parole rehaussé, le module du spectre sera $|\chi[n,l,m]|$ remplacé par son estimé rehaussé $|S[n,l,m]|$en utilisant la règle de soustraction. Le spectre rehaussé sera obtenu par

$$|\hat{s}[n,l,m]| = \begin{cases} \left(|\chi[n,l,m]|^{\gamma} - \rho|\widehat{D}[n,l,m]|^{\gamma}\right)^{\frac{1}{\gamma}}, si\ |\chi[n,l,m]|^{\gamma} - \rho|\widehat{D}[n,l,m]|^{\gamma} \geq \beta|\widehat{D}[n,l,m]|^{\gamma} \\ (\beta|\widehat{D}[n,l,m]|^{\gamma})^{\frac{1}{\gamma}} \end{cases}$$

. (3.15)

L'estimation du module du spectre des modulations du bruit $|\check{D}[n,l,m]|$ est obtenue à partir des décisions prises par un algorithme de détection d'activité vocale (DAV), appliqué dans le domaine des modulations du spectre. Le DAV va classifier chaque trame et leur assigner une valeur de 1 en présence de bruit et 0 en cas de silence, en utilisant la règle suivante

$$\Phi[n,l] = \begin{cases} 1, si\ \phi[n,l] \geq \theta \\ 0 cinon \end{cases} \quad (3.16)$$

Où $\Phi[n,l]$ correspond au RSB d'un segment des modulations du spectre défini par

$$\Phi[n,l] = 10\log_{10}\left(\frac{\sum_m |\chi[n,l,m]|^2}{\sum_m |\chi[n,l,m]|^2}\right) \quad (3.17)$$

et où Θ correspond à un seuil de présence de parole déterminé de manière empirique.

L'estimation du bruit est renouvelée selon l'équation de moyennage

$$|\widehat{D}[n,l,m]|^{\gamma} = \lambda|\widehat{D}[n-1,l,m]|^{\gamma} + (1-\lambda)|\chi[n,l,m]|^{\gamma} \quad (3.18)$$

Où λ est un facteur d'oubli défini de façon empirique selon la stationnarité du bruit. Le module rehaussé sera combiné avec la phase non modifiée pour donner l'estimation rehaussée du spectre modulatoire

$$y[n,l,m] = |\hat{s}[n,l,m]|e^{j\angle\chi[n,l,m]} \quad (3.19)$$

Pour obtenir le module du spectre rehaussé, la partie synthèse du modèle AMS est modifiée pour passer du domaine des modulations du spectre au domaine spectral. L'équation de synthèse est donnée par

$$|\hat{s}[n,l]| = \sum_c \left\{ v[l-lz] \sum_{m=0}^{N-1} y[n,l,m] e^{j\frac{2\pi(l-lz)m}{N}} \right\} \quad (3.20)$$

Une fois le module du spectre obtenu, il est utilisé pour obtenir le signal temporel rehaussé.

3.2.4 Rehaussement par estimateur EQMM des modulations du Spectre

Il a été montré dans le domaine spectral que le rehaussement par estimateur EQMM offre de meilleures performances que celui obtenu par la soustraction spectrale. Il a donc été proposé de développer un estimateur EQMM opérant dans le domaine des modulations du spectre plutôt que dans le domaine spectral [38]. De la même façon que pour le domaine spectral, le rehaussement par estimateur EQMM n'opère que sur le module du spectre modulatoire et ne tient pas compte de sa phase. De la même façon que pour l'estimateur obtenu par soustraction dans le domaine des modulations spectrales, une deuxième TFD est appliquée sur le module du spectre pour obtenir le signal donné en (3.18). L'approche EQMM consiste à déterminer une fonction de gain g$[n,l,m]$ permettant de minimiser la différence entre le module du spectre modulatoire de la parole et son estimation. L'estimateur EQMM ainsi obtenu est :

$$\left|\hat{s}[n,l,m]\right| = \varsigma[n,l,m]\left|\chi[n,l,m]\right| \tag{3.21}$$

Où $\left|\hat{S}[n,l,m]\right|$ correspond au module rehaussé et $\varsigma[n,l,m]$ est la fonction de gain

$$\varsigma[n,l,m] = \frac{\sqrt{\pi}}{2}\frac{\sqrt{v[n,l,m]}}{\gamma[n,l,m]}\mathrm{A}[v[n,l,m]] \tag{3.22}$$

Où A[.] correspond à

$$A[\theta] = e^{-\frac{\theta}{2}}\left[(1+\theta)\,I_0\left(\frac{\theta}{\theta}\right) + \theta\, I_1\left(\frac{\theta}{\theta}\right)\right] \tag{3.23}$$

Où I_0(.) et I_1(.) correspondent aux fonctions modifiées de Bessel d'ordre zéro et un respectivement. Ensuite, $v[n,l,m]$ est donné par

$$v[n,l,m] = \frac{\varepsilon[n,l,m]}{1+\varepsilon[n,l,m]}\gamma[n,l,m] \tag{3.24}$$

$\epsilon[n,l,m]$et $\gamma[n,l,m]$ sont respectivement le RSB a priori et a posteriori et sont définis Par

$$\varepsilon[n,l,m] = \frac{E\left[\left|s[n,l,m]\right|^2\right]}{D[n,l,m]^2} \tag{3.25}$$

$$\gamma[n,l,m] = \frac{\left|\chi[n,l,m]\right|^2}{\left|D[n,l,m]\right|^2} \tag{3.26}$$

