UNIVERSITE BLIDA-1

Faculté de Technologie

THESE DE DOCTORAT

en Electronique

CONTRIBUTION A L'INFERENCE D'IDENTITE EN UTILISANT UN SYSTEME DE RECONNAISSANCE DU LOCUTEUR GMM-UBM.

Par

Abdennour ALIMOHAD

devant le jury composé de :

A. BENALLAL	Professeur, U. de Blida	Président
D. BERKANI	Professeur ENP, Alger	Examinateur
M. OULDZMIRLI	MCA, Université de Médéa	Examinateur
Z.A BENSELAMA	MCA, Université de Blida	Examinateur
A. GUESSOUM	Professeur, U. de Blida	Directeur de thèse
A. BOURIDANE	Professeur, U. Northumbria, UK	Co-Directeur de thèse

Blida, septembre 2015

الملخص

جذب التعرف على المتكلم الاهتمام من عدد من الباحثين، وهذا لتصميم النظام الأكثر كفاءة من أجل تحديد بدقة هوية الشخص من خلال صوته. لكن تبين في الواقع أن المهمة ليست دائما سهلة خاصة في وجود أنواع مختلفة من التقلبات بسبب البيئة و أدوات التسجيل أو عوامل أخرى. وفي مثل هذه الظروف يصبح نظام التعرف غير فعال. يدخل عمل هذه الأطروحة في إطار التعويض عن هذه التغيرات التي تحدث جراء كون طريقة التدريب والاختبار ليست نفسها و لهذا تم اعتماد طريقة المعاملات الثابتة.

بدأنا من خلال دراسة هذه المعايير والعملية التي أدت لاستخراجها ثم قمنا بإدخال بعض التغييرات عليها من أجل استخدامها في سياق التعرف على المتكلم. أجرينا تجارب من أجل التحقق من صحة نهجنا واختبار أدائها من خلال مقارنتها بالطرق التقليدية مثل المعاملات MFCC أو PLP . النتائج التي تم الحصول عليها سمحت لنا بالتأكد من توقعاتنا، بحيث أن المعاملات الثابتة قابلة للتطبيق في أنظمة التعرف على المتحدث، كما كان أداءها أفضل من الطرق التقليدية في الحالات العادية (التطابق بين التدريب والاختبار) أو وجود التغيرات (عدم تطابق).

لوحة المفاتيح : التعرف على المتكلم معاملات MFCC ،معاملات PLP ، المعاملات الثابتة ، عدم تطابق

RESUME

La reconnaissance du locuteur a suscité l'intérêt d'un certain nombre de chercheurs, afin de concevoir le système le plus performant en termes de détermination, avec exactitude, de l'identité d'une personne à travers sa voix. En réalité, cette tâche de reconnaissance n'est pas toujours facile en présence de différents types de variabilités, dues à l'environnement, aux moyens d'enregistrement, ou à d'autres facteurs. Dans ces conditions, la reconnaissance du locuteur devient entachée d'erreurs et le système de reconnaissance est peu performant.

Notre travail de thèse s'inscrit dans le cadre de la compensation de la variabilité de session, pour laquelle les conditions d'apprentissage et de test ne sont pas les mêmes. Pour cela, les paramètres invariants ont été adoptés. Nous avons commencé par étudier ces paramètres et le processus menant à les extraire. Nous avons, ensuite procédé à quelques modifications de ces paramètres pour pouvoir les utiliser dans notre contexte de reconnaissance de locuteurs.

Des expérimentations ont été effectuées afin de valider notre approche et tester ses performances en la comparant aux méthodes classiques comme les coefficients MFCC ou PLP. Les résultats obtenus nous ont permis de confirmer nos prédictions, puisque les paramètres invariants sont applicables dans les systèmes de reconnaissance du locuteur et sont plus performants que les méthodes classiques, dans les cas normaux (correspondance entre l'apprentissage et le test) ou en présence de variabilité (non correspondance).

Mots clés : Reconnaissance du locuteur, Coefficients MFCC, Coefficients PLP, Paramètres invariants, Non correspondance.

ABSTRACT

Speaker recognition has attracted interest from a number of researchers, to design the most efficient system in terms of determination, accurately, of person identity through his voice. However, in reality this recognition task is not always easy in presence of different types of variability due to environment, recording tools, or other factors. In these conditions, the speaker recognition becomes tainted by errors and the recognition system is inefficient.

Our thesis work falls in the session variability compensation framework, for which the training and testing conditions are not the same. For this, invariant features were adopted. We have started by studying these features and the process leading to extracting them. Then, we proceed to some modifications on these parameters in order to use them in speaker recognition context.

Experiments were performed to validate our approach and to test its performance by comparing it to classical methods as MFCC or PLP coefficients. Obtained results confirmed our predictions, since invariant features are applicable in speaker recognition systems and are more efficient then the classical methods, in normal case (match between train and test) or in presence of variability (mismatch case).

Keywords : Speaker recognition, MFCC Coefficients, PLP Coefficients, Invariant features, mismatch.

REMERCIEMENTS

Je commence mes mots, par remercier ALLAH le tout puissant, pour m'avoir aidé et guidé pour atteindre ce point et terminer ce travail de doctorat.

Puis, je tiens à remercier mon directeur de thèse, le professeur Abderrezak GUESSOUM, sans lequel cette thèse n'aurait pas pu être concrétisée. Je le remercie pour ses conseils, sa gentillesse et sa disponibilité durant toutes ces longues années.

Mes sincères remerciements vont à mon co-directeur de thèse, le professeur Ahmed BOURIDANE, pour son aide continue, sa disponibilité, et son support ici en Algérie ou en Angleterre où il m'a si bien accueilli dans son laboratoire. C'est grâce à ce séjour à l'université Northumbria de Newcastle Upon Tyne en grande Bretagne que j'ai pu faire l'essentiel de ce travail de thèse.

Je remercie profondément, le professeur Ahmed BENALLAL en sa qualité de président de ce jury, les membres du jury, le professeur Daoud BERKANI, le docteur Mohamed OULD ZMIRLI, et le docteur Zoubir Abdeslem BENSELAMA, ainsi que les experts pour avoir accepté de juger ce travail.

Je remercie, spécialement, mes amis Dr. Idir Mechai et Dr. Fouad Khelifi pour m'avoir aidé à réaliser ce travail.

Un grand merci à ma mère et à mon père pour leur soutien et leur prières, ainsi qu' à tous les membres de ma famille.

Cette section sera incomplète sans avoir remercié tous mes amis et frères qui se reconnaitront, dont le support moral, était ma source d'énergie dans l'accomplissement de ce travail.

SOMMAIRE

CHAPITRE 2 : SYSTEME DE RECONNAISSANCE DE LOCUTEUR

CHAPITRE 3 : ETUDE DES PARAMETRES INVARIANTS POUR LA COMPENSATION DE LA VARIABILITE EN RECONNAISSANCE DU LOCUTEUR.

CHAPITRE 4 : EXPERIMENTATIONS ET RESULTATS

LISTE DES ILLUSTRATIONS, GRAPHIQUES ET TABLEAUX

LISTE DES SYMBOLES ET ABREVIATIONS

AMDF : Absolute Magnitude Difference function

ANN : Artificial Neural Network

APS : Active-Pixel Sensor

AR : Auto Régressive

CCD : Charge-Coupled Device

CMVN : Cepstral Mean and Variance Normalization

DCT : Discrete Cosine Transform

DET : Detection Error Tradeoff

DFT : Discrete Fourier Transform

DTW : Dynamic Time Warping

EER : Equal Error Rate

EM : Expectation Maximisation

FAR : False Accept Rate

FFT : Fast Fourier Transform

FM : Feature Mapping

FRR : False Reject Rate

GMM : Gaussian Mixture Models

HMM : Hidden Markov Model

IIF : Invariant Integration Features

JFA : Joint Factor Analysis

LPC : Linear Predictive Coefficients

LR : Likelihood Ratio

MAP : Maximum A Posteriori

MFCC : Mel Frequency Cepstral Coefficients

minDCF : Minimum Detection Cost Function

MLP : Multi Layer Perceptron

NAP : Nuisance Attribute Projection

PLP : Perceptual Linear Prediction

ROC : Receiver Operating Characteristic

SVM : Support Vector Machine

TF : Transformée de Fourier

UBM : Universal Background Model

VAD : Voice activity detection

INTRODUCTION GENERALE

Le signal de parole est porteur de plusieurs types d'informations comme le message, la langue, les émotions, ou même l'environnement. Ces informations sont considérées comme source pour divers domaines d'applications, nous citons à titre d'exemple, la reconnaissance de la parole, la reconnaissance du langage naturel, et la reconnaissance du locuteur. Cette dernière application est une modalité biométrique qui utilise la voix (ou la parole) pour identifier les personnes.

Durant ces dernières décennies, la recherche dans le domaine de la reconnaissance des locuteurs a connu des avancées considérables, et des systèmes de reconnaissance basés sur la voix ont été introduits dans certains domaines. C'est ainsi que la technologie de la reconnaissance du locuteur a permis l'utilisation de la voix pour contrôler l'accès des personnes à des services restreints (comme l'accès aux services bancaires), ou effectuer des transactions financières par téléphone.

Cependant, même avec cette rapidité de développement des technologies de la reconnaissance des locuteurs, beaucoup de problèmes restent, jusqu' à ce jour, non complètement résolus. Un de ces problèmes est la robustesse des systèmes de reconnaissance automatique des locuteurs à la variabilité.

Récemment, les travaux dans la reconnaissance du locuteur se sont concentrés sur le traitement de ce problème de variabilité. Cette variabilité est causée par le canal de transmission, le bruit, et les caractéristiques temporelles du locuteur (comme l'humeur ou la fatigue). Le terme "variabilité de session" se rapporte à tous les phénomènes qui causent une différence de son de deux enregistrements provenant d'un même locuteur. Les non correspondances (mismatch) des capteurs, de la langue, du style de parole, et de l'environnement entre la phase d'apprentissage et la phase test (ces deux phases sont à la base de tout système biométrique, comme le système de reconnaissance du locuteur) affectent dramatiquement les performances d'un système de reconnaissance.

Plusieurs méthodes ont été développées afin de palier à ce problème de variabilité. Ces méthodes agissent sur les différentes parties constituant le système de reconnaissance du locuteur. On parle, alors, de compensation de la variabilité au niveau de l'étage d'extraction des paramètres, au niveau de l'étage de modélisation, et au niveau de l'étage de prise de décision.

De par leur position dans tout système de reconnaissance, Il est clair qu'une sélection appropriée des paramètres peut augmenter, significativement, le taux de classification. L'extraction des paramètres invariants est une technique alternative qui a été appliquée avec succès dans beaucoup d'applications de l'analyse des images [1]. Les structures invariantes ont été proposées pour la reconnaissance de la parole pour surmonter le problème des facteurs non linguistiques [2,3]. Les résultats expérimentaux sur les énoncés de voyelles en langue japonaise, montrent que ces paramètres réalisent de meilleurs taux de reconnaissance.

Müller et al [4] ont utilisé les paramètres d'intégration invariants (invariant integration features IIF) dans la reconnaissance de la parole. Les bons résultats obtenus montrent l'efficacité de ces paramètres en termes de réduction des erreurs de reconnaissance. Dans leur travail, ils ont proposé des paramètres qui sont invariants à la translation de l'espace des sous bandes de fréquence, où l'objectif était de reconnaitre la même parole, même si elle était prononcée par différents locuteurs (reconnaissance de la parole indépendante du locuteur).

Dans notre travail, nous proposons d'utiliser les paramètres invariants, pour reconnaitre le même locuteur, même si sa parole est enregistrée par différents moyens ou dans des environnements différents, qui sont des cas de variabilité. En plus du fait que ces paramètres sont invariants en fréquence, ils doivent être, aussi, invariants en amplitude.

L'idée de ces paramètres invariants est de trouver une transformation T capable d'extraire des paramètres similaires correspondant à des observations x d'une classe équivalente (le locuteur dans notre cas).

Considérons un signal de parole pur corrompu par un bruit additionnel et un canal, tel que représenté par la figure suivante :

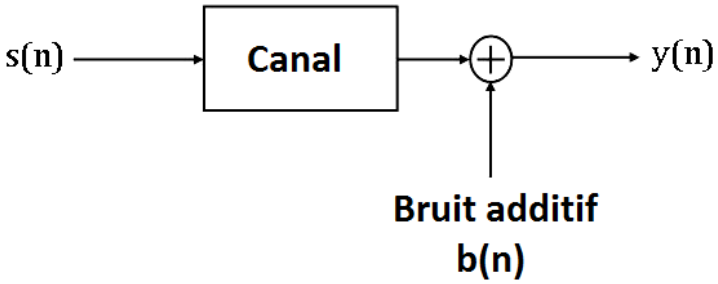


Figure 1 : Processus d'obtention d'un signal de parole observé

Dans ce cas la sortie y(n) peut s'écrire comme :

$$y(n) = s(n) * h(n) + b(n), \quad (1)$$

où $y(n)$ représente le signal de parole observé, $h(n)$ la réponse impulsionnelle du canal, $s(n)$ est le signal de parole pur, et $b(n)$ le bruit additionnel. $*$ indique l'opération de convolution.

En prenant la transformée de Fourier de cette équation, nous aurons :

$$Y[k] = S[k]H[k] + B[k], \quad (2)$$

où k indique l'indice de fréquence.

L'équation (2) montre l'effet du canal (exprimé par $H[k]$) sur le signal de parole résultant.

Dans ce travail, nous considérons le traitement du problème de la variabilité causé par l'utilisation de différents canaux, couramment appelé variabilité de non correspondance (mismatch variability), en utilisant les paramètres invariants dont nous sommes les premiers à les utiliser dans le contexte de la reconnaissance du locuteur [5]. Cependant, les paramètres proposés sont extraits en tenant compte de deux actions, le changement de l'amplitude et la translation de la fréquence.

Le reste de cette thèse est organisé comme suit :

Le premier chapitre décrit les concepts généraux de la biométrie et de la parole. Après avoir définit la biométrie, nous allons donner la structure générale d'un système biométrique, les modalités les plus rencontrées et les principales applications de la biométrie. Nous passons, ensuite, à la modalité parole (voix) avec plus de détails puisqu'elle servira de base à la conception de notre système de reconnaissance du locuteur.

Dans le deuxième chapitre, nous s'intéresserons au système de reconnaissance du locuteur. Sachant que le système travaille en apprentissage et en reconnaissance (test), les modules qui composent ces deux parties seront donnés avec détails. C'est ainsi qu'on verra l'extraction des paramètres regroupant le prétraitement et les paramètres couramment utilisés comme les coefficients de prédiction linéaire LPC (Linear Predictive Coefficients), et les paramètres cepstraux MFCC (Mel Frequency Cepstral Coefficients) et PLP (Perceptual Linear Prediction). Le système de modélisation, que nous allons utiliser tout au long de ce travail, est basé sur les modèles Gaussiens GMM - UBM (Gaussian Mixture Models - Universal Background Model), sera aussi étudié dans ce chapitre. Enfin, nous terminerons par décrire le module de prise de décision et les mesures utilisées pour l'évaluation de notre système de reconnaissance.

Le troisième chapitre comportera l'essentiel de notre contribution dans cette thèse, nous consacrerons une partie de ce chapitre pour définir le problème de variabilité et les différentes façons de le traiter. Par la suite, nous détaillerons notre méthode de compensation de la variabilité basée sur les paramètres invariants.

Le chapitre quatre sera dédié a l'expérimentation des différentes situations de la reconnaissance du locuteur. Pour cela, nous commençons par la description des données utilisées pour effectuer les diverses expériences, et les protocoles d'évaluation des systèmes de reconnaissance du locuteur. Des expériences visent à valider les concepts théoriques relatifs a notre système, d'autres visent à tester l'approche des invariants adoptée pour la compensation de la variabilité dans le domaine de la reconnaissance du locuteur. La discussion des résultats va nous

permettre de juger l'efficacité des paramètres invariants comparés aux paramètres les plus utilisés dans ce domaine, à savoir, la méthode MFCC.

Nous terminerons notre travail par une conclusion générale.

CHAPITRE 1

VUE D'ENSEMBLE SUR LA BIOMETRIE ET LA PAROLE

1.1. Introduction

La biométrie est la science de mesure des caractéristiques physiologiques comme les empreintes digitales, l'iris, ou le visage, et comportementales comme la démarche, la signature, ou la façon de taper sur un clavier. Toutes ces caractéristiques peuvent êtres utilisées pour l'identification des individus. La figure 1.1 montre quelques applications biométriques basées sur ces caractéristiques physiologiques et comportementales.