Comme il n'est pas possible de récupérer le RSB a priori et a posteriori, $\epsilon[n,l,m]$et $\gamma[n,l,m]$sont estimés par la même technique de décision dirigée pour le domaine spectral décrit en [34], mais modifiée pour être appliquée dans le domaine des modulations de spectre. L'estimé de $\epsilon[n,l,m]$est donné par

$$\hat{E}[n,l,m] = \alpha \frac{\left|\hat{s}[n,l,m-1]\right|^2}{\lambda[n,l,m-1]} + (1-\alpha)\max[\hat{y}[n,l,m]-1,0] \tag{3.27}$$

Où α est un paramètre empirique permettant de choisir le compromis entre la réduction de bruit et la distorsion transitoire du signal et où $\lambda[n,l,m]$ est l'espérance de l'énergie du spectre modulatoire de l'estimation du bruit. Ensuite, le RSB a posteriori est estimé par

$$\hat{y}[n,l,m] = \frac{\left|\chi[n,l,m]\right|^2}{\lambda[n,l,m]} \tag{3.28}$$

La limitation de la valeur minimale du RSB a priori ajoute un résidu de bruit après rehaussement, ce qui cause une distorsion non négligeable [34] [39]. Pour pallier à ce problème, un seuil est déterminé pour éviter d'obtenir une valeur causant ces distorsions. Le RSB a priori est donc donné par

$$\hat{y}[n,l,m] = \max\left\lfloor \hat{y}[n,l,m], \gamma_{\min} \right\rfloor \tag{3.29}$$

Où γ_{min} corresponds au seuil minimal à atteindre. Une valeur empirique de -25 dB est utilisée pour ce seuil, sans précision par les auteurs de la démarche ayant permis d'atteindre cette valeur. Finalement, le module du spectre modulatoire rehaussé sera utilisé dans l'équation (3.19) qui donnera le spectre rehaussé, permettant d'obtenir le signal de parole rehaussé.

3.2.5 Rehaussement avec la technique de Fusion

Les travaux en [38] proposent une méthode de rehaussement combinant le rehaussement fait dans le domaine spectral avec celui fait dans le domaine des modulations du spectre. Cette combinaison se fait avec une fonction de poids permettant d'utiliser un domaine plus que l'autre selon le RSB a posteriori de la trame à rehausser. Cette

combinaison se fait dans le domaine spectral, ce qui veut dire que le rehaussement fait dans le domaine des modulations du spectre doit être converti vers le domaine spectral avant de pouvoir être utilisé. La fonction de poids est donnée par

$$|\hat{s}[n,l]| = \left(\Psi[\omega[l]] \left| \hat{S}_{spectral}[n,l] \right|^{\lambda} + (1-\psi[\omega[l]]) \left| \hat{S}_{\text{mod } ulatoire}[n,l] \right|^{\lambda} \right)^{\frac{1}{\lambda}} \tag{3.30}$$

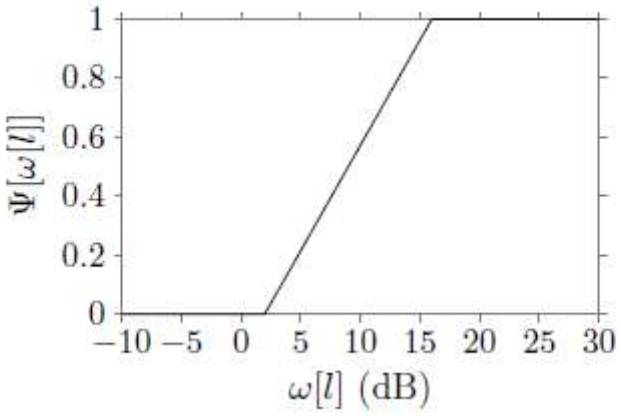


Figure.3.1 Fonction de poids, $\Psi[\omega[l]]$

Où le module du spectre rehaussé $|\hat{S}[n,l,]|$ est obtenu par une combinaison du module du spectre rehaussé $\hat{S}_{spectral}[n,l]$ et du module des modulations spectrales $\hat{S}_{modulatoire}[n,l].\Psi[\omega[l]]$ correspond au facteur de poids donné pour la combinaison e $\Psi[l]$correspond au RSB de la trame l. $\Psi[\omega[l]]$ est donné par

$$\psi[\omega[l]] = \begin{cases} 0si, g[\omega[l]] \le 2 \\ \dfrac{g[\omega[l]-2}{14} si, 2\langle g[\omega[l]]\langle 16 \\ 1si, g[\omega[l]] \ge 16 \end{cases} \tag{3.31}$$

Où $g[\omega[l]] = \log_{10}[\omega[l]]$ et où $[\omega[l]]$correspond au RSB après rehaussement. La fonction de poids est illustrée à la fig. 3.1. Les trames ayant un RSB plus bas que 2 dB seront rehaussées uniquement dans le domaine des modulations du spectre, celles ayant un RSB plus grand que 16 dB seront rehaussés uniquement dans le domaine spectral tandis que les trames entre ces deux limites seront rehaussées par une combinaison des deux domaines.