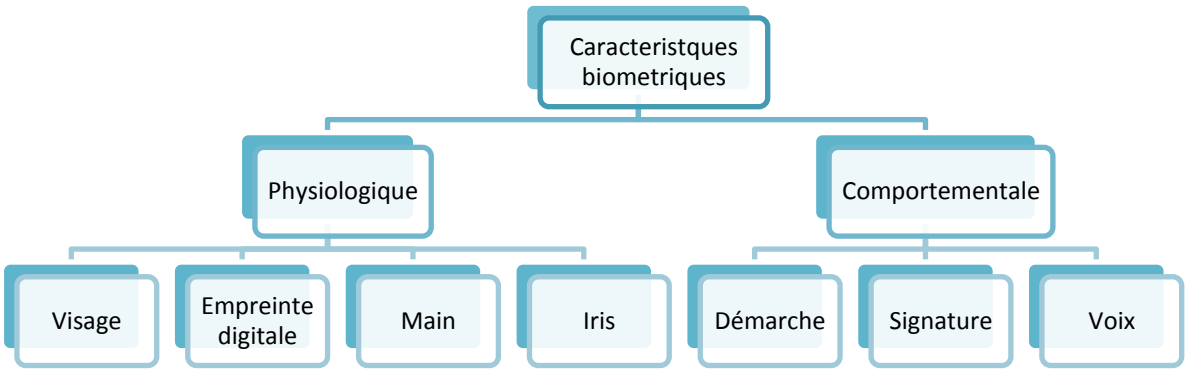


Figure 1.1 : Quelques exemples de caractéristiques biométriques

Le développement d'applications nécessitant la détermination avec exactitude des identités des personnes, comme les opérations bancaires onlines, ou l'ouverture de centres décentralisés offrant des services à la clientèle, ont renforcé la nécessité de faire appel aux techniques biométriques.

Habituellement, la sécurité a toujours été liée à quelque chose qu'on "possède" (comme une carte d'identité ou une clé) ou quelque chose qu'on "sait" (comme un mot de passe ou un code PIN). Cependant, ces types de moyens de sécurité peuvent être perdus, volés ou oubliés [6].

Les techniques biométriques courantes établissent l'identité en se basant sur se qu'on "est" (comme l'empreinte digitale, la voix, la main, ou le visage). Les résultats indiquent que les techniques biométriques sont beaucoup plus précis que les techniques traditionnelles [7]. La particularité de la biométrie est que deux personnes ne possèdent jamais le même visage ou la même voix. Cette propriété est utilisée pour vérifier ou identifier les personnes.

Fondamentalement, un système biométrique peut opérer en mode vérification appelé aussi authentification ou en mode identification connu aussi par le mode de reconnaissance. La vérification répond à la question " Es tu la personne qui prétend l'être? ", le système répond à cette question en comparant la donnée biométrique déjà enregistrée avec sa propre caractéristique biométrique.

Pour retenir une caractéristique physiologique ou comportementale comme paramètre de reconnaissance biométrique, elle doit être :

- Universelle : Tel que toutes les personnes doivent avoir cette caractéristique.
- Permanente : Tel que la caractéristique ne doit pas changée au cours du temps.
- Distinctive : Tel que chaque personne doit être différenciable des autres.
- Acceptable : Tel que la caractéristique doit avoir une large acceptabilité dans les sociétés.
- Performante : Tel que la technologie utilisée soit précise, rapide et robuste.
- Mesurable : Tel que les moyens d'acquisition de la caractéristique soient faciles.

Les systèmes biométriques ne possèdent pas toutes ces propriétés, ou du moins les possèdent avec des degrés différents.

Dans la littérature, les caractéristiques biométriques sont aussi appelées traits, indicateurs, ou modalités. Bien que les systèmes biométriques ont leurs propres

limitations comme la performance, la non universalité de certaines modalités (certaines catégories de population ne peuvent pas être identifiées par certaines caractéristiques), ils ont beaucoup d'avantages par rapport aux méthodes traditionnelles puisqu'ils ne sont pas faciles à voler ou partager, ils sont aussi plus commodes car l'utilisateur n'a pas besoin de concevoir ni de retenir de mots de passe.

1.2. Système biométrique

Essentiellement, un système biométrique est composé de deux parties, la partie apprentissage et la partie test (voir figure 1.2). Ces deux parties se partagent les opérations de captation de la modalité choisie, le prétraitement visant à améliorer la qualité des données acquises, et l'extraction des paramètres pertinents qui représentent le mieux la modalité considérée. Ces paramètres seront enregistrés dans la base de données durant l'étape d'apprentissage.

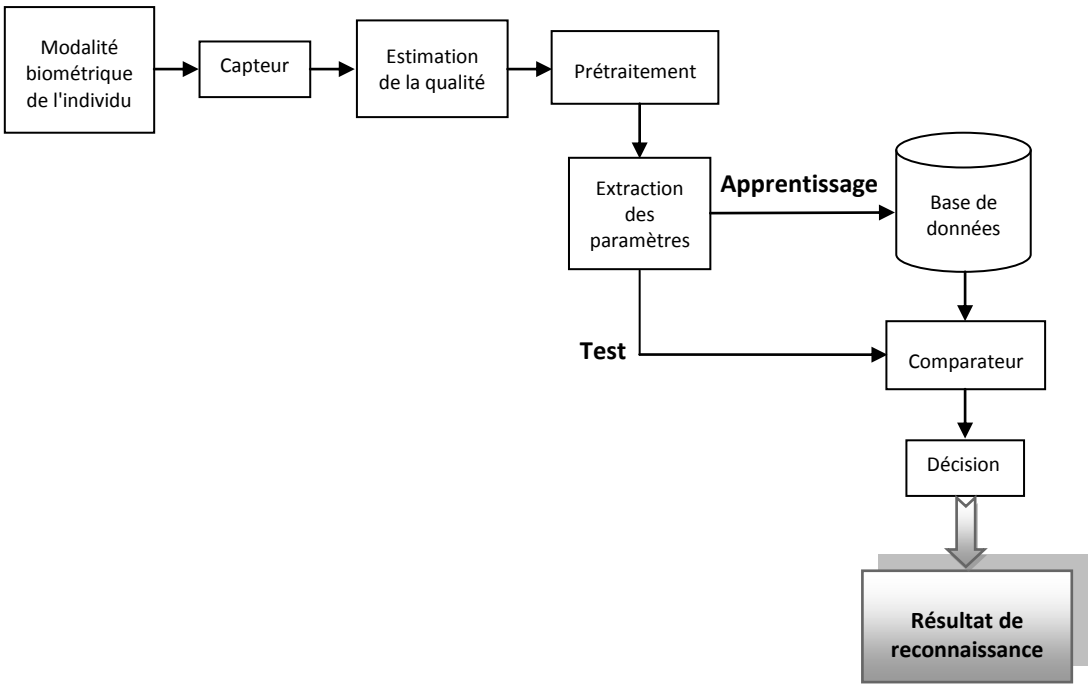


Figure 1.2 : Schéma blocs du système biométrique.

Durant l'étape de test, les paramètres correspondant à l'utilisateur du système seront comparés à un ou plusieurs modèles déjà enregistrés dans la base de données. Deux modes de reconnaissance existent.

- Le mode vérification, dans lequel l'utilisateur prétend une identité et le système vérifie si cette prétention est authentique en comparant l'entrée biométrique au modèle du même utilisateur contenu dans la base de données.
- Le mode identification, dans lequel l'entrée biométrique de l'utilisateur est comparée aux modèles de toutes les personnes enregistrées dans la base de données. Dans ce cas, le système retourne soit, l'identité de la personne ayant le plus grand taux de similarité avec son modèle parmi tous les autres modèles, soit une décision de non existence de cette personne dans cette base de données.

1.3. Spécifications de conception du système biométrique

1.3.1. Précision

Deux types d'erreurs peuvent être commis dans un système biométrique; la fausse rejection et la fausse acceptation. Ces erreurs se produisent quand la variation intra-utilisateur est large. Par conséquent, on définit la précision d'un système biométrique par le taux de faux rejet (False reject rate FRR) et le taux de fausse acceptation (False accept rate FAR). Le taux de faux rejet est la probabilité produite par le système en termes de rejection de personnes authentiques. Alors que le taux de fausse acceptation est la probabilité produite par le système en termes d'acceptation de personnes non authentiques (appelées imposteurs).

La précision requise par un système biométrique dépend essentiellement de l'application. Par exemple, pour le contrôle d'accès le système biométrique a comme objective de refuser l'entrée aux personnes étrangères. Il est clair que dans ce cas, le taux de fausse acceptation doit être bas même si ceci se fait au répit d'un plus grand taux de faux rejet.

Pour une autre application comme la surveillance des enfants perdus par une camera dans un lieu publique, le système biométrique doit avoir un taux de faux rejet très bas, même si le taux de fausse acceptation, dans ce cas, sera élevé.

1.3.2. Débit

Il correspond au nombre de transactions qu'un système biométrique peut traiter par unité de temps. En vérification, le débit ne pose pas un grand problème, puisque cette opération n'implique qu'une comparaison de l'échantillon de test contre le modèle correspondant enregistré dans la base de données. Le problème devient plus sérieux en identification puisque dans ce mode, l'échantillon test doit être comparé à chacun des modèles enregistrés dans la base de données. Lorsque cette base de données est large le système biométrique doit comporter des procédés appropriés comme l'indexation ou le filtrage permettant de faciliter la recherche dans la base de données pour améliorer le débit.

1.3.3. Intimité

La biométrie permet une authentification facile en établissant une liaison avec l'identité de la personne. Néanmoins, des problèmes de violation de la vie privée des personnes peuvent surgir. Ceci est dû à un détournement d'usage où la donnée biométrique subit des abus pour des raisons mal intentionnées. La donnée enregistrée par le système biométrique peut, par exemple, faciliter la poursuite des utilisateurs sans qu'ils le sachent. Dans ce cas, le système biométrique doit effectuer des vérifications afin de protéger l'intimité des utilisateurs.

1.3.4. Coût

Le coût total d'un système biométrique inclut les composants de ce système. Il inclut aussi les opérations récurrentes pour la maintenance, ou la mise à niveau du système biométrique. Souvent, un compromis entre le coût des composants et les performances du système biométrique est fait. Les coûts résultants des erreurs commises par le système biométrique doivent aussi être évalués.

1.4. Types de technologies biométriques

1.4.1. Empreinte digitale

L'empreinte digitale est la modalité la plus largement utilisée parmi les modalités biométriques existantes. Les lecteurs d'empreintes digitale ont atteint les dimensions, le prix, et les performances nécessaires pour être intégrés dans beaucoup d'appareils d'accès logiques comme les ordinateurs portables, les smartphones, les claviers, ...etc. Cependant, des facteurs comme l'association de

cette modalité aux applications criminalistiques, des blessures au niveau des doigts durant les travaux manuels, peuvent affecter l'efficacité de cette modalité.

On définit l'empreinte digitale comme l'impression laissée par l'apposition de la peau du doigt. Cette empreinte apparait comme une série de crêtes et de vallées (figure 1.3).

Figure 1.3 : Empreinte digitale.

L'analyse globale de la forme de l'empreinte digitale présente des régions distinctives formées par les lignes de crêtes. On retrouve des régions sous forme de delta, ou de boucle.

A un niveau local, on caractérise l'empreinte digitale par la minutie. Cette dernière invoque les différentes façons par lesquelles les crêtes peuvent être discontinues. Une crête peut, par exemple, passer à une terminaison, ou se diviser en deux crêtes (bifurcations).

Historiquement, l'acquisition d'images des empreintes digitales se faisait par l'étalement de l'encre sur le doigt, et la pression de ce dernier sur un papier. Ce papier sera ensuite scanné pour donner une représentation numérique de l'empreinte. De nos jours, l'empreinte digitale est obtenue par des capteurs électroniques, on retrouve:

- Les capteurs optiques : Dans ce type de capteurs, les crêtes et les vallées sont obtenues par l'absorption et la réflexion de la lumière, respectivement. Cette lumière réfléchie est dirigée vers des capteurs CCD (Charge-coupled device) ou APS (active-pixel sensor).
- Les capteurs ultrasons : Des signaux acoustiques sont envoyés, puis les échos de ces signaux sont capturés selon la surface de l'empreinte du doigt. La qualité des images acquises dans cas est bonne puisque les signaux acoustiques sont capables de traverser la saleté ou la graisse pouvant se présenter dans la main.
- Les capteurs de silicium : Ces capteurs sont constitués d'un ensemble de pixels, où chaque pixel n'est rien d'autre qu'un petit capteur. En touchant la surface du silicium, un courant électrique est produit. L'effet peut résulter d'origine capacitive, thermique, champ électrique, ou piézoélectrique.

1.4.2. Visage

Le visage est le mode le plus naturel de reconnaissance de personnes chez l'être humain. Il est robuste et non intrusif. Une illustration est montrée dans la figure 1.4.

Cependant, ce type d'identification reste encore un très difficile problème. Ceci est du aux variabilités de visages des personnes dans des situations opérationnelles, qui ne sont pas toujours métrisables. Ces différentes situations peuvent provenir de différence d'illumination, de rotations, d'expressions, d'âge, de lunettes, etc.

Les performances du système biométrique utilisant la modalité de visage est dans ces cas très dégradées. Ce qui empêche sa large utilisation dans les applications réelles.

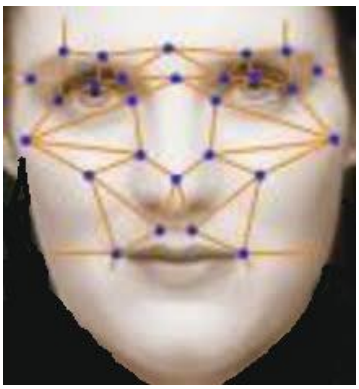


Figure 1.4 : Géométrie du visage.

En reconnaissance faciale utilisant le visage comme modalité biométrique, deux techniques sont adoptées. La première est basée sur l'apparence du visage alors que la deuxième est basée sur son géométrie. Dans la première technique on considère les propriétés du visage entier. Dans la deuxième technique, basée sur la géométrie du visage, on considère des relations comme la surface, la distance, ou l'angle entre certains points du visage comme paramètres descriptifs de la modalité faciale. Ces différents coefficients seront suivis par d'autres méthodes comme, l'analyse en composantes principales, l'analyse discriminante linéaire, ou l'analyse en composantes indépendantes, afin de réduire la dimension du vecteur des coefficients représentant le visage et de produire des coefficients décorrélés.

L'acquisition des données de visage se fait par des caméras. Selon l'application, la représentation du visage peut être des images en 2-dimensions acquise directement de la camera ou des images construite en 3-dimensions. On peut aussi acquérir des données à travers une séquence vidéo. La qualité du signal obtenu diffère d'une caméra à une autre. Le choix du niveau d'illumination ambiant, les effets du bruit, et la vitesse de mouvement des objets dans une vidéo doit être soigneusement fait.

1.4.3. Iris

L’iris est une membrane circulaire de la face antérieure du globe oculaire. Elle est percée en son centre (pas exactement au centre) d’un orifice ou trou noir appelé la pupille par laquelle la lumière pénètre vers la rétine. L’iris sert à adapter cette quantité de lumière en se réfractant ou se dilatant suivant les conditions de luminosité. Par exemple, quand la luminosité ambiante est forte, l'iris se contracte, ce qui diminue l'intensité lumineuse qui vient frapper le centre de la rétine, et vice-versa.

L’iris est un organe qui doit sa couleur, qu’elle soit grise, verte, bleue, marron ou noire au pigment responsable de la coloration : la mélanine.

Le fait que l'iris est très riche en informations, stable durant la vie de l'être humain, et rarement sujet aux dommages pouvant toucher les autres parties du corps comme les doigts, fait que cette modalité biométrique soit très convenables et efficace [8]. Deux vrais jumeaux ont des iris distincts [9].