La technique de rehaussement utilisée pour le domaine spectral pour obtenir $\hat{S}_{spectral}[n,l]$ dans [38] est l'algorithme de rehaussement par estimateur EQMM du module du spectre tandis que l'algorithme pour le rehaussement du module des modulations du spectre $\hat{S}_{modulatoire}[n,l]$ est celui du rehaussement par soustraction spectrale.

Cette technique de combinaison offre une amélioration des performances étant donné que les résultats avec le rehaussement spectral sont meilleurs que ceux avec le rehaussement des modulations du spectre à fort RSB tandis qu'à faible RSB, le rehaussement des modulations du spectre performe mieux que le rehaussement spectral. Le tableau 2.1 ci-dessous contient les résultats obtenus par les auteurs de [37] pour la comparaison de leur algorithme avec le rehaussement spectral et des modulations du spectre. Les performances sont mesurées à partir du score PESQ.

SNR (dB)	*MMSE*	*ModSpecSub*	*SpecSub*	*Fusion*
0	2.00	2.22	1.76	2.25
5	2.10	2.49	2.18	2.63
10	2.62	2.76	2.62	2.79
15	2.97	3.02	3.01	3.05

Tableau 3.1 Score PESQ avec un bruit aditif blanc gaussien (BABG).

Le terme MMSE est la traduction anglaise de EQMM (Minimum Mean Square Error). Il correspond à la technique présentée en section 3.3.2. Le terme ModSpecSub correspond à la technique présentée en section 3.3.3. La technique SpecSub est décrite à la section 3.3.1 et Fusion est la technique de rehaussement utilisant la fonction de combinaison présentée ci-dessus.

3.2.6. Méthodes de rehaussement utilisant des réseaux de neurones profond

Le réseau de neurones profond (DNN) est un nouveau domaine de recherche qui a connu une énorme croissance en popularité ces dernières années.

Les méthodes de rehaussement de la parole utilisant des réseaux de neurones profonds ont d'abord eu un succès notable dans le contexte monocanal. Elles s'appuient sur différentes architectures de réseau, basées sur des couches totalement connectées (FF, feed-forward)

[40] ou RNN [41], dont LSTM (long short-term memory) [42,43] et LSTM bidirectionnelles (biLSTM) [44], le réseau peut être entraîné à reconstruire le signal source lui-même [45], parfois de façon générative [46], ou bien un masque qui permet d'estimer la proportion de signal cible présente dans le mélange en chaque point temps-fréquence [47]. Nous nous intéressons ci-dessous plus en détail à l'utilisation de réseaux de neurones dans le cadre multicanal.

Rehaussement d'un locuteur seul

L'information multicanale a été utilisée sous diverses formes pour faciliter l'apprentissage du réseau : en combinant l'ITD et l'ILD à une information monocanale [48, 45] ; en comparant les différences de phase effectives et celles, théoriques, correspondant à la direction supposée connue de la source d'intérêt [49] ; ou en mesurant le caractère diffus du signal [50]. Bien que certains travaux cherchent à estimer directement le signal cible s(t; f) en sortie du réseau [45], la plupart reposent sur l'estimation de masques temps-fréquence comme étape intermédiaire [51]. Ces masques peuvent être appliqués tels quels à chaque canal du signal [48], mais ils induisent alors des distorsions défavorables à la RAP.

Filtrage multicanal utilisant un masque temps-fréquence Pour éviter les distortions résultant d'un traitement monocanal, il a été proposé d'utiliser les masques estimés par un réseau de neurones pour l'estimation de filtres multicanaux [52,53]. Le principe de cette approche est présenté dans la Figure 3.2 : le masque $\boldsymbol{M_s = (t, f)}$ permet d'estimer le signal cible ***S(t, f)*** et le signal interférent ***n(t, f)*** :

$$\hat{s}(t, f) = \boldsymbol{M}_s(t, f)x(t, f) \quad (3.32)$$

$$\hat{n}(t, f) = \left(1 - \boldsymbol{M}_s(t, f)x(t, f)\right) \quad (3.33)$$

Les matrices de covariance peuvent ensuite être estimées en moyenne sur un intervalle de temps

$$\boldsymbol{R}_{ss}(f) = \sum_t \hat{s}(t, f)\,\widehat{\boldsymbol{s}}(\boldsymbol{t}, \boldsymbol{f})^H \quad (3.34)$$

$$\boldsymbol{R}_{nn}(f) = \sum_t \hat{n}(t, f)\,\widehat{\boldsymbol{n}}(\boldsymbol{t}, \boldsymbol{f})^H \quad (3.35)$$

et être utilisées pour calculer un des filtres, qui est alors invariant sur cet intervalle de temps. Certains travaux calculent un seul masque pour l'ensemble des canaux [54]. D'autres estiment un masque par canal, puis un unique masque médian est appliqué à tous les canaux pour plus de robustesse [55].



Figure3.2 Filtrage multicanal utilisant des masques temps-fréquence estimés par un réseau de neurones.

Cette formulation permet de relâcher les hypothèses de stationnarité du bruit et d'indépendance des signaux, ainsi que la nécessité d'avoir accès à une estimation préalable de l'activité vocale, conditions auparavant indispensables à l'utilisation des filtres . Elle a permis d'améliorer significativement les performances de rehaussement d'une source de parole dans un bruit diffus, y compris en environnements réels [56].

Estimation directe des paramètres de filtrage Enfin, certaines méthodes proposent d'utiliser les réseaux de neurones pour estimer directement les paramètres des filtres, sans passer par le calcul des matrices de covariance. Par exemple, Xiao et al. [57] optimisent directement les coefficients des filtres selon des critères de RAP.

Nugraha et al. [58] utilisent des DNNs pour estimer les spectres des signaux sur lesquels repose le MWF dans le cadre probabiliste gaussien, permettant le calcul de filtres multicanaux variant dans le temps.

3.3 Rehaussement de la parole et l'apprentissage profond

Pour que ces différentes techniques de traitement de la parole puissent fonctionner correctement, il est important de disposer en entrée d'un signal de parole clair et intelligible. Or, dans la pratique, la parole est acquise en présence de bruits, qui viennent dégrader la performance de ces techniques et qui peut même les rendre non fonctionnelles. Il est donc nécessaire de devoir rehausser cette parole bruitée pour que ces techniques puissent être utilisées de manière efficace. Ce rehaussement consiste à retirer l'énergie du bruit pour atteindre l'intelligibilité du signal de parole propre. Pour ce faire, plusieurs méthodes de rehaussement de la parole ont été développées afin de rendre les techniques de traitement de la parole fonctionnelles et performantes. Les premières techniques développées consistaient à supprimer le bruit en utilisant la soustraction spectrale [32]. Des techniques plus complexes, en utilisant des estimateurs statistiques sur le spectre [34] et plus récemment sur la modulation du spectre [38] [37], ont permis d'améliorer la capacité à réduire le bruit dans un segment de parole.

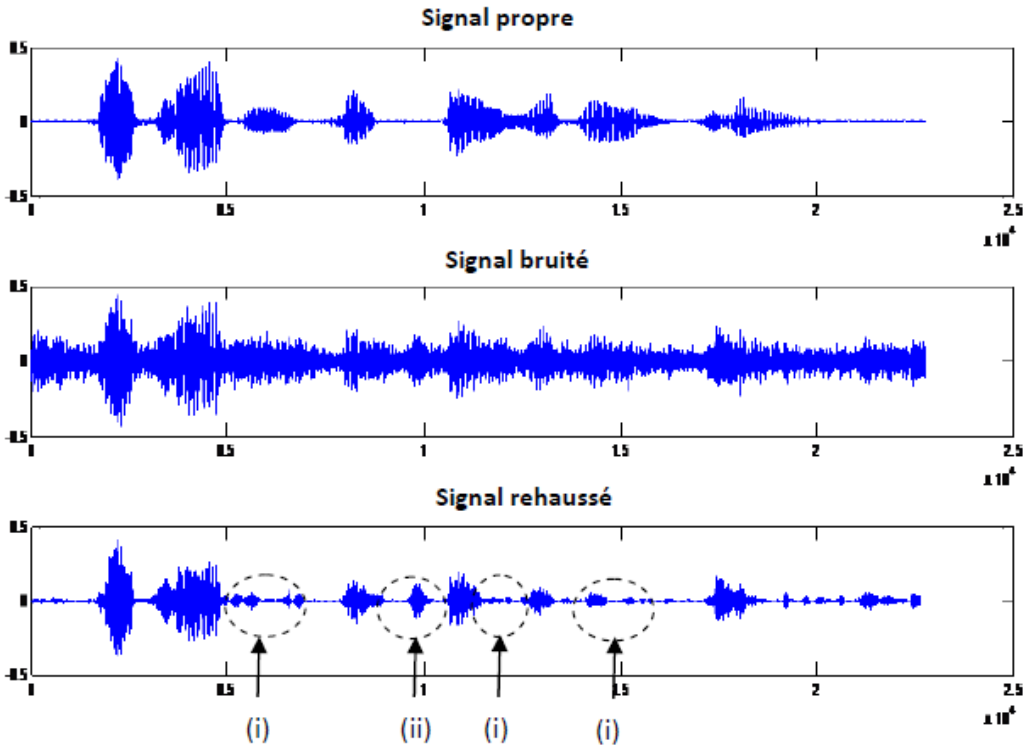


Figure 3.3 Signal propre, corrompu par un bruit et rehaussé.

3.4 Conclusion

Ce chapitre a présenté une vue d'ensemble sur le rehaussement de la parole. Dans ce chapitre nous avons présenté plusieurs méthodes de rehaussement de la parole : rehaussement par soustraction spectrale, rehaussement par estimateur spectral EQMM, rehaussement par soustraction spectrale dans le domaine des modulations du spectre, rehaussement par estimateur EQMM des modulations du Spectre et le rehaussement avec la technique de Fusion.

Chapitre 4
Expériences et résultats

4.1. Introduction

Dans ce chapitre, nous allons présenter les expériences et les résultats obtenus après l'évaluation de notre système de reconnaissance du locuteur. Notre objectif étant de développer un système de reconnaissance du locuteur basé sur l'approche DNN-i-vecteur dans un environnement mobile.

4.2. Logiciels et langages utilisés dans l'application:

a) Logiciel Matlab:

Nous avons utilisé un logiciel MATLAB (matrix laboratory) version 16.02 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.