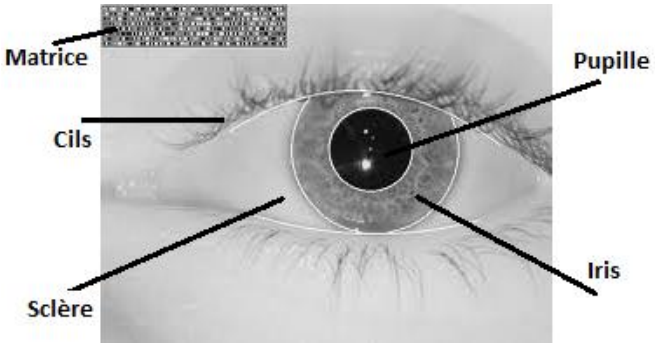


Figure 1.5 : Image de l'iris avec sa matrice de données.

Cependant, des facteurs peuvent rendre l'acquisition des images de l'iris pas facile à faire comme :

- Sa taille très petite (≈1cm) pour une distance d'environ 1m.
- La situation de l'iris derrière la cornée, qui est considérée comme une surface très humide et réfléchissante, perturbe les images à acquérir.
- La paupière est souvent tombante, ce qui masque partiellement l'iris.
- L'iris est une cible en mouvement.

Le système de reconnaissance de personnes basé sur la modalité biométrique iris est composé de quatre modules; l'acquisition, la segmentation, la normalisation, et le module de comparaison. Le rôle du module d'acquisition est d'obtenir des images 2D de l'œil en utilisant des caméras CCD monochromes sensibles aux infrarouges proches (near-infrared NIR) du spectre électromagnétique. Souvent, ces systèmes de reconnaissance exigent la coopération de l'utilisateur et de placer leurs yeux à proximité de la camera. Dans le module de segmentation, on localise les limites spatiales de l'iris dans l'image en l'isolant des autres parties voisines. Ces parties incluent la pupille, la sclère, la paupière, et les cils. Une fois les limites de l'iris estimées, le module de normalisation se charge de transformer la texture de l'iris, située dans la région annulaire, des coordonnées cartésiennes aux coordonnées pseudo-polaires. Cette procédure aboutit à une matrice dont les lignes correspondent à la direction angulaire de l'iris et les colonnes aux directions radiales.

La normalisation permet d'assurer que les iris des différents individus soient représentés dans un domaine commun d'images. Dans la dernière étape deux matrices correspondant chacune à un iris peuvent être directement comparées en utilisant des filtres de corrélation. Cependant, un module d'extraction de paramètres peut être inséré avant celui de la comparaison. Parmi les approches les plus adoptées dans ce cas, on retrouve ceux basées sur les filtres de Gabor 2D qui extraient la phase de la texture de l'iris dans une analyse multi-échelles.

1.4.4. Démarche

La démarche est définie comme le style ou la manière avec laquelle un individu marche. Dans une situation de suivi et de protection d'une large zone, un problème de reconnaissance d'une personne se situant à une centaine de mètres peut se poser. Utiliser des modalités comme l'iris ou l'empreinte digitale n'est plus possible. De même, utiliser le visage n'est pas très recommandé car les images capturés dans de telles environnements et à des distances lointaines ne seront sans doute pas de bonne qualité.

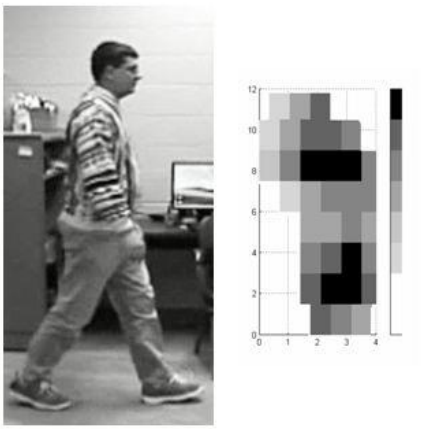


Figure 1.6 : La modalité de démarche.

La démarche est dans ce cas une alternative pour l'identification en utilisant la forme et le mouvement dans une vidéo d'une personne en marche. La démarche d'une personne est déterminée par sa structure musclo-squelettique. C'est ainsi que Cutting et al [10] montrent expérimentalement la possibilité d'identifier une personne à partir de sa manière de marcher.

Dans un système biométrique basé sur la démarche, une camera vidéo est utilisée pour capturer les personnes en marche dans son champ de vision. Si la camera travail dans des conditions contrôlées la qualité de la vidéo est suffisamment bonne pour extraire les paramètres pertinents à la reconnaissance comme l'angle de la jambe, ou les trajectoires de la tête et des pieds. Dans les cas non contrôlés, les paramètres à extraire peuvent être des silhouettes extraites d'images binaires avec soustraction de l'arrière plan.

La plupart des paramètres utilisés dans cette modalité biométrique doivent inclure un traitement permettant de les rendre invariants (au moins en échelle et en translation). Par la suite une mesure de similarité sera faite entre le vecteur de paramètres de la démarche d'un individu de test avec un modèle de la base de données. L'algorithme de déformation temporelle dynamique (Dynamic Time Warping DTW) peut, par exemple, comparer deux séquences de paramètres. Cet algorithme est basé sur la programmation dynamique, il calcule la meilleure normalisation temporelle non linéaire de la séquence test avec la séquence d'apprentissage. La méthode statistique de modélisation des chaines de Markov cachés est aussi utilisée dans ce domaine. Les paramètres de la démarche sont la sortie des états cachés, dont les transitions sont supposées markoviennes.

1.4.5. Voix

La voix est une combinaison de caractéristiques biométriques physiques et comportementales. Les caractéristiques physiques reflètent la forme et la taille du système phonatoire (P.ex. le conduit vocal, la bouche, le conduit nasal, et les lèvres). Les caractéristiques biométriques comportementales reflètent la façon de parler due au contexte sociolinguistique, l'éducation, ou l'environnement socio-économique.

Même si la modalité voix n'est pas considérée aussi précise que d'autres modalités comme l'empreinte digitale ou l'ris [7], deux principaux avantages font qu'elle soit très utilisée. Le premier est que la parole est la modalité la plus naturelle et souvent la moins intrusive (aucun contact physique n'est requis) entre toutes les modalités biométriques. Le second avantage est le résultat de la prolifération et la grande utilisation des différents types de téléphones (terrestres, mobiles et IP). Cette

infrastructure fait que la voix n'a pas besoin d'autres appareils ou de systèmes de transmission.

Parmi les applications de la biométrie basées sur la voix on trouve l'authentification de locuteur (pour le contrôle d'accès à distance par téléphone), la détection de locuteur, et la reconnaissance de locuteur en criminalistiques (Forensic Speaker Recognition).

Figure 1.7 : La voix comme modalité d'identification.

Dans les applications utilisant les ordinateurs, l'acquisition des signaux de parole se fait par de simples cartes sons et des microphones. En téléphonie, on n'a pas besoin d'appareils d'acquisition du moment que chaque appareil téléphonique peur être utilisé n'importe où.

La parole peut être caractérisée par des paramètres de "niveau haut" ou des paramètres de "niveau bas". La première catégorie d'information reflète les caractéristiques comportementales (niveau haut) comme la prosodie, la prononciation, ou la phonétique. La seconde catégorie reflète les propriétés spectrales de la parole (niveau bas) liées à la nature physique d'appareil phonatoire. Cette deuxième catégorie est la plus utilisée en reconnaissance de locuteur. L'extraction des paramètres dans ce cas utilise la transformée de Fourier courte. Il est à noter que les paramètres de niveau haut ont reçu un intérêt croissant ces dernières années afin d'améliorer les résultats de reconnaissance [11].

Les modèles de mélange gaussien (Gaussian Mixture Models GMM) sont devenus, ces dernières années, l'approche dominante en reconnaissance de locuteur. Ils estiment la densité de probabilité permettant de capturer toutes les variations du signal parole.

Dans l'étape de prise de décision, si on est dans le mode vérification, on calcule le rapport de vraisemblance, qui est le rapport entre les fonctions de densités de probabilité du signal parole test associées au modèle locuteur en question en numérateur et le modèle des autres locuteurs en dénominateur. Ce rapport sera comparé à un seuil pour accepter ou rejeter ce locuteur. En mode identification, On calcule les rapports de vraisemblance du signal de test pour chaque modèle des locuteurs. On identifie alors le locuteur par celui ayant le rapport de vraisemblance le plus élevé.

1.5. Quelques détails sur le mécanisme de production de la parole

La parole est produite par les organes de l'appareil phonatoire montrés dans la figure 1.8. Les poumons constituent la source principale de la parole. Lors de la parole, l'air est chassé des poumons, traverse le larynx, puis passe entre les cordes vocales pour arriver aux trois principales cavités du conduit vocal, à savoir, le pharynx, la cavité orale et/ou la cavité nasale. A partir des cavités orale et nasale, ce flux d'air sort à travers le nez et la bouche.

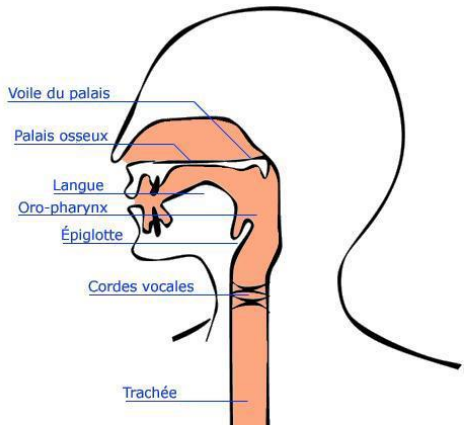


Figure 1.8 : Appareil phonatoire.

La forme d'ouverture entre les cordes vocales, appelé glotte, est la source sonore du système vocal (figure 1.9). Les cordes vocales peuvent agir par différentes manières durant la parole. Dans le but de moduler le flux d'air, les cordes vocales s'ouvrent et se ferment rapidement, produisant un bourdonnement comme pour les voyelles. L'occlusion totale peut se produire si, soudainement, les cordes vocales qui étaient totalement fermées s'ouvrent.
D'un autre coté, la production des consonnes non voisées comme /s/ se fait lorsque les cordes sont complètement ouvertes. Une position intermédiaire peut donner lieu à un faible chuchotement comme pour le son \h\.

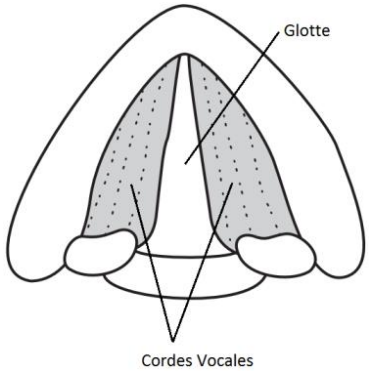


Figure 1.9 : Cordes vocales.

Ces distinctions des sons de parole sont déterminées non seulement par la source, mais aussi par les différentes configurations du conduit vocal, et la combinaison de ces configurations avec les différentes sources (périodique, impulsive, etc.).

Le conduit vocal est lié par des structures de tissus doux et d'autres durs. Ces structures sont soit immobiles, comme le palais dur et les dents, soit mobiles. Les structures mobiles associées à la production de la parole sont appelées articulateurs. La langue, les lèvres, la joue et la voile sont des articulateurs. Leurs mouvements comptent pour la majorité des variations de la forme du conduit vocal associée à la parole.

La plus petite unité de parole est appelée phonème. L'association d'un ou plusieurs phonèmes forme la syllabe. Le mot est formé d'un ou plusieurs syllabes.

1.6. Modèle source filtre de la parole

Le modèle source filtre de la parole [12] est au cœur de plusieurs méthodes d'analyse da la parole. L'idée de ce modèle est que les sons de la parole sont produits par l'action d'un filtre, le conduit vocal, sur la source sonore qui peut être la glotte ou une autre constriction dans le conduit vocal (Figure 1.10).

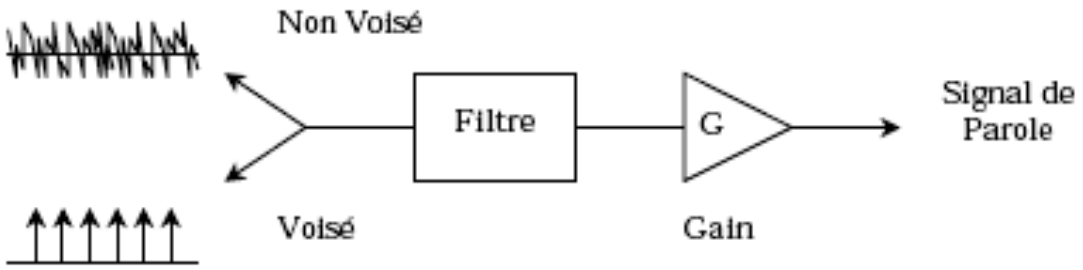


Figure 1.10 : Modèle source - filtre de la parole.

Le modèle se base sur la supposition que la source et le filtre sont indépendants. Cette supposition implique que la modification des propriétés des filtres ne change pas les propriétés de la source et vice versa. Quoique ce ne soit pas strictement vrai dans tous les cas, en pratique, cette supposition donne un modèle d'une grande utilité et largement précis da la production de la parole.

1.6.1. Source

D'un point de vue acoustique, les sources peuvent correspondre aux sons voisés de parole et non voisés.

La source de la parole voisée est la vibration des cordes vocales en réponse à un courant d'air provenant des poumons. Cette vibration est périodique. Son examen montre qu'elle est constituée d'une série de larges pointes (Figure 1.11).

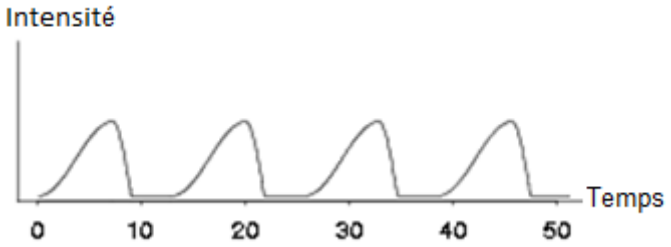


Figure 1.11 : Forme périodique de la source de parole.

Le spectre de la source glottique est constitué de pics de fréquences correspondant aux harmoniques de la fréquence fondamentale de vibration des cordes vocales (figure 1.12).

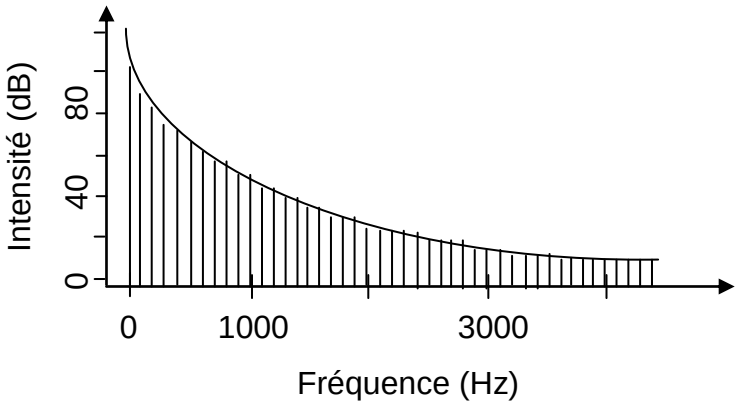


Figure 1.12 : Spectre de la source glottique.

L'amplitude du spectre décroit avec l'augmentation de la fréquence. La fréquence fondamentale de vibration des cordes vocales dépend de la masse et la tension de ceux ci. Elle est d'environ 100 Hz, 200 Hz, et 300 Hz pour les hommes, les femmes, et les enfants, respectivement.

La source de parole non voisée est créée par une vibration non régulière des cordes vocales. Ces vibrations sont causées par un flux d'air turbulent.

1.6.2. Filtre

En général, un filtre est un système qui altère la composition fréquentielle d'un signal d'entrée. En production de la parole, le filtre est le conduit vocal. Ce dernier est considéré comme un tube acoustique constitué d'un assemblage de sous tubes de sections variables qui est fermé d'un côté (par la glotte) et mesure, en moyenne, environ 17.5 cm pour l'homme (figure 1.13).

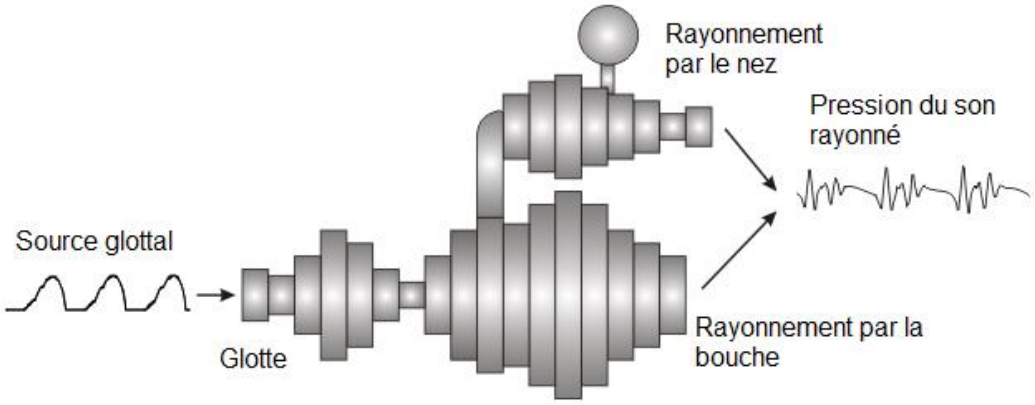


Figure 1.13 : Conduit vocal à tubes.