4.3 Description des bases de données utilisées

Les expériences ont été effectuées en utilisant les bases de données NIST-2008 et la base de bruit standard.

a) La base de données NIST:

L'évaluation de système de reconnaissance du locuteur fait une partie des séries continue 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 attaques à l'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 300 locuteurs à partir de cette base. La base de données est passée ensuite à travers le codec standard AMRWB pour aboutir à la base de données transcodée mobile.

b) La base de données NOISEX-92

NOISEX-92 est une base de données de bruits standards. Elle fournit différents bruits enregistrées militaire dans des environnements réels avec de taux d'échantillonnage 19.9KHz.

4.4. Étapes de la réalisation des systèmes de reconnaissance du locuteur:

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

- Créer l'ensemble d'apprentissage: la base de données que nous avons utilisée est constituée de 1500 fichiers, prononcés par 300 locuteurs de différents sexes qui répètent la même phrase 5 fois.
- Effectuer l'analyse acoustique: l'ensemble des fichiers de la base de données ont été convertis en une série de vecteurs.
- Utilisation le DNN entre le signal propre et signal bruité
- Faire l'apprentissage des modèles de mélanges de lois gaussiennes: il s'agira de construire le modèle du monde « GMM-UBM » et le modèle du locuteur cible.
- L'apprentissage d'un sous-espace de variabilité totale à partir des statistiques de base « GMM-UBM».
- Créer un modèle GPLDA Gaussian 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 reconnaissance du locuteur

4.5. La structure générale de Système de reconnaissance du locuteur:

La structure générale d'un système de reconnaissance automatique du locuteur est représentée sur la (Figure 4 .1).

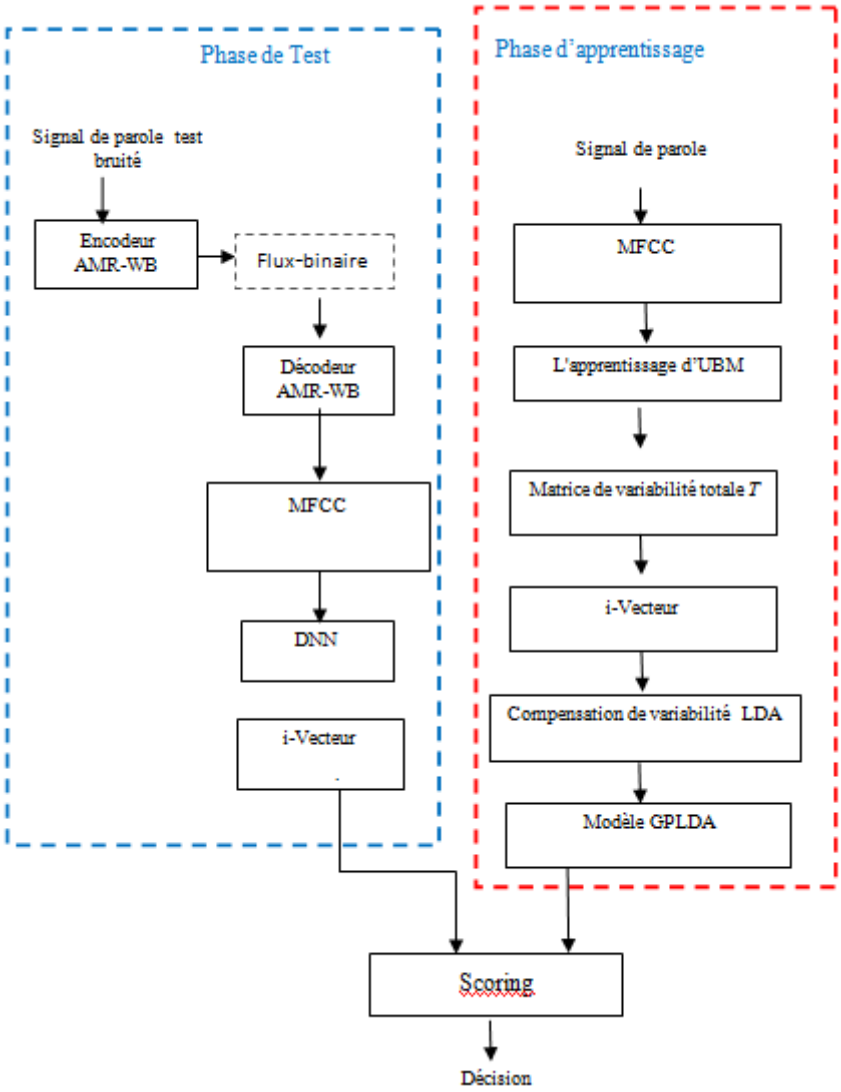


Figure 4.1: Structure générale d'un système RAL.

4.6. Extraction des paramètres

4.6.1. Paramétrisation des coefficients MFCCs

Le signal de parole est paramétré en utilisant les coefficients MFCCs. Ces paramètres sont extraits toutes les 10ms, sur des fenêtres de 25 ms. La fenêtre de pondération utilisée est la fenêtre de Hamming et puis appliquer Douze (12) coefficients MFCCS, ils sont calculés pour chaque trame dans l'échelle fréquentielle Mel. Les coefficients différentiels du premier et du second ordre (ΔΔ), sont ensuite calculés pour former un vecteur de dimension 36 (12MFCC + 12 Δ MFCC+ 12 Δ Δ MFCC).