Comme pour tout filtre, ce tube est caractérisé par un spectre. Essentiellement, ce spectre change quand la forme du conduit vocal change durant la production de parole. Par conséquent, différents sons sont produits par le changement de la forme du conduit vocal, qui donne un ensemble particulier de caractéristiques du filtre. L'espace du conduit vocal, composé de cavités orale et nasale, peut être vu comme un filtre acoustique à temps variable qui amplifie certaines fréquences du spectre et attenue d'autres. Les fréquences de résonance du conduit vocal sont appelées formants. Ces formants dépendent des phonèmes, mais aussi, de la forme générale, de la longueur et du tissu du conduit vocal. Les sons voisés sont composés d'environ 3 à 5 formants.

La flexibilité du conduit vocal, dans lequel les articulateurs sont facilement ajustés pour former une variété de formes, donne un potentiel de produire une large gamme de sons.

1.7. Analyse du signal de parole

L'analyse des signaux de parole tient en considération les façons par lesquelles les sons de parole sont produits. Comme décrits dans la section 1.4.e, on dira que les signaux de parole résultent de l'interaction de deux facteurs ; la source glottique et le conduit vocal. Par conséquent, le signal de parole est obtenu par une source passant dans un filtre linéaire à temps variable. Le filtre peut dériver des modèles de production de la parole basés sur la théorie acoustique. Dans cette théorie, la source représente le flux d'air au niveau des cordes vocales et le filtre représente les résonnances du conduit vocal qui change dans le temps. La figure 1.14 illustre le processus source - filtre de la parole.

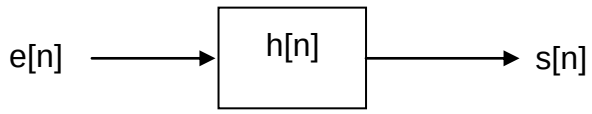


Figure 1.14 : Système d'obtention du signal de parole.

La pression du flux d'air rayonné par la bouche et / ou le nez est convertie en un courant électrique à travers un microphone. Ce signal de parole électrique utilise essentiellement des fréquences allant de 100 Hz à quelques 8000 Hz et des amplitudes variant de 30 à 90 dB.

L'analyse des signaux de parole est le processus d'estimation des paramètres d'un modèle, variant dans le temps, dans le but d'extraire des informations sur la production de ce signal. Les méthodes modernes d'analyse des signaux sont basées sur le traitement numérique des signaux. La première raison de la numérisation est la facilitation des techniques de traitement sophistiquées qui sont très difficiles voir impossible à réaliser en analogique. La seconde raison est que le traitement

numérique est beaucoup plus efficace et peut être effectué dans des circuits compacts.

La numérisation implique l'échantillonnage pour discrétiser le temps et la quantification pour discrétiser l'amplitude. Le taux avec lequel le signal analogique est échantillonné est appelé fréquence d'échantillonnage. Dans les réseaux de télécommunications les signaux de parole analogiques sont limités entre 300 et 3400 Hz, la fréquence d'échantillonnage est dans ce cas de 8000 Hz, conformément au théorème de Nyquist. Pour une qualité supérieure, la bande passante fréquentielle de la parole est limitée entre 0 et 7000 Hz, dans ce cas la fréquence d'échantillonnage est choisie égale à 16 KHz.

Le signal échantillonné est ensuite quantifier en amplitude en utilisant un convertisseur analogique - digital permettant de représenter chaque échantillon réel en un nombre limité de bits. 12 bits sont nécessaires en pratique pour assurer un rapport signal sur bruit supérieur à 35 dB.

1.7.1. Analyse temporelle

Les fonctions constituant le modèle source - filtre de la parole sont toutes variables dans le temps. L'excitation est souvent considérée comme un bruit blanc, un signal périodique, ou un mélange des deux. Un signal périodique résulte d'une excitation voisée qui se produit lorsque les cordes vocales vibrent. Le bruit aléatoire se manifeste dans le cas d'une excitation non voisée au niveau d'une constriction dans le conduit vocal créant un flux d'air turbulent. On appelle voisement le mélange des ces deux excitations.

La figure 1.15 montre une représentation temporelle d'une voyelle. Il est clair que seule, une excitation périodique, permet de modéliser ce signal de parole.

L'excitation périodique est caractérisée par une fréquence fondamentale appelée "pitch".

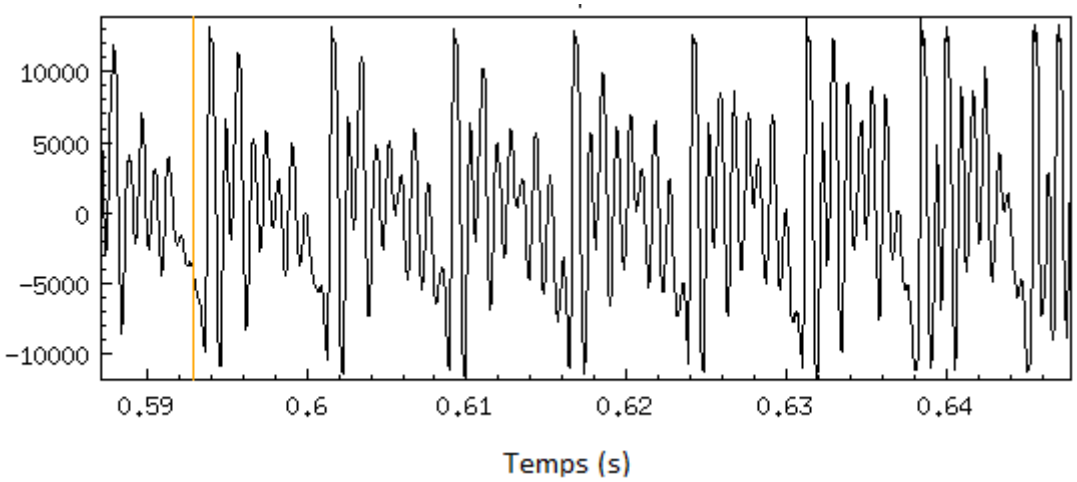


Figure 1.15 : Représentation d'une voyelle d'un signal de parole.

Le taux de passage par zéro est une caractéristique importante du signal de parole dans le domaine temporel. Il est défini comme le nombre de fois où le signal change de signe dans une durée limitée. Pour les sons voisés ce taux est relativement faible.

La fonction d'autocorrélation est un autre moyen d'estimation du degré de voisement. Pour un segment de parole S composé de N échantillons, la fonction d'autocorrélation est donnée par :

$$Cor(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} s(n).s(n-\tau), \qquad (1.1)$$

où $s(n)$ est la valeur du $n^{\text{ième}}$ échantillon du segment S.

Le voisement peut être estimé par la fonction d'autocorrélation car pour un signal périodique cette fonction est périodique. La figure 1.16 est un exemple de fonction d'autocorrélation d'un signal voisé. La mesure de l'intervalle entre deux pics successifs permet de déterminer la fréquence fondamentale.

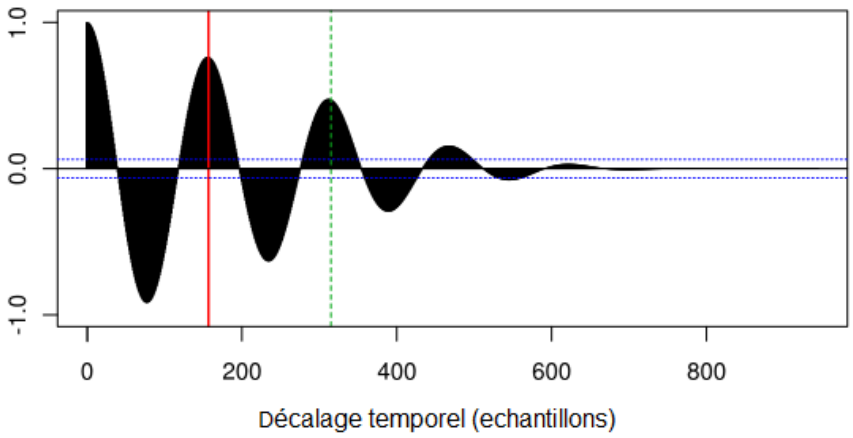


Figure 1.16 : Fonction d'autocorrélation d'un signal de parole voisé.

Une autre fonction est souvent utilisée pour l'estimation du pitch dans les sons voisés, cette fonction est appelé valeur absolue de la différence des signaux (absolute magnitude difference function AMDF). Elle est donnée par :

$$A(\tau) = \frac{1}{N}\sum_{n=0}^{N-1}|s(n) - s(n-\tau)|. \qquad (1.2)$$

La rapidité de calcul de cette fonction, du fait qu'elle n'implique aucune multiplication, et son efficacité en termes d'indication du pitch font que cette fonction soit couramment utilisée.

1.7.2. Analyse fréquentielle

En traitement de la parole, la plupart des paramètres se trouvent dans le domaine spectral. Le signal de parole est plus facile à analyser et plus systématique en fréquence qu'en temps. Aussi, le spectre de sortie du modèle de production de la parole n'est rien d'autre que le produit de la réponse fréquentielle du conduit vocal et le spectre de l'excitation. Donc, il est évident que cette sortie spectrale reflète les propriétés des réponses fréquentielles de l'excitation et du conduit vocal. Un des processus les plus communément utilisée pour obtenir le spectre d'un signal

temporel est la transformé de Fourier (TF). Mathématiquement, on définit la TF d'une séquence comme :

$$S(e^{j\omega}) = \sum_{n=-\infty}^{\infty} s(n)e^{-j\omega n}, \qquad (1.3)$$

où $s(n)$ représente l'échantillon de la séquence.

Vu la nature non stationnaire des signaux de parole, due au changement du système phonatoire dans le temps, il est impératif de considérer des séquences courtes en temps telle que les caractéristiques du conduit vocal restent inchangées. On montre qu'une durée de 10 à 30 ms permet de vérifier cette supposition, on dira que le signal est quai stationnaire.

On définit, alors, une nouvelle forme de la transformée de Fourier dépendant du temps par :

$$S_m(e^{j\omega}) = \sum_{n=-\infty}^{\infty} w(m-n).s(n)e^{-j\omega n}, \qquad (1.4)$$

où $w(n)$ est une fonction de fenêtrage qui est nulle sauf pour une durée finie. La variable m est utilisée pour sélectionner la partie de la séquence à analyser. La fonction S_m est appelée transformée de Fourier à court terme.

Pour des raisons de calcul, on échantillonne la fréquence de l'équation (1.4), on obtient, par conséquent, la transformée de Fourier discrète à court terme (Short Time Fourier Transform STFT) comme suit :

$$S_m[k] = \sum_{n=0}^{N-1} w[m-n].s[n]e^{-j2\pi nk/N}, \quad pour \quad k = 0, \ldots, M-1 \qquad (1.5)$$

L'amplitude du spectre du signal de parole, définie par la DFT, permet d'extraire plusieurs caractéristiques utiles comme la fréquence fondamentale d'un son voisé à travers la séparation des pics. Les plus larges pics sont appelés "formants", ils résultent des résonnances du conduit vocal.

Pour réduire la distorsion spectrale due au fenêtrage, la fonction w(n) doit satisfaire deux conditions. Premièrement, un lobe principal étroit et fort. Deuxièmement, une large atténuation des lobes secondaires. Comme ces deux conditions sont contradictoires, un compromis doit être fait, ce qui a donné lieu a plusieurs fenêtres. Le tableau 1.1 résume les propriétés de fenêtres utilisées dans la pratique.

Bien que la fenêtre rectangle soit la plus simple à construire, ses niveaux de lobes secondaires très élevées font qu'elle soit très peu utilisée. Dans le domaine des signaux de parole, la fenêtre la plus utilisée est celle de Hamming. Elle présente le meilleur compromis entre les deux conditions.

Tableau 1.1 : Propriétés de quelques fenêtres connues [13].

Nom de la fenêtre	$w[n]$ dans $-\frac{N-1}{2} \leq n \leq \frac{N-1}{2}$	Largeur du lobe principal $\left(\times \frac{\pi}{N}\right)$	Niveau des lobes secondaires [dB]
Rectangle	1	4	-13.3
Hanning	$\frac{1}{2}\left(1+\cos\left(\frac{2\pi n}{N}\right)\right)$	8	-31.5
Hamming	$0.54+0.46\cos\left(\frac{2\pi n}{N}\right)$	8	-42.7
Blackman	$0.42+0.5\cos\left(\frac{2\pi n}{N}\right)$ $+0.08\cos\left(\frac{4\pi n}{N}\right)$	12	-58.1

1.7.3. Analyse prédictive

L'analyse prédictive linéaire appelée aussi modélisation auto régressive (AR) fait partie des techniques les plus puissantes pour l'analyse des signaux de parole. Elle est utilisée pour estimer les paramètres de la parole comme le pitch, les formants, ou les fonctions d'air du conduit vocal. L'intérêt de cette méthode réside dans sa précision d'estimation et sa relative vitesse de calcul des différents paramètres de la parole [14]. Cette technique repose sur l'approximation de l'échantillon de parole par la combinaison linéaire des échantillons passés de parole.

Une bonne approximation des signaux de parole est assurée par un filtre tout pôle avec un nombre suffisant de pôles [15]. La réponse fréquentielle du système de la figure 1.13 est dans ce cas donnée par :

$$H(z) = \frac{S(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)}, \tag{1.6}$$

où p est l'ordre des coefficients de l'analyse de prédiction linéaire. Le filtre inverse $A(z)$ est définie par :

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}. \tag{1.7}$$

La transformée en z inverse de l'équation (1.6) donne :

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + e[n], \tag{1.8}$$

La prédiction de l'échantillon présent à partir de la combinaison linaire des p échantillons passés donne :

$$\tilde{s}[n] = \sum_{k=1}^{p} a_k s[n-k], \tag{1.9}$$

L'erreur de prédiction de cette approximation est alors :

$$e[n] = s[n] - \tilde{s}[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k], \tag{1.10}$$

L'approche de base utilisée pour trouver les coefficients de prédiction est la minimisation de l'erreur quadratique moyenne E_m d'un court segment de parole $s_m[n]$. Cette erreur est donnée par:

$$E_m = \sum_n e_m^2[n] = \sum_n \left(s_m[n] - \sum_{k=1}^{p} a_k s_m[n-k] \right)^2 , \tag{1.11}$$

En calculant la dérivée de l'équation (1.11), puis en la mettant égale a 0. Nous obtenons [16] :

$$\sum_n e_m[n] s_m[n-i] = 0 , \qquad pour \quad 1 \le i \le p \tag{1.12}$$

On obtient alors un produit scalaire nul entre les vecteurs d'échantillons e_m et s_m. Cette condition est connue par le principe d'orthogonalité.

En remplaçant e_m par sa valeur, l'équation (1.12) devient :

$$\sum_n s_m[n-i] s_m[n] = \sum_{j=1}^{p} a_j \sum_n s_m[n-i] s_m[n-j] . \;\; i = 1,2,\dots,p \tag{1.13}$$

En utilisant les coefficients de corrélation ϕ_m, l'équation (1.13) peut s'écrire telle que:

$$\sum_{j=1}^{p} a_j \phi_m[i,j] = \phi_m[i,0], \qquad i = 1,2,\dots,p, \tag{1.14}$$

avec :

$$\phi_m[i,j] = \sum_{n} s_m[n-i]s_m[n-j]. \tag{1.15}$$

Les équations (1.14) sont appelées les équations de Yule- Walker.