4.6.2. Phase d'apprentissage

Cette phase consiste en la génération des modèles acoustiques du locuteur (les modèles GPLDA) à partir de la matrice des variabilités totales (i-vecteur) et l'apprentissage de l'UBM à 512 composantes.

4.6.3. Phase de test

la phase de test via les modèles GPLDA consiste à évaluer un score mesurant le degré de la ressemblance entre i-vecteur test et le modèle GPLDA .

4.7. Expériences et résultats

Pour évaluer la robustesse des systèmes de reconnaissance du locuteur aux conditions bruitées. Nous avons réalisé une série des expériences de reconnaissance du locuteur pour trois types de bruits (stationnaire et non-stationnaire) qui sont respectivement: bruit de la foule bruit de voiture et bruit d'usine. Pour l'entrainement des modèles acoustiques et des modèles additifs a été réalisé avec la base de données NIST 2008 non bruitée, le bruit a été seulement ajouté pour la base de test : (Babble, Car et Factory) issus de la base de données de bruits standards NOISEX'92, en utilisant les niveaux SNR suivants : -5dB, 0dB, 5dB, 10dB. Nous implémentons notre système de reconnaissance du locuteur en utilisant la boîte à outils d'identité MSR.

4.7.1. Expérience 1

Dans la première expérience, nous présentons les performances des systèmes de reconnaissance en utilisant les données d'apprentissage propres et bruitées avec le bruit de foule à différents niveaux de SNR (-5, 0, 5, 10 dB). Les résultats sont présentés dans le **Tableau.4.1**

Tableau. 4.1 Performance des systèmes i-vecteur-GPLDA et DNN-i-vecteur en utilisant de données de test bruitées pour le bruit de foule

	EER (%)			
SNR	**-5dB**	**0dB**	**5dB**	**10dB**
i-vecteur GPLDA	15.26	12.37	9.30	7.87
DNN-i-vecteur	11.79	8.92	7.70	5.56

4.7.2. Expérience 2

Dans la deuxième expérience, nous présentons les performances des systèmes de reconnaissance en utilisant les données d'apprentissage propres et bruitées avec le bruit de voiture à différents niveaux de SNR (-5, 0, 5, 10 dB). Les résultats sont présentés dans le **Tableau.4.2**

Tableau. 4.2 Performance des systèmes i-vecteur-GPLDA et DNN-i-vecteur en utilisant de données de test bruitées pour le bruit de voiture

	EER (%)			
SNR	**-5dB**	**0dB**	**5dB**	**10dB**
i-vecteur GPLDA	19.25	14.89	11.20	9.23
DNN-i-vecteur	16.87	12.45	10.40	8.62

4.7.3. Expérience 3

Dans la troisième expérience, nous présentons les performances des systèmes de reconnaissance en utilisant les données d'apprentissage propres et bruitées avec le bruit d'usine à différents niveaux de SNR (-5, 0, 5, 10 dB). Les résultats sont présentés dans le **Tableau.4.3**

Tableau. 4.3 Performance des systèmes i-vecteur-GPLDA et DNN-i-vecteur en utilisant de données de test bruitées pour le bruit d'usine

	EER (%)			
SNR	**-5dB**	**0dB**	**5dB**	**10dB**
i-vecteur GPLDA	25.45	22.65	18.45	16.10
DNN-i-vecteur	24.67	20.23	17.62	13.34

Les résultats dans les **Tableau.4.1, Tableau 4.2 et Tableau 4.3** montrent que le système DNN-i-vecteur donne de meilleurs résultats de reconnaissance en terme EER (%) par rapport au système i-vecteur-PLDA à tous les niveaux de SNR. Ceci peut s'expliquer par le fait que l'approche DNN inclue dans la partie du signal de parole bruité permettent d'estimer d'une manière plus robuste le spectre de la parole dans des conditions bruitées.

4.8. Conclusion

Dans ce chapitre de mise en œuvre deux systèmes de reconnaissance du locuteur basé sur i-vecteur-GPLDA et DNN-i-vecteur pour réduire l'effet de bruit dans le signal transmis d'un environnement mobile. Des expériences en milieu bruite ont été réalisées, et d'après l'analyse des résultats obtenus montre que l'utilisation du système de reconnaissance DNN-i vecteur donne de meilleurs résultats par rapport aux paramètres au système de base i-vecteur DPLDA en terme de taux d'erreur EER.

Conclusion Générale

Dans le cadre de ce travail, nous avons abordé le problème de la robustesse des systèmes de reconnaissance du locuteur dans l'environnement mobile bruité et rehaussement de la parole dégradé par un bruité additif prélevé à partir de la base de données NOISEX-92. Nous nous sommes intéressés en particulier au module d'extraction des paramètres acoustiques dans le système de reconnaissance du locuteur pour améliorer sa performance.

Au coure de ce travail, nous avons proposé une technique de rehaussement de la parole est basé sur les réseaux de neurone profond. L'évaluation de cette technique a montré leur robustesse, nous avons noté une nette amélioration de la performance du système dans les différentes conditions réelles.