La solution du système des p équations permet de déterminer les p coefficients LPC qui minimisent l'erreur de prédiction. Les méthodes d'autocorrélation ou de covariance peuvent être utilisées pour la résolution des équations de Yule- Walker. Une des variantes de la méthode d'autocorrélation est connue pour son efficacité de traitement, elle est basée sur l'algorithme Levinson-Durbin. Dans cette méthode nous supposons que $s_m[n]$ est nul a l'extérieur de l'intervalle 0 ≤ n < N.

$$E_m = \sum_{n=0}^{N+p-1} e_m^2[n]. \tag{1.16}$$

Avec cet intervalle, l'équation (1.15) devient :

$$\phi_m[i,j] = \sum_{n=0}^{N+p-1} s_m[n-i]s_m[n-j] = \sum_{n=0}^{N-1-(i-j)} s_m[n]s_m[n+i-j], \tag{1.17}$$

ou bien :

$$\phi_m[i,j] = R_m[i-j], \tag{1.18}$$

où R_m est l'autocorrélation de la séquence $s_m[n]$.

Le remplacement du résultat donné par l'équation (1.18) dans l'équation (1.14) donne :

$$\sum_{j=1}^{p} a_j R_m[|i-j|] = R_m[i]. \tag{1.19}$$

L'écriture matricielle correspondant à cette équation est alors :

$$\begin{pmatrix} R_m[0] & R_m[1] & R_m[2] & & R_m[p-1] \\ R_m[1] & R_m[0] & R_m[1] & \cdots & R_m[p-2] \\ R_m[2] & R_m[1] & R_m[0] & & R_m[p-3] \\ & \vdots & & \ddots & \vdots \\ R_m[p-1] & R_m[p-2] & R_m[p-3] & \cdots & R_m[0] \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} R_m[1] \\ R_m[2] \\ R_m[3] \\ \vdots \\ R_m[p] \end{pmatrix} \tag{1.20}$$

Cette matrice est symétrique et tous les éléments de la diagonale sont égaux. On l'appelle Toeplitz. La méthode récursive de Levinson-Durbin exploite ces caractéristiques pour aboutir à un algorithme efficace. Cet algorithme se développe comme suit :

- **Initialisation :**

$$E^0 = R[0] \tag{1.21}$$

- **itérativement :**

$$k_i = \left(R[i] - \sum_{j=1}^{i-1} a_j^{i-1} R[i-j] \right) / E^{i-1}, \tag{1.22}$$

$$a_i^i = k_i \tag{1.23}$$

$$a_j^i = a_j^{i-1} - k_i a_{i-j}^{i-1}\,, \qquad 1 \le j < p \tag{1.24}$$

$$E^i = (1 - k_i^2) E^{i-1} \tag{1.25}$$

- **Solution finale :**

$$a_j = a_j^p\,, 1 \le j \le p \tag{1.26}$$

où k_i sont appelés les coefficients de réflexion. Ces coefficients sont bornés entre -1 et 1.

1.7.4. Analyse cepstrale

L'étude de la production de la parole nous a conduit à une supposition fondamentale qui représente le signal de parole comme sortie d'un système linéaire invariant dans le temps. Par conséquent, l'excitation et le filtre du système sont reliés par le produit de convolution.

La transformation homomorphique permet de convertir la convolution en une sommation, par :

$$\hat{e}[n] = D(s[n]) = D(e[n] * h[n]) = \hat{e}[n] + \hat{h}[n]. \tag{1.27}$$

L'analyse cepstrale est une transformation homomorphique qui vise à séparer la source du filtre d'un système.

En parole, le son voisé $x[n]$ peut être considéré comme la réponse articulatoire d'un conduit vocal $h[n]$, qui équivaut au filtre, a l'excitation d'une source pseudopériodique $g[n]$. On aura alors ;

$$x[n] = \sum_{m=0}^{L} g[m].\, h\,[n-m]. \tag{1.28}$$

Où L est la taille du résultat de la convolution. Cette équation (1.28) est équivalente à :

$$X(\omega) = G(\omega)H(\omega). \tag{1.29}$$

où $X[k], G[k], et\ H[k]$ sont les transformées de Fourier discrète (TFD) de $x[n]$, $g[n]$, et $h[n]$ respectivement.

L'application du logarithme au module de $X[k]$ donne :

$$log|X[k]| = log|G[k]| + log|H[k]|. \tag{1.30}$$

Le cepstre, qui est la transformée de Fourier inverse du $log|X(\omega)|$, est :

$$c(\tau) = TF^{-1}log|X[k]| = TF^{-1}log|G[k]| + TF^{-1}log|H[k]|. \tag{1.31}$$

Principalement, le premier terme de la partie droite de l'équation (1.31) montre la formation d'un pic dans la région haute quéfrence (qui est le domaine temporel dans l'analyse cepstrale), et le second terme représente une concentration dans la région basse quéfrence (de 0 a 2 ou 4 ms) [15].

Dans le cas des signaux de parole, le cepstre est calculé sur un segment de parole de longueur N obtenu par fenêtrage.

En appliquant le logarithme sur l'amplitude de la transformée de Fourier de $x[n]$ le cepstre devient :

$$c_n = \frac{1}{N}\sum_{k=0}^{N-1} log|X[k]|\, e^{j2\pi kn/N} \qquad 0 \leq n \leq N-1 \qquad (1.32)$$

1.8. Applications de la biométrie

De nos jours, les technologies biométriques sont de plus en plus adoptées dans différents secteurs gouvernementaux ou commerciaux. Cela est du à la nécessité de validation et de confirmation d'identités de personne avec une grande sécurité et une très bonne efficacité. Les technologies biométriques offrent la possibilité d'élever le niveau de sécurité et de convenance au dessus des méthodes conventionnelles. En général, les applications de la biométrie peuvent être divisées en quatre parties qui sont :

- le contrôle d'accès qui peut être lui même subdivisée en contrôle d'accès virtuel (P.ex. pour accéder a une ressource de données ou un PC) et en contrôle d'accès physique (P.ex. pour accéder a des lieux sécurisés).
- L'authentification des personnes.
- L'enregistrement et l'identification de personnes.
- La personnalisation.

A partir de ces grandes familles d'applications biométriques, une répartition plus détaillée de ces applications existe dans la réalité.

1.8.1. Répression

Appelée aussi la science forensique. Elle désigne la contribution des sciences, en particulier les sciences de la nature, à la justice. Ainsi, des démarches scientifiques et des méthodes sont appliquées dans l'étude des traces qui prennent leur origine dans une activité criminelle, ou un litige civile, réglementaire ou administrative [17].

Les modalités biométries sont souvent utilisées dans ce domaine criminalistique. Les empreintes digitales sont utilisées dans une application particulière, pour confirmer ou déterminer l'identité d'un individu parmi un ensemble de personnes enregistrées dans un centre de police. Le Bureau fédéral d'investigations (FBI) détient actuellement une des plus larges bases de données existantes. Elle contient des dizaines de millions d'enregistrements d'empreintes digitales appartenant à des personnes civiles et des criminelles. L'analyse de la voix est aussi utilisée en forensiques. Son utilisation pour des investigations policières repose sur le fait que chaque personne peut être identifiée à partir d'un échantillon de sa voix. Un suspect peut laisser des enregistrements de sa voix sur le téléphone, le Voice Mail, un répondeur, ou dans un enregistreur caché, et par la suite, il peut être utilisé comme preuve [18].

1.8.2. Vérification d'antécédents

Les technologies biométriques sont utilisées pour faire des vérifications d'antécédents comme condition de recrutement de personnes dans divers professions gouvernementales ou commerciales. Même si la vérification se fait dans les mêmes bases de données utilisées pour la recherche des criminelles, la différence dans cette application est que les caractéristiques biométriques des personnes ne seront pas conservées, ils seront détruites juste après avoir transmis les résultats de vérification aux demandeurs d'information.

1.8.3. Surveillance

La biométrie permet de localiser, de suivre, et d'identifier les personnes dans une zone bien déterminée. Ainsi, la reconnaissance du visage permet d'automatiser le processus de surveillance en utilisant des caméras CCTV par exemple (figure 1.17).

Figure 1.17 : Camera de surveillance CCTV.

Ces systèmes de surveillance permettent d'alerter les autorités concernées de la présence de l'individu d'intérêt.

1.8.4. Le contrôle aux frontières

L'utilisation de la biométrie pour le contrôle aux frontières est née pour réduire le temps d'attente des passagers et des personnels qui ne cessent d'accroitre. Elle vise aussi à accroitre la sécurité globale. Selon les standards internationaux, les passeports biométriques seront achevés en 2015. La figure 1.18 illustre la procédure de contrôle automatique dans un aéroport.

Figure 1.18 : Le contrôle aux frontières automatisé.

Ces passeports utilisent l'iris, l'empreinte digitale, et la reconnaissance du visage pour ses applications de contrôle. Ainsi, le bon déploiement de ces ressources permet de bien contrôler les passagers à risque.

1.8.5. Lutte contre la fraude

Dans le secteur public, la technologie biométrique peut servir à réduire la fraude en empêchant les individus de réclamer des prestations sous multiples identités. La reconnaissance basée sur l'iris et l'empreinte digitale sont des exemples de moyens utilisés pour leur dissuader contre un éventuel cumul.

1.8.6. Gestion du temps et de la présence des employés

La biométrie peut servir, dans des applications commerciales, à aider les gestionnaires des employés. Dans cette application, les appareils sont utilisés pour poursuivre la présence des travailleurs.

Figure 1.19 : Appareil biométrique pour la gestion d'accès
et de présence des travailleurs.

On retrouve dans le marché, la reconnaissance par les empreintes digitales et la géométrie qui servent à la gestion des heures de travail et de la paie (figure 1.19).

1.8.7. Reconnaissance du consommateur

L'authentification conventionnelle utilisait les cartes d'accès, les codes PIN, ou les signatures pour effecteur des transactions commerciales. Ces dernières peuvent êtres réalisées par une application biométrique se basant sur l'identité d'une personne. La biométrie permet donc de diminuer les mots de passe ainsi que les cartes. La reconnaissance des empreintes digitales est communément utilisée dans ce type d'applications.

1.8.8. Reconnaissance à distance

Cette application biométrique offre des méthodes sûres d'authentification pour l'accès à distance à des informations importantes par des utilisateurs d'appareils mobiles. Les biométriques utilisées jusqu'ici sont la voix et l'empreinte digitale. La figure 1.20 montre l'utilisation du doigt pour effectuer des paiements à distance.

Figure 1.20 : Accès a distance.

1.8.9. Protection des biens

La protection des informations numériques, ainsi que d'autres appareils sensibles d'utilisateurs non autorisés, constitue une tache très importante en biométrie. La reconnaissance d'empreinte digitale est une application très sûre pour protéger les documents sensibles.

Les traits biométriques servent aussi comme complément de sécurité aux méthodes déjà en place. L'utilisation de mots de passe et l'identification des utilisateurs sont un exemple de ce type d'applications.

1.9. Conclusion

Nous avons commencé notre thèse par décrire des notions générales sur la biométrie et la parole. Dans la première partie nous avons défini la biométrie, puis nous avons établit le schéma blocs d'un système biométrique général en vérification et en identification. Nous avons, aussi, cité les modalités les plus rencontrées en biométrie comme l'empreinte digitale, le visage, l'iris, la démarche, et la voix (parole). Cette dernière modalité a été détaillée car elle constitue la base de notre travail de reconnaissance du locuteur. Nous avons terminé notre chapitre par passer en revue les différentes applications de la biométrie.

CHAPITRE 2

SYSTEME DE RECONNAISSANCE DU LOCUTEUR

2.1. Introduction

Une des technologies biométriques émergentes est la reconnaissance du locuteur. Elle offre une approche prometteuse en termes de sécurité comparativement aux méthodes classiques basées sur les cartes d'accès ou les mots de passe. La reconnaissance du locuteur peut être définie comme la tache d'établir l'identité du locuteur en se basant sur sa voix. Elle est appelée aussi reconnaissance de la voix. La reconnaissance du locuteur est divisée en deux grandes classes : l'identification et la vérification (Figure 2.1). En identification, aucune identité n'est prétendue par le locuteur. Le système doit déterminer, automatiquement, celui qui parle. Si le locuteur appartient à un ensemble prédéfini de locuteurs connus, cette approche est dite identification du locuteur dans un ensemble fermé. En général, le système doit traiter des cas où les locuteurs peuvent ne pas êtres modélisés dans la base de données. Dans ce cas l'approche est dite identification du locuteur en ensemble ouverte.

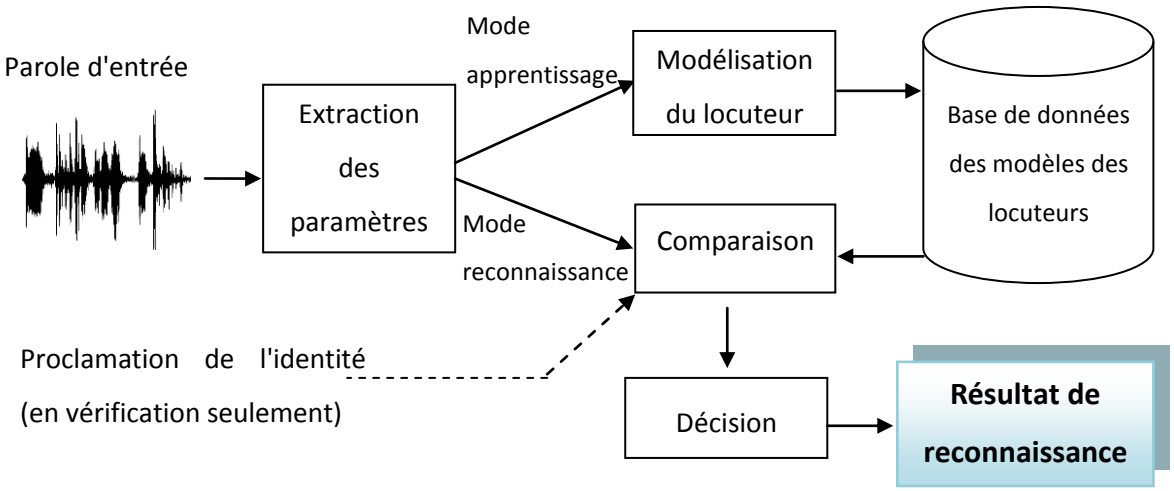


Figure 2.1 : Système de reconnaissance du locuteur.

En vérification de locuteur, le système doit déterminer si oui ou non la personne est celle qui prétend l'être. Cela implique que l'utilisateur doit fournir une identité, le système accepte ou rejette cette personne selon que la vérification a réussi ou échoué.

Une autre classification des systèmes de reconnaissance de locuteur divise ceux ci en systèmes dépendants ou indépendants du texte. Les systèmes de reconnaissance dépendants du texte sont dits "avec contrainte". Dans ce cas, on demande à l'utilisateur de prononcer soit un mot fixe (comme un mot de passe) ou une phrase. Cette information peut améliorer les performances du système de reconnaissance. Le système de reconnaissance indépendant du texte est sans contrainte. Dans ce cas, le locuteur doit être reconnu quelque soit le texte prononcé, ce qui constitue un problème plus difficile à résoudre.

2.2. Théorie de la reconnaissance de locuteur

Comme tout système de reconnaissance de formes, un système de reconnaissance du locuteur peut travailler en deux modes; appelés l'apprentissage et le test. La représentation et le prétraitement des signaux de parole sont des étapes communes aux deux modes de travail. Dans le mode d'apprentissage, le vecteur de paramètres représentant le signal de parole d'un locuteur sert à produire le modèle du locuteur en utilisant des algorithmes de classification appropriés. Les modèles résultants de l'apprentissage seront ensuite enregistrés dans une base de données. Dans le mode de test, le vecteur de paramètres correspondant à un locuteur inconnu sera comparé à un ou plusieurs modèles de la base de données. Ainsi, sur la base des scores de comparaisons faites, le système de reconnaissance décidera pour quelle personne appartient cette voix.

Dans les paragraphes suivants nous détaillerons les différentes fonctions d'un système de reconnaissance de locuteur.

2.3. Extraction des paramètres

De par sa production, la parole est variable dans le temps. Les modèles de paramétrisation sont, par conséquent, eux aussi variables dans le temps.

L'estimation de ces paramètres nécessite, dans ce cas, une analyse à court terme. Le signal de parole change tous les 10 à 30 millisecondes. Dans cet intervalle, le signal est supposé rester stationnaire, un vecteur de paramètres est extrait durant des segments de temps courts appelées trames. En général, avant de procéder à d'autres étapes d'extraction de paramètres, le signal de parole subit des prétraitements qui incluent :

- **Détection de la parole :** Le signal de parole peut contenir du silence dans différentes positions comme le début du signal, entre les mots d'une phrase, à la fin du signal. etc. Ce silence n'apporte aucune contribution à la reconnaissance du locuteur. Il doit être supprimé avant la poursuite des traitements. Plusieurs méthodes sont utilisées pour réaliser cette objective. La plus connue est la détection d'activité vocale (Voice activity detection VAD), surtout dans les transmissions téléphoniques où le silence constitue la grande partie du temps de transmission, dans chaque direction. Typiquement, la VAD exploite deux types de caractéristiques : (a) la différence spectrale entre le bruit et la parole et (b) les variations temporelles de l'énergie en court terme.

  Dans la technique VAD, la première étape consiste à calculer les énergies de toutes les trames, puis à sélectionner la valeur maximale. Le seuil de détection est, alors, fixer au dessous de ce maximum. Un autre seuil est nécessaire pour annuler les trames ayant une énergie absolue faible.

  La VAD basée sur la périodicité est une technique alternative aux techniques basées sur l'énergie, elle a été étudiée dans [19].

- **Segmentation et chevauchement :**

  Le but de la segmentation est de découper le signal de parole en petites tranches (chacune de durée 10 à 30 ms) où il peut être considéré localement comme quasi-stationnaire.

  En outre, et pour profiter de l'évolution lente du signal vocal, la segmentation permet le traitement en temps réel et facilite l'analyse des signaux sur la machine. Les ressources d'une machine étant limitées, le signal ne peut pas être traité dans sa globalité.

Si la trame est de longueur N échantillons, telles que les trames adjacentes sont séparées par un nombre d'échantillons (de 5 à 15 millisecondes environ), cela est illustré à la Figure 2.2. L'opération de chevauchement vise à ne pas perdre de données.

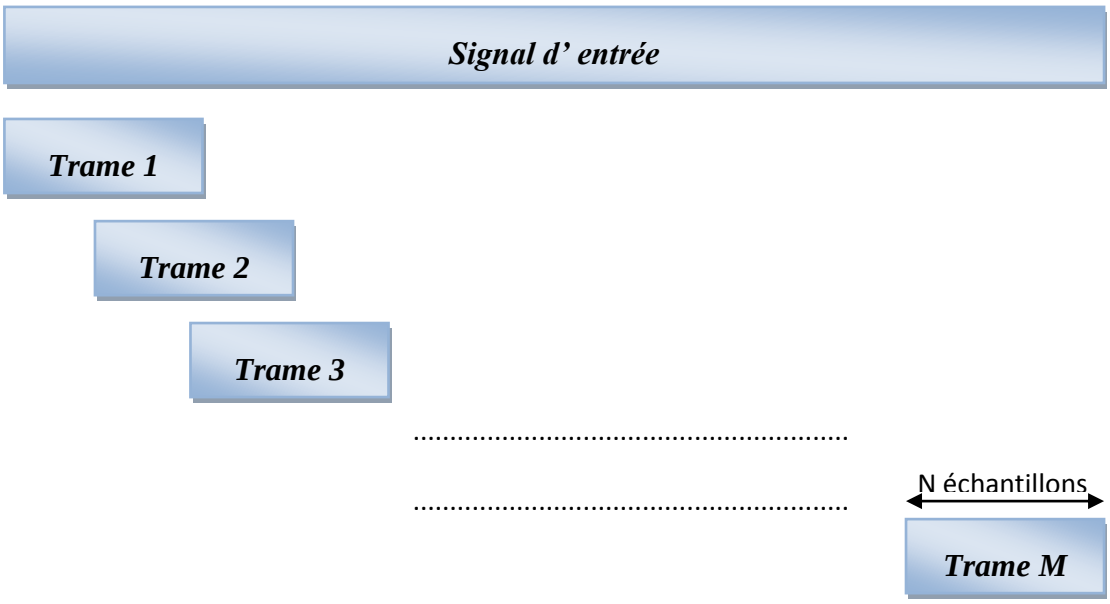


Figure 2.2 : Découpage du signal de parole en trames.

- **Fenêtrage :** Chaque trame est multipliée par une fonction fenêtre. La fenêtre Hamming est devenue la plus populairement utilisée pour son efficacité à lisser les discontinuités abruptes aux bords des trames. Comme la plupart des paramètres sont de nature spectrale, la transformée de Fourier est employée. L'effet multiplicatif de la fonction de fenêtrage dans le domaine temporel est un produit de convolution en fréquence. La pondération du signal de parole par la fenêtre de Hamming vise, d'une part, à concentrer la répartition de l'énergie sur les basses fréquences et, d'autre part, d'amoindrir les fortes variations du signal sur les bords de la fenêtre, qui risque d'entraîner une mauvaise estimation des coefficients du filtre si elles n'étaient pas atténuées.
  Cette fenêtre est donnée comme suit :

$$w[k] = \begin{cases} 0.54 + 0.46\cos\left(\dfrac{2\pi k}{N-1}\right) & si \quad 0 \le k \le N-1 \\ 0 & sinon \end{cases} \qquad (2.1)$$

Une autre méthode, plus récente [20], utilise plusieurs fenêtres superposées de différentes formes (multitaper analysis).

2.3.a. Paramètres long terme

L'information long terme fait référence aux paramètres qui sont extraits sur des régions plus longues que la trame. Ils sont considérés comme des caractéristiques en haut niveau qui capturent les informations phonétiques, prosodiques, et lexiques. Les paramètres prosodiques capturent les variations dans l'intonation, le timing, et l'intensité, qui sont spécifiques au locuteur [21]. Comme de tels paramètres s'étends au delà d'une trame, ils font partie des paramètres long terme. Ces paramètres ne sont pas toujours faciles à extraire, et n'ont attiré l'attention des chercheurs que durant ces dernières années.

Le pitch et la dynamique de l'énergie sont des exemples de paramètres utilisés dans la reconnaissance des locuteurs. Le pitch d'un locuteur est influé par la longueur et la masse des cordes vocales dans le larynx. Les locuteurs varient dans la plage de fréquences qu'ils sont capables de produire et la plage de fréquences qu'ils utilisent dans leurs discours quotidiens. Donc, le pitch peut être considéré comme un paramètre discriminant de locuteurs. Le changement d'intensité ou d'énergie est beaucoup moins relié a l'anatomie de la personne, il est essentiellement pertinent au marquage de stress et l'expression des émotions.

2.3.b. Paramètres spectraux court terme

De nos jours, la plupart des systèmes de reconnaissance de locuteur utilisent les paramètres spectraux à court terme. Ces paramètres sont extraits à partir de petits segments du signal de parole. La paramétrisation spectrale de la parole peut être divisée en méthodes paramétriques et non paramétriques. Les méthodes paramétriques supposent que les données suivent un certain modèle, et le problème d'extraction de paramètres revient à l'estimation des paramètres de ce modèle. Si le modèle ne correspond pas aux données, les performances du système utilisant ces paramètres seront dégradées et l'estimation est dans ce cas biaisée. La méthode d'extraction de paramètres LPC est l'une des méthodes les plus utilisées dans cette catégorie. Les méthodes non paramétriques sont basées sur les bancs de filtres.

Pour générer les paramètres à partir du spectre du signal de parole, on tient compte de la résolution non linéaire des fréquences, tout comme le fait l'appareil auditive de l'être humain. Les coefficients mel cepstraux (Mel Frequency Cepstral Coefficients MFCC) et les coefficients de prédiction linéaire perceptuelle (Perceptual Linear Prediction PLP) sont les méthodes les plus largement utilisées en reconnaissance de la parole. Leur succès découle de l'utilisation des bancs de filtres non linéairement espacés basés sur le fonctionnement perceptuel humain et de la robustesse des coefficients MFCC, ainsi que la flexibilité obtenue dans l'analyse cepstrale.

2.3.c. Coefficients mel cepstraux

Les coefficients MFCC ont été introduits au début des années 1980 pour les applications de la reconnaissance de la parole et depuis sont aussi adoptés pour d'autres applications comme la reconnaissance du locuteur [22]. Les principales étapes impliquées dans le calcul des paramètres MFCC sont montrées dans la figure 2.3.

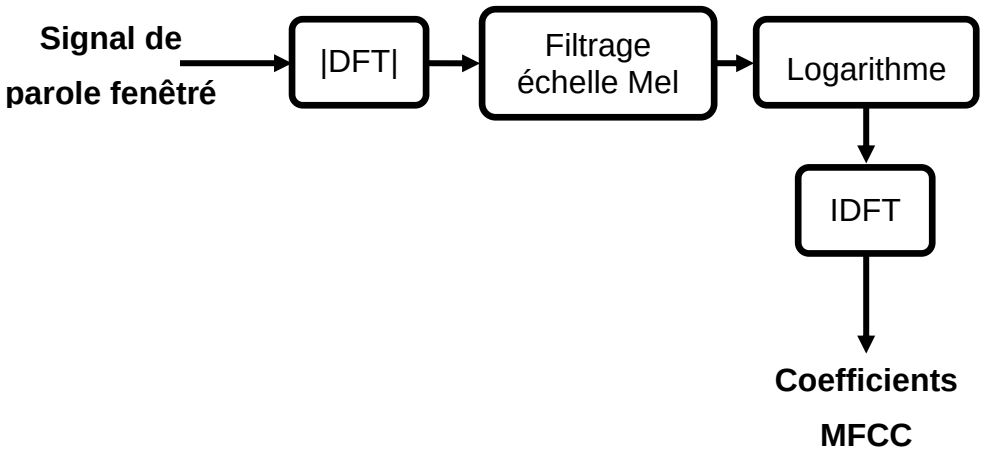


Figure 2.3 : Etapes de calcul des paramètres MFCC.

La transformée de Fourier discrète (TFD) est appliquée a chaque trame du signal parole donnant des valeurs spectrales complexes. L'information de phase de ces nombres complexes est ignorée pour ne garder que l'information de l'amplitude du

spectre. L'implémentation rapide de la TFD appelée transformée de Fourier rapide (TFR) (Fast Fourier Transform FFT) est utilisée en pratique. La forme globale de l'amplitude du spectre, connue par l'enveloppe du spectre, contient des informations sur les propriétés de résonnance du conduit vocal. Il a été montré que l'enveloppe spectrale est la partie la plus informative du spectre dans la reconnaissance du locuteur [23]. Le modèle le plus simple de l'enveloppe spectrale utilise un ensemble de filtres passe bande. En général, pour le calcul des paramètres MFCC, les filtres les plus utilisées sont de forme triangulaire.

Soit un banc de filtres constitué de M filtres ($m = 1, 2, \ldots, M$), tel que le filtre m est de forme triangulaire, donné par :

$$H_m(k) = \begin{cases} 0 & k < f[m-1] \\ \dfrac{2(k - f[m-1])}{(f[m+1]f[m-1])(f[m]f[m-1])} & f[m-1] \le k \le f[m] \\ \dfrac{2(f[m+1] - k)}{(f[m+1]f[m-1])(f[m+1]f[m])} & f[m] \le k \le f[m+1] \\ 0 & k > f[m+1] \end{cases} \quad (2.2)$$

Ces filtres calculent la moyenne du spectre autour de chaque fréquence centrale avec l'augmentation des bandes fréquentielles, ces bandes sont montrées dans la figure 2.4.

Soient, f_b et f_h les fréquences basse et haute, respectivement, du banc de filtres exprimées en Hz, F_e la fréquence d'échantillonnage en Hz, M le nombre de filtres, et N la taille de la FFT. Les points $f[m]$ sont uniformément espacés sur l'échelle Mel, avec :

$$f[m] = \left(\frac{N}{F_e}\right) Mel^{-1}\left(Mel(f_b) + m\frac{Mel(f_h) - Mel(f_b)}{M+1}\right) \quad (2.3)$$

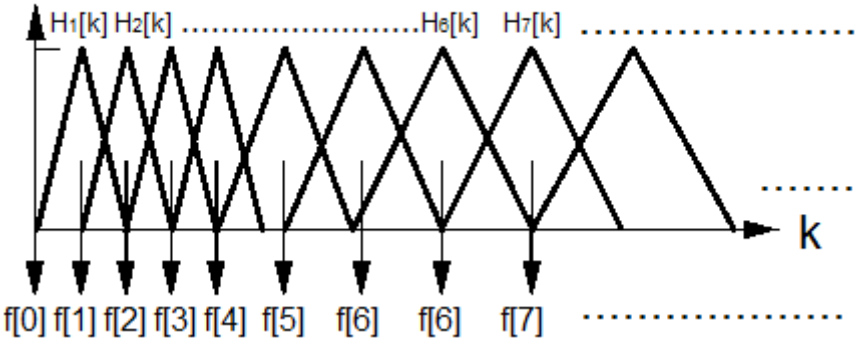


Figure 2.4 : Représentation des filtres triangulaires utilisés dans la méthode MFCC.

où l'échelle Mel est donné par :

$$Mel(f) = 1125\, ln(1 + f/700) \tag{2.4}$$

Justifiée par les études psycho-acoustiques, la plage des basses fréquences est souvent représentée par une plus haute résolution, en lui allouant plus de filtres avec des bandes de fréquences étroites [24]. Bien que les sous bandes d'énergies résultantes aient été directement utilisés comme paramètres [25, 26], généralement, d'autres transformations sont utilisées pour réduire encore la dimensionnalité des paramètres. Ceci est nécessaire puisque un spectre de puissance de dimension 256 par exemple, donne une représentation trop détaillée sur le spectre, avec un lissage du spectre obtenu en prenant une dizaine de valeurs par trame, ou plus, la représentation spectrale devient plus efficace. En outre, un lissage additionnel peut être réalisé d'une manière significative en appliquant le logarithme. La dernière étape consiste à convertir les K valeurs spectrales résultant du logarithme des bancs de filtres S_k, en n coefficients ceptraux C_n en utilisant la transformée en cosinus discrète (Discrete Cosine Transform DCT), tel que :

$$C_n = \sum_{k=1}^{K} (logS_k).cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right] \quad (2.5)$$

Contrairement aux paramètres spectraux qui sont très corrélés, les paramètres produits par analyse cepstrale sont moins corrélés. En plus, on peut inclure le coefficient C_0, obtenu en remplaçant n par la valeur nulle dans l'équation (2.5), qui représente la moyenne de l'énergie logarithmique de la trame. Comme cette moyenne n'est pas porteuse d'une grande information discriminante spécifique au locuteur, il est commun de l'exclure des autres paramètres.

2.3.d. Coefficients de prédiction linéaire perceptuelle

La technique PLP combine divers concepts psycho-acoustiques du système auditif humain. Comme pour l'obtention des coefficients de prédiction linéaire (LPC), la méthode récursive de Levinson-Durbin est aussi utilisée en PLP [27]. La différence ici est qu'au lieu de calculer l'autocorrélation dans le temps, elle est calculée comme la transformée de Fourier inverse de l'énergie spectrale $|X(f)|^2$. Pour les raisons perceptuelles des fréquences dans l'oreille humaine, une analyse par un banc de filtres, appelées "bandes critiques", est effectuée. Ainsi, ce banc de filtres est appliquée dans l'échelle Bark qui s'exprime en fonction de fréquence par :

$$B(f) = 6ln\left(\frac{f}{600} + \left(\frac{f}{600}+1\right)^{1/2}\right) \quad (2.6)$$

f étant la fréquence en Hz.

Alors qu'en MFCC les filtres avaient des formes triangulaires, en PLP cette forme est donnée par l'équation suivante :

$$\psi = \begin{cases} 0 & f_{c(Bark)} < -2.5 \\ 10^{(f_{Bark} - f_{c(Bark)} + 0.5)} & -2.5 \leq f_{Bark} - f_{c(Bark)} \leq -0.5 \\ 1 & -0.5 < f_{Bark} - f_{c(Bark)} < 0.5 \\ 10^{-0.25(f_{Bark} - f_{c(Bark)} + 0.5)} & 0.5 \leq f_{Bark} - f_{c(Bark)} \leq 1.3 \\ 0 & f_{Bark} - f_{c(Bark)} > 1 \end{cases} \quad (2.7)$$

où $f_{c(Bark)}$ est la fréquence centrale d'un filtre en échelle de Bark.

La Figure 2.5 illustre le banc de filtres en échelle de Bark sur une échelle de fréquence linéaire [15].