Il serait très intéressant de voir dans des travaux futurs, l'étude des effets relatifs au canal de transmission n'a pas été prise en considération dans ce travail.

[1] K.Ould Ammar, M.Benzeghimi. Annulation d'Echo Acoustique par Les Algorithmes Rapides. Mémoire de Master, Université Blida 1. 2011-2012.

[2] Y.Aziza. Modélisation AR et ARMA de la parole pour une vérification robuste du locuteur dans un milieu bruité en mode dépendant du texte. Mémoire de Magister. Université FERHAT ABBAS Setif.2013.

[3] TIWARI, Vibha. MFCC and its applications in speaker recognition. International Journal on Emerging Technologies, 2010, vol. 1, no 1, p. 19-22.

[4] IMPEDOVO, Donato et REFICE, Mario. Frame length selection in speaker verification task. Transaction on Systems, 2008, vol. 7, no 10, p. 1028-1037.

[5] MEIGNIER, S. Indexation en locuteurs de documents sonores: Segmentation d'un document et Appariement d'une collection. These de doctorat, Université d'Avignon et des Pays de Vaucluse, 2002.

[6] MEIGNIER, Sylvain, BONASTRE, Jean-Françcois, FREDOUILLE, Corinne, et al. Modèle de Markov évolutif pour les tâches de suivi de locuteurs. JEP, 2000, p. 69-72.

[7] SAYOUD, Halim. Reconnaissance automatique du locuteur approche connexionniste. 2003. Thèse de doctorat.

[8] BLOUET, Raphaël. Approche probabiliste par arbres de décision pour la vérification automatique du locuteur sur architectures embarquées. 2002. Thèse de doctorat. Rennes 1.

[9] BOCCARDI, Federico et DRIOLI, Carlo. Sound morphing with Gaussian mixture models. In : Proc. DAFx. 2001. p. 44-48.

[10] BOITE, René. Traitement de la parole. PPUR presses polytechniques, 2000.

[11] JOUSSE, Vincent. Identification nommée du locuteur: exploitation conjointe du signal sonore et de sa transcription. 2011. Thèse de doctorat. Université du Maine.

[12]Vapnik, V.N. (1998). « Statistical Learning Theory. » 1st ed. Wiley.

[13] Kenny, P. Ouellet, P. Dehak, N. Gupta, V. et Dumouchel, P. (2008). « Study of Inter-Speaker Variability in Speaker Verification. » IEEE Transactions on Audio, Speech and Language Processing.

[14] Bishop CM. (2007). « Pattern Recognition and Machine Learning. » Information Science and Statistics. 1st ed. Springer.

[15] Prince, S. J. D. et Elder, J. H. (2007). « Probabilistic Linear Discriminant Analysis for Inferences about Identity. » in Proc. 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, pp. 1–8.

[16] Kenny, P. Reynolds, D. et Castaldo, F. (2010b). « Diarization of Telephone Conversations using Factor Analysis. » IEEE Journal of Selected Topics in Signal Processing.

[17] Senoussaoui, M. Kenny, P. Brummer, N. de Villiers, E. et Dumouchel, P. (2011a). « Mixture of PLDA Models in I-Vector Space for Gender-Independent Speaker Recognition. » Proc. Interspeech 2011, Florence.

[18] Laflamme, C., et al. "On reducing computational complexity of codebook search in CELP coder through the use of algebraic codes." Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on. IEEE, 1990.

[19] ALENCAR, Marcelo S. et DA ROCHA, Valdemar C. Communication systems. Springer Science & Business Media, 2005.

[20] Lefort L., Merlin T., Bonastre J.-F. et Nocera P., Le projet MTM reconnaissance de la parole et du locuteur sur une plateforme embarqu´ee, dans XXIV`emes Journ´ees d'Etudes sur la Parole JEP'2002, pages 265–268, Nancy, France, 2002.

[21] Cornu E., Destrez N., Dufaux A., Sheikhzadeh H. et R.Brennan, An ultra low power, ultra miniature voice command system based on hidden markov models, dans Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2002), tome 4, pages 3800–3803, Orlando, Florida, USA, 2002.

[22] Wang D., Zhang L., Liu J. et Liu R., Emnedded speech recognition system on 8 bits MCU core, dans Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2004), tome 5, pages 301–304, Montreal, Canada, 2004.

[23] Deligne S., Eide E., Gopinath R.-A., Kanevksy D., Maison B., Olsen P., Printz H. et Sedivy J., Low-resource speech recognition of 500-word vocabularies, dans Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech'2001), pages 1833–1836, Aalborg, Denmark, 2001.

[24] Obermaier B. et Rinner B., A tms320c40 based speech recognition system for embedded applications, dans The 2nd European DSP Education & Research Conference, Paris, France, 1998.

[25] Astrov S. et Andrassy B., Large vocabulary speaker independent isolated word recognition for embedded systems, dans Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech' 2003), pages 1137–1140, Geneva, Switzerland, 2003.

[26] Astrov S., Bauer J.-G. et Stan S., High performance speaker and vocabulary independent ASR technology for mobile phones, dans Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2003), pages 281–284, Hong Kong, 2003.

[27] Gong Y. et Kao Y.-H., Implementing a hich accuracy speakerindependent continuous speech recognize on a fixed-point, dans Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2000), tome 6, pages 3686–3689, Istanbul, Turkey, 2000.