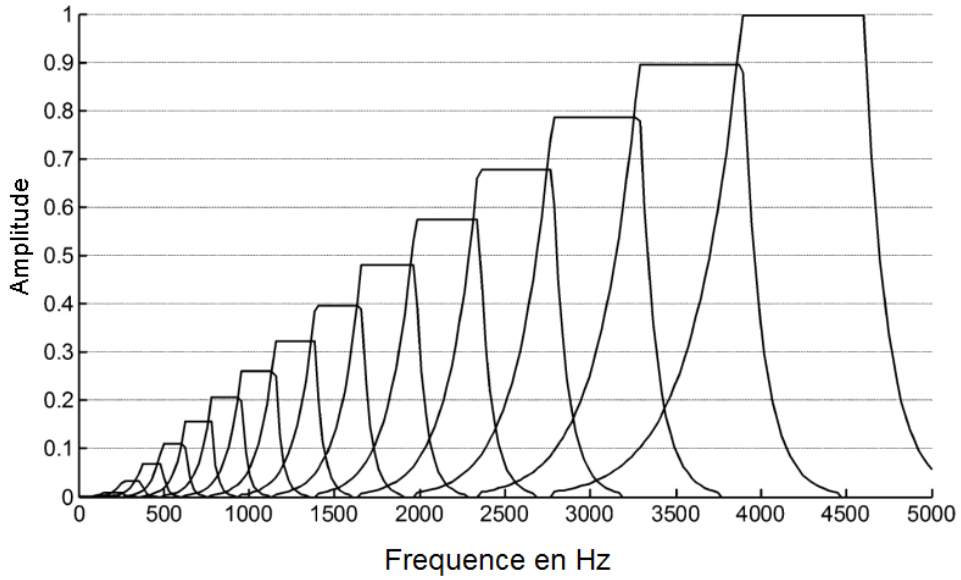


Figure 2.5 : Représentation de 16 filtres en échelle de Bark
pour une largeur de bande de 5000 Hz [3].

Pour compenser la sensibilité inégalée de l'oreille humaine aux différentes fréquences, les courbes d'isosonie peuvent être utilisées.

Dans l'étape suivante, une compression d'intensité est appliquée sur les bandes d'énergie spectrales. Elle décrit la relation non linéaire entre l'intensité du signal et de l'intensité sonore perçue. Dans la technique PLP, cette relation est une relation de racine cubique.

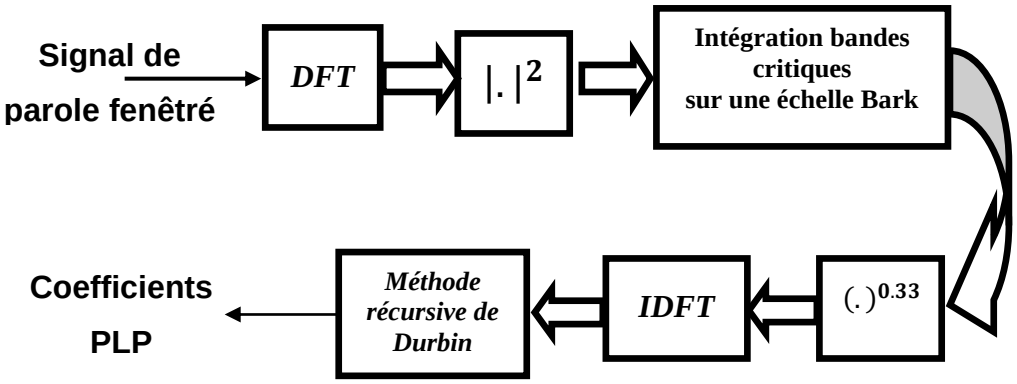


Figure 2.6 : Etapes de calcul des paramètres PLP.

2.3.e. Paramètres dynamiques de la parole

Les dérivées dans le temps des paramètres spectraux de la parole sont nommées les paramètres dynamiques. Ainsi, les performances du système de reconnaissance du locuteur peuvent être améliorées par l'ajout de ces paramètres dynamiques aux paramètres statiques classiques. Les dérivées d'ordre un sont connues par les paramètres delta. Ils peuvent être calculés par la pondération linéaire des différences entre les vecteurs statiques voisins [28].

$$\nabla_t = \frac{\sum_{l=1}^{L}(C_{t+l} - C_{t-l})}{2\sum_{l=1}^{L} l^2} \tag{2.8}$$

où ∇_t est le coefficient delta à l'instant t, calculé en fonction des coefficients statiques $C_{t+l}\ et\ C_{t-l}$, l représente la taille de la fenêtre delta.

Les paramètres double delta sont définis en utilisant l'operateur delta, ultérieurement définit, en l'appliquant à la dérivée première des coefficients.

2.3.f. Log-Energie

L'inclusion de l'énergie de chaque trame du signal parole apporte une augmentation significative au taux de reconnaissance. Souvent, le moyen le plus

utilisé pour mesurer cette quantité est le calcul du logarithme de l'énergie ($log\ Ei$) des données d'une trame de la parole dans le domaine temporelle. Telle que :

$$log\ E_i = log \sum_{n=1}^{N} s_n^2 \tag{2.9}$$

où s_n et N sont le $n^{ième}$ échantillon et le nombre d'échantillons de la trame i, respectivement.

L'utilisation fréquente de ce paramètre avec les autres paramètres cités est due à sa faible variabilité.

2.4. Approches de modélisation

La tâche de reconnaissance du locuteur implique la comparaison d'un locuteur inconnu avec des locuteurs connus dans une base de données. Sur la base de cette comparaison le locuteur correspondant sera choisi. L'utilisation directe de vecteurs de paramètres représentant chaque locuteur n'est pas très pratique lorsque les vecteurs d'apprentissage sont larges. La modélisation est un moyen efficace de compression des données permettant d'obtenir un petit ensemble de points et qui doit bien représenter le locuteur.

Essentiellement, les techniques de modélisation peuvent appartenir à : l'approche vectorielle, l'approche connexionniste, ou l'approche statistique.

Une autre partition divise la modélisation en approches paramétriques et approches non paramétriques. Les modèles paramétriques supposent que la distribution des données suit une forme connue a priori comme le modèle Gaussien. Dans le cas des modèles non paramétriques, aucune hypothèse sur la distribution des données n'est faite, comme pour la quantification vectorielle. Dans ce qui suit nous décrirons les approches de modélisation les plus répandues.

2.4.a. Quantification vectorielle

L'utilisation de cette approche est due, principalement, à sa simplicité d'implémentation et sa bonne précision. La quantification vectorielle est le processus de représenter les vecteurs de paramètres, qui se trouvent dans un large espace vectoriel et qui correspond à un certain locuteur, par un nombre fini de régions dans l'espace. Chaque région est appelée " cluster " qui peut être représenté par son centre appelé " mot code ". L'ensemble de ces mots codes forme un dictionnaire. Cette approche est une méthode de compression avec perte basée sur le principe de codage en blocs.

La figure 2.7 illustre la formation des mots codes appelés aussi " centroïdes ", en cercles noirs pour le locuteur 1 et en triangles noirs pour le locuteur 2, à partir d'échantillons des deux locuteurs.

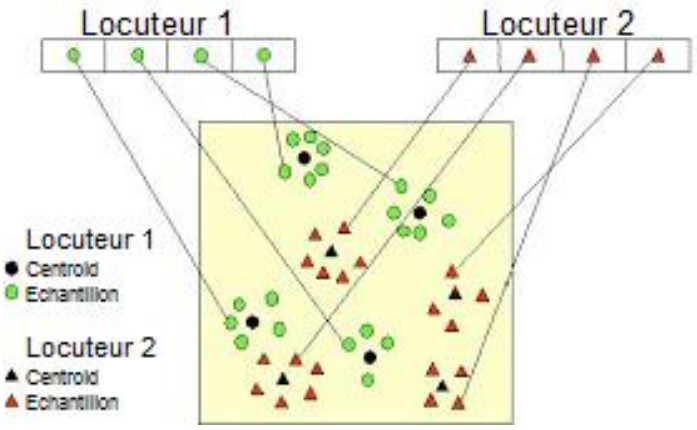


Figure 2.7 : Illustration de l'utilisation de la quantification vectorielle.

En phase d'apprentissage, nous obtenons donc des clusters. La distance entre le vecteur et le centroïde du dictionnaire est appelée "distorsion de la quantification vectorielle". Dans la phase de reconnaissance, le signal d'entrée d'un locuteur inconnu est vectoriellement quantifié avec tout le dictionnaire vectoriel. Par conséquent, une distorsion totale est calculée. Le locuteur, représenté dans le dictionnaire et ayant la distorsion totale la plus petite, est identifié comme celui qui a prononcé la parole d'entrée.

L'augmentation de la taille du dictionnaire permet de mieux représenter le locuteur, mais le système devient moins rapide et plus gourmant en mémoire. Il est nécessaire dans ce cas qu'un bon compromis entre ces deux facteurs soit fait. Pour construire le dictionnaire, les méthodes LBG (due to Linde, Buzo and Gray) et K moyennes (K-means) sont souvent utilisées [29].

2.4.b. Alignement temporel dynamique (dynamic time warping DTW)

C'est une méthode non paramétrique séquentielle, applicable en mode dépendant du texte [30]. Cette technique a été largement utilisée en reconnaissance du locuteur afin d'obtenir la distorsion complète entre deux énoncés de parole. Cette mesure de distorsion complète est calculée à partir de l'accumulation des distances entre deux vecteurs de paramètres. Chaque énoncé est représenté par une séquence de vecteurs caractéristiques. La variation temporelle de l'énoncé de référence et de l'énoncé de test est normalisée par alignement non linéaire de la séquence des vecteurs caractéristiques en utilisant l'algorithme de programmation DTW.

La figure 2.8 représente l'utilisation de l'algorithme DTW pour comparer deux énoncés, qui sont des trames de coefficients de paramètres de parole, la fonction d'alignement sert à la normalisation des énoncés de référence et de test. La distance euclidienne cumulée peut servir à la classification.

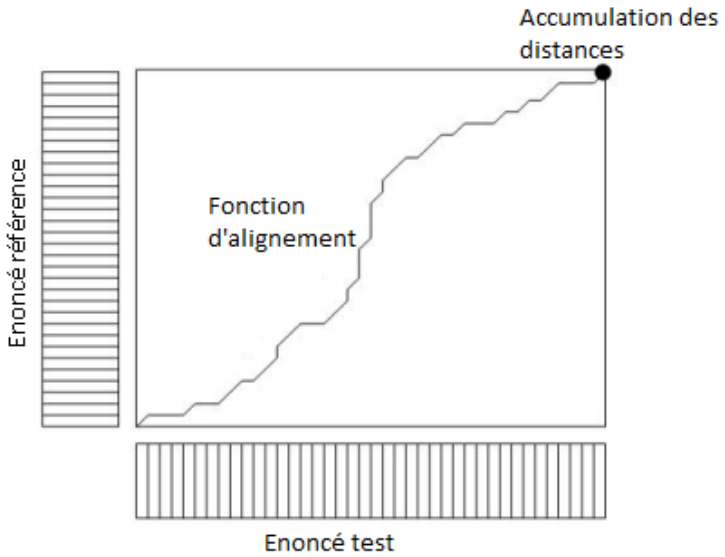


Figure 2.8 : Comparaison de deux séquences avec utilisation d'alignement temporel

Malgré ses performances en termes de taux de reconnaissance, l'utilisation de la technique d'alignement temporel dynamique en reconnaissance de locuteur a été abandonnée au profit des techniques statistiques comme les modèles de mélanges de gaussiennes.

2.4.c. Réseaux de neurones artificiels (Artificial Neural Network ANN)

L'approche ANN ou réseaux connexionnistes s'inspirent fortement du fonctionnement du cerveau humain. Dans les réseaux de neurones artificiels de nombreux processeurs appelés cellules ou unités, capables de réaliser des calculs élémentaires, sont structurés en couches successives capables d'échanger des informations au moyen de connexions qui les relient. On dit de ces unités qu'elles miment les neurones biologiques. La flexibilité de calcul du cerveau humain est due à son grand nombre de neurones dans une maille d'axones et de dendrites. Pour communiquer, un neurone envoi des signaux vers d'autres neurones à travers l'axone, qui est une fibre. Enfin, c'est à travers des formations spéciales au bout de l'axone, les synapses, que le signal provenant du neurone arrive aux dendrites, ou à d'autres neurones.

En reconnaissance du locuteur, l'approche connexionniste se résume à une tache de classification. Un modèle client se présente sous la forme d'un ou plusieurs réseaux de neurones pour lequel la séquence de vecteurs d'apprentissage du client concerné ainsi que celles des autres clients du système sont fournies en entrée. Lors de la reconnaissance, la vraisemblance pour qu'une séquence de vecteurs de test soit produite par un réseau de neurones est calculée. Les modèles de réseaux de neurones utilisent une topologie particulaire pour les interactions et les corrélations des connections des unités neuronales.

Le perceptron multicouche (Multi Layer Perceptron MLP) est un réseau de neurone très utilisé qui est composé d'une couche en entrée, une ou plusieurs couches cachées et une couche en sortie. Un exemple de MLP est donné dans la figure 2.9.

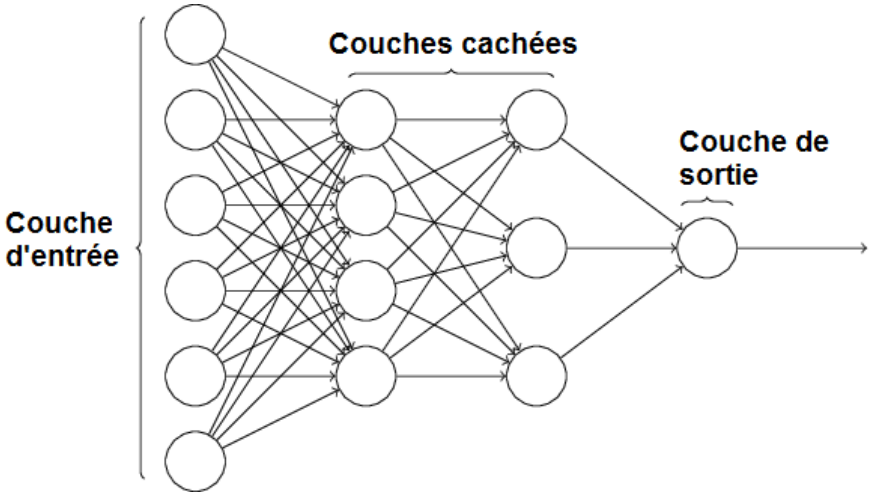


Figure 2.9 : Les perceptrons multicouche à 4 couches. Une couche d'entrée, deux couches cachées, et une couche de sortie.

Ce réseau est composé d'une couche d'entrée, d'une couche de sortie, et de deux couches au milieu qui sont cachées. Ces dernières peuvent êtres des extracteurs de paramètres.

Le principal inconvénient de la modélisation par les réseaux de neurones est la complexité d'apprentissage. En outre, elle pose le problème de l'ajout d'un nouveau client qui nécessite le réapprentissage de tous les modèles. En effet, une nouvelle phase de classification est nécessaire afin de prendre en compte le nouveau client au sein du processus de discrimination entre les locuteurs [31].

2.4.d. Machines à vecteurs de support (Support Vector Machine SVM)

Les SVM sont des classificateurs binaires et statiques. C'est à dire qu'ils permettent de créer une surface de décision entre deux classes définies dans un même espace. Pour cela, ils construisent une frontière de décision par projection des caractéristiques provenant d'un espace d'origine dans un espace de caractéristiques de dimension supérieure (voir infini) dans le but de rendre les classes linéairement séparables.

L'hyperplan choisi est celui qui maximise la marge de séparabilité entre les deux ensembles de données (Figure 2.10).

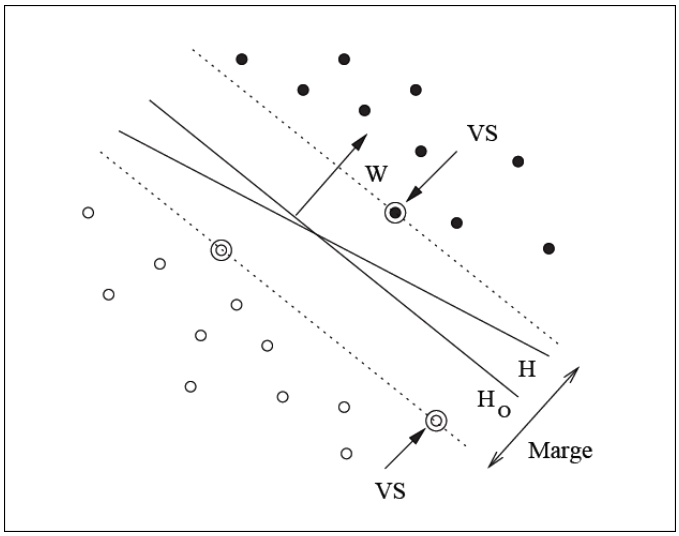


Figure 2.10 : Hyperplans séparateurs : H est un hyperplan quelconque, H_O est l'hyperplan optimal, VS : sont les vecteurs de support.