[28] Ishikawa S., Yamabana K., Isotani R. et Okumura A., Parallel LVCSR algorithm for cellphone-oriented multicore processor, dans Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2006), pages 177–180, Toulouse, France, 2006.

[29] Maes S.-H., Cohen G., Hoory R. et Chazan D., Conversational networking: Conversational protocols for transport, coding and control, dans Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000), tome 2, pages 198–201, Beijing, China, 2000.

[30] Ramaswamy G.-N. et Gopalakrishnan P.-S., Compression of acoustic features for speech recognition in network environments, dans Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1998), tome 2, pages 977–980, Seattle, Washington, USA, 1998.

[31] Srinivasamurthy N., Ortega A. et Narayanan S., Efficient scalable encoding for distributed speech recognition, IEEE Transactions on Speech and Audio Processing, 2005.

[32] Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech and Signal Processing, volume 27, numéro 2, p. 113–120.

[33] Berouti, M., Schwartz, R. et Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. Dans IEEE International Conference on Acoustics, Speech and Signal Processing, volume 4. p. 208–211.

[34] Ephraim, Y. et Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics,
Speech and Signal Processing, volume 32, numéro 6, p. 1109–1121.

[35] Zadeh, L. (1950). Frequency analysis of variable networks. Proceedings of the IRE, volume 38, numéro 3, p. 291–299.

[36] Atlas, L. et Shamma, S. A. (2003). Joint acoustic and modulation frequency. EURASIP Journal on Advances in Signal Processing, volume 2003, numéro 7, p. 1–8.

[37] Paliwal, K., Wojcicki, K. et Schwerin, B. (2010). Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Communication, volume 52, numéro 5, p. 450–475.

[38] Paliwal, K., Schwerin, B. et Wojcicki, K. (2012). Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. Speech Communication, volume 54, numéro 2, p. 282–305.

[39] Cappe, O. (1994). Elimination of the musical noise phenomenon with the Ephraim and malah noise suppressor. IEEE Transactions on Acoustics, Speech and Signal Processing, volume 2, numéro 2, p. 345–349.

[40] Y. Wang and D. Wang. Towards scaling up classi_cation-based speech separation. IEEE TASLP, 21(7) :1381_1390, 2013.

[41] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Deep learning for monaural speech separation. In Proc. of ICASSP, pages 1562_1566, 2014.

[42] F. Weninger, F. Eyben, and B. Schuller. Single-channel speech separation with memory-enhanced recurrent neural networks. In Proc. of ICASSP, pages 3709_ 3713, 2014.

[43] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller. Discriminatively trained recurrent neural networks for single-channel speech separation. In Proc. Of GlobalSIP, pages 577_581, 2014.

[44] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In Proc. of ICASSP, pages 708_712, 2015.

[45] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, et al. Exploring multi-channel features for denoising-autoencoder-based speech enhancement. In Proc. of ICASSP, pages 116_120, 2015.

[46] S. Pascual, A. Bonafonte, and J. Serrà. SEGAN : speech enhancement generative adversarial network. In Proc. Interspeech, pages 3642_3646, 2017.

[47] A. Narayanan and D.Wang. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In Proc. of ICASSP, pages 7092_7096, 2013.

[48] Y. Jiang, D. Wang, R. Liu, and Z. Feng. Binaural classi_cation for reverberant speech segregation using deep neural networks. IEEE TASLP, 22(12) :2112_2121, 2014.

[49] P. Pertilä and J. Nikunen. Microphone array post-_ltering using supervised machine learning for speech enhancement. In Proc. of Interspeech, pages 2675_2679, 2014.

[50] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann. Spatial di_useness features for DNN-based speech recognition in noisy and reverberant environments. In Proc. of ICASSP, pages 4380_4384, 2015.

[51] Y. Wang, A. Narayanan, and D. Wang. On training targets for supervised speech separation. IEEE TASLP, 22(12) :1849_1858, 2014.

[52] J. Heymann, L. Drude, and R. Haeb-Umbach. Neural network based spectral mask estimation for acoustic beamforming. In Proc. of ICASSP, pages 196_200, 2016.

[53] L. Perotin, R. Serizel, E. Vincent, and A. Guérin. Multichannel speech separation with recurrent neural networks from high-order Ambisonics recordings. In Proc. Of ICASSP, pages 36_40, 2018.

[54] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani. Robust MVDR beamforming using time-frequency masks for online/o_ine ASR in noise. In Proc. of ICASSP,pages 5210_5214, 2016.

[55] K. .molíková, M. Delcroix, K. Kinoshita, T. Higuchi, et al. Speaker-aware neural network based beamformer for speaker extraction in speech mixtures. In Proc. Of Interspeech, pages 2655_2659, 2017.

[56] Z. Wang, E. Vincent, R. Serizel, and Y. Yan. Rank-1 constrained multichannel Wiener _lter for speech recognition in noisy environments. Computer Speech & Language, 49 :37_51, 2018.

[57] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, et al. Deep beamforming networks for multi-channel speech recognition. In Proc. of ICASSP, pages 5745_5749, 2016.

[58] A. A. Nugraha, A. Liutkus, and E. Vincent. Multichannel audio source separation with deep neural networks. IEEE TASLP, 24(9) :1652_1664, 2016.