La sélection de l'hyperplan dans un espace de caractéristiques nécessite d'évaluer un produit scalaire dans cet espace. Ce qui peut être très coûteux en temps et en complexité si l'espace est de très grande dimension. Heureusement, ce calcul n'est pas obligatoire grâce à une opération mathématique appelé noyau. Le noyau calcule le produit scalaire de deux points dans l'espace de dimension supérieure sans avoir à les projeter [32].

Pour la reconnaissance du locuteur (cas de la vérification par exemple), les SVM recherchent une surface de décision optimale, déterminée par certains points de l'ensemble d'apprentissage appelés vecteurs de support, en projetant les données d'entrée non-linéairement séparables dans un espace de plus grande dimension appelé espace de caractéristiques (Figure 2.11). Cette surface, qui est dans l'espace des caractéristiques, peut être considérée comme un hyperplan optimal de décision. Elle est obtenue par la résolution d'un problème de programmation quadratique dépendant des paramètres de régularisation.

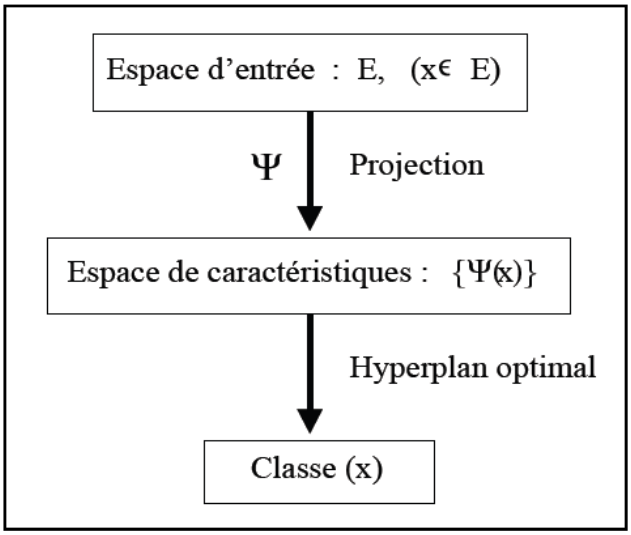


Figure 2.11 : Principe des SVM.

De par leur nature, les méthodes SVM sont binaires c'est-à-dire s'appliquent à deux classes. Cependant, les problèmes du monde réel sont dans la plupart des cas multi-classes, les méthodes SVM, réduisent le problème multi-classes à une composition de plusieurs hyperplans bi-classes permettant de tracer les frontières de décision entre les différentes classes [33]. Ces méthodes décomposent l'ensemble d'exemples en plusieurs sous-ensembles représentant chacun un problème de classification binaire.

2.4.e. Modèle de Markov caché (Hidden Markov Model HMM)

Le modèle HMM est une méthode paramétrique statistique très puissante en termes de caractérisation d'échantillons de données sous forme de séquences temporelles. En parole, cette technique a été utilisée avec succès pour la reconnaissance automatique de la parole, suivi du pitch et des formants, synthèse de la parole, la reconnaissance du locuteur dépendant du texte, etc. [16].

Un modèle de Markov caché est une chaine Markovienne dans laquelle l'état à l'instant t ne dépend que de l'état aux k instants précédents. Il est défini par une structure composée d'états, de transitions et par un ensemble de distributions de probabilités sur les transitions.

En HMM chaque état correspond à une observation qui est une fonction probabiliste. Le modèle produit peut être vu comme un double processus, l'un gère la transition d'un état à une autre, mais l'état du système n'est pas directement observable, l'autre processus gère l'obtention d'une séquence de paramètres que nous pouvons observer.

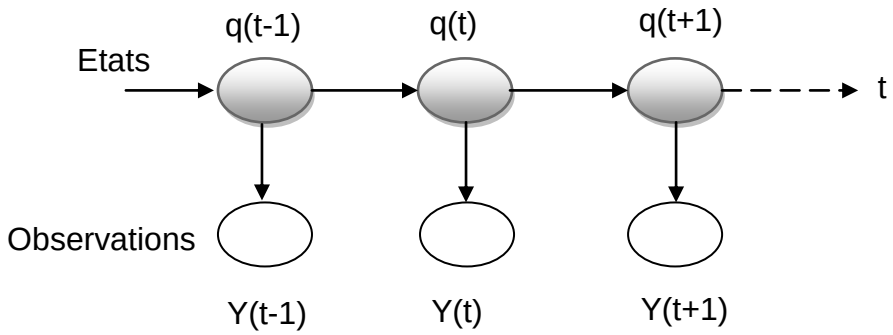


Figure 2.12 : Représentation d'un HMM à trois états.

La sortie d'un modèle de Markov caché est donc une variable aléatoire Y générée selon la fonction de probabilité de sortie de chaque état. La figure 2.12 résume la modélisation HMM à travers un exemple à 3 états.

Pour construire le modèle de la séquence, qui va servir ensuite en reconnaissance comme référence de comparaison avec des séquences inconnues, une méthode statistique d'apprentissage est utilisée. L'apprentissage vise à déterminer les paramètres du modèle de Markov (matrice de probabilités de transitions, matrice de probabilités de sortie ou d'émission, et la matrice de probabilités initiales) permettent de maximiser la probabilité à posteriori d'émettre les observations ayant le modèle recherché. C'est le principe de maximum de vraisemblance. L'algorithme utilisé pour l'apprentissage est l'algorithme de Baum-Welch [34].

En reconnaissance, le maximum de vraisemblance d'une séquence de paramètres test est calculée par rapport au modèle HMM. En mode dépendant du texte, toutes les phrases et tous les phonèmes sont modélisés en utilisant une structure multi-états. Cette structure se réduit à un seul état en mode indépendant

du texte. Les modèles sont, dans ce cas, appelés mélanges de gaussiennes (Gaussian Mixture Models GMM). Dans ce travail, ces modèles sont adoptés comme modèles de base de toutes les expériences effectuées. Ils seront, par conséquent, présentés avec plus de détails dans le paragraphe suivant.

2.4.f. Modèle de mélanges de Gaussiennes (Gaussian Mixture Models GMM)

Au cours des dernières années, les modèles GMM sont devenus l'approche de modélisation la plus dominante pour les systèmes de reconnaissance du locuteur en mode indépendant du texte [35]. Reynolds et al. [36] ont présenté le système de vérification du locuteur basé sur les modèles GMM qui utilise le modèle du monde (Universal Background Model UBM) comme une représentation du locuteur alternatif. Une adaptation Bayesienne est effectuée afin d'obtenir les modèles des locuteurs à partir du modèle UBM.

La figure 2.13 illustre la technique de modélisation GMM-UBM.

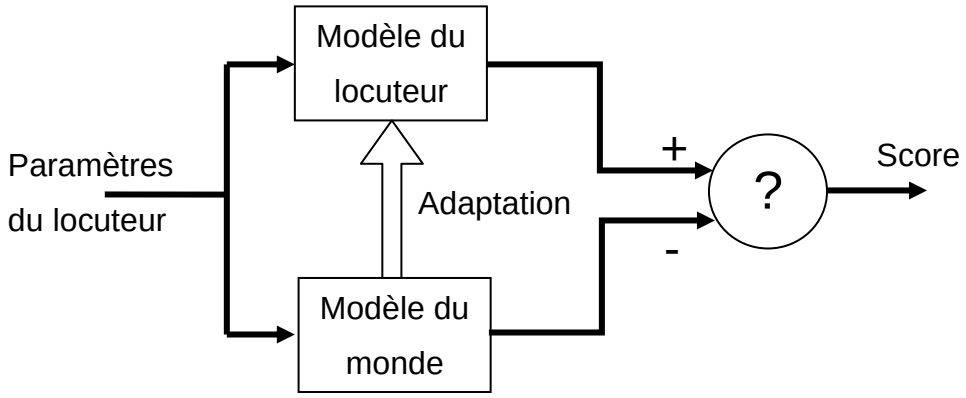


Figure 2.13 : Système GMM-UBM.

La fonction de vraisemblance d'un vecteur de paramètres **x** de dimension D est définie par la somme pondérée de M composantes de distributions gaussiennes, comme donné par l'équation suivante :

$$P(\boldsymbol{x}/\lambda) = \sum_{i=1}^{M} \omega_i \, p_i(\boldsymbol{x}) \tag{2.10}$$

Le mélange des poids satisfait la contrainte :

$$\sum_{i=1}^{M} \omega_i = 1. \tag{2.11}$$

Le mélange des gaussiennes λ représente le modèle du locuteur, il est caractérisé par trois paramètres : les poids ω_i, les vecteurs de moyennes μ_i, et les matrices de covariances Σ_i (en général pleine, mais dans la plupart des cas pratiques juste diagonale). Un exemple de mélange de trois gaussiennes est illustré dans la figure 2.14.

La densité correspondant a la i^{ème} composante gaussienne est donnée par :

$$p_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\mu_i)^T \Sigma_i^{-1} (\boldsymbol{x}-\mu_i)} \tag{2.12}$$

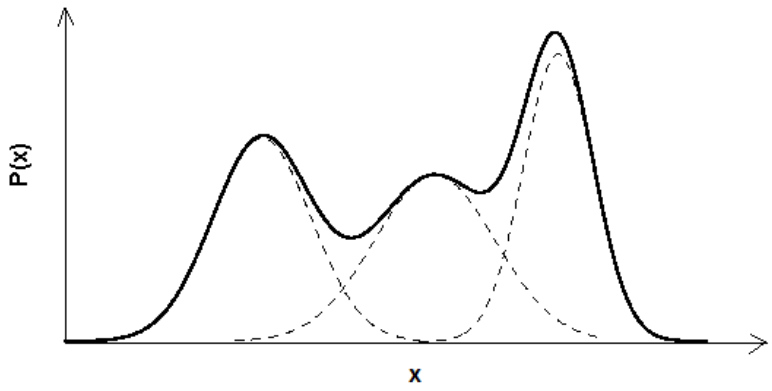


Figure 2.14 : Exemple de mélange de trois gaussiennes.

Soit $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_N}\}$ une séquence de N vecteurs de paramètres représentant un signal de parole. En supposant que les vecteurs de paramètres de $\boldsymbol{X}$ sont mutuellement indépendants, la vraisemblance logarithmique (appelé aussi score de vraisemblance logarithmique) se calcule par :

$$P(\boldsymbol{X}/\lambda) = \frac{1}{N}\sum_{j=1}^{N} P(\boldsymbol{x_j}/\lambda) \tag{2.13}$$

Souvent, la moyenne du score de vraisemblance logarithmique est utilisée (en divisant par N). Puisqu'il est moyenné sur toues les trames des paramètres, ce score est indépendant de la durée du vecteur test.

Le système GMM-UBM est construit selon les étapes suivantes :

1. un seul modèle indépendant des locuteurs, appelé modèle du monde, est utilisé pour représenter $P(\boldsymbol{X}/\lambda_{UBM})$. L'apprentissage du modèle UBM se fait en utilisant un ensemble large de données de parole obtenu par la concaténation d'une large population de locuteurs. C'est un large modèle GMM appris pour représenter la distribution de paramètres indépendants de locuteurs. Plus précisément, on désire sélectionner des échantillons de parole qui reflètent la parole provenant d'une partie alternative nécessaire à la reconnaissance. Ceci s'applique au type et à la qualité de parole, ainsi qu'à la composition des locuteurs.

   Un algorithme itératif appelé espérance-maximisation (Expectation Maximisation EM) est utilisé pour estimer le maximum de vraisemblance du modèle pour les vecteurs de paramètres d'apprentissage [37].

   L'estimation des paramètres du modèle de mélange des gaussiennes $\lambda_{UBM} = \{\omega_{UBM}, \mu_{UBM}, \Sigma_{UBM}\}$ est améliorée d'une itération a l'autre jusqu'a atteindre la convergence. L'apprentissage peut être arrêté lorsque le changement de vraisemblance entre deux itérations est au dessous d'un certain seuil ou après avoir fait un nombre prédéterminé d'itérations.

2. La dérivation des modèles des locuteurs se fait, dans le système GMM-UBM de façon adaptative. Les signaux de parole d'apprentissage de chaque locuteur serrent a adapter les paramètres du modèle du monde (UBM) en utilisant l'algorithme d'estimation du maximum a posteriori (MAP) [38]. Cet algorithme adapte itérativement le modèle UBM pour arriver au modèle du locuteur comme illustré dans la figure 2.15.

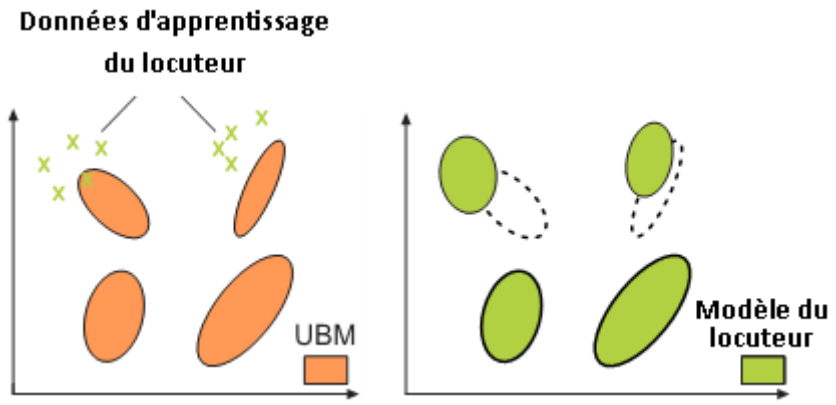


Figure 2.15 : Obtention du modèle du locuteur par la méthode d'adaptation MAP.

Considérons la séquence de T vecteurs de paramètres $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_T}\}$. Pour chaque mélange i du modèle du monde, $Pr(i/\boldsymbol{x}_t)$ est calculée par :

$$Pr(i/\boldsymbol{x}_t) = \frac{\boldsymbol{\omega}_i \, p_i(\boldsymbol{x}_t)}{\sum_{j=1}^{M} \boldsymbol{\omega}_j \, p_j(\boldsymbol{x}_t)} \tag{2.14}$$

$Pr(i/\boldsymbol{x}_t)$ et $\boldsymbol{x}_t$ seront utilisées, ensuite, pour calculer les statistiques suffisantes pour les paramètres de poids, de moyenne, et de variance. Telles que :

$$\eta_i = \sum_{t=1}^{T} Pr(i/\boldsymbol{x}_t) \tag{2.15}$$

$$E_i(\boldsymbol{x}) = \frac{1}{\eta_i}\sum_{t=1}^{T} Pr(i/\boldsymbol{x}_t).\boldsymbol{x}_t \tag{2.16}$$

$$E_i(\boldsymbol{x^2}) = \frac{1}{\eta_i}\sum_{t=1}^{T} Pr(i/\boldsymbol{x}_t).\boldsymbol{x}_t{}^2 \tag{2.17}$$

Finalement, ces nouvelles statistiques suffisantes, qui correspondent aux données d'apprentissage, seront utilisées pour réactualiser les statistiques suffisantes UBM précédentes du mélange i pour créer les paramètres adaptés avec :

$$\widehat{\omega}_i = [\alpha_i \eta_i / T + (1-\alpha_i)\omega_i]\gamma \tag{2.18}$$

$$\hat{\mu}_i = \alpha_i E_i[\boldsymbol{x}] + (1-\alpha_i)\mu_i \tag{2.19}$$

$$\hat{\sigma}_i{}^2 = \alpha_i E_i[\boldsymbol{x^2}] + (1-\alpha_i)(\sigma_i{}^2 + \mu_i{}^2) - \hat{\mu}_i{}^2 \tag{2.20}$$

$\widehat{\omega}_i$, $\hat{\mu}_i$, et $\widehat{\sigma}_i{}^2$ sont les nouveaux poids, moyennes, et variances. Le facteur γ est calculé après que tous les poids aient été adaptés pour assurer que leur somme soit égale à 1.

α_i est un coefficient d'adaptation contrôlant la balance entre les paramètres estimés des précédent et nouveau modèles. Ce coefficient est défini par :

$$\alpha_i = \frac{\eta_i}{\eta_i + r}, \tag{2.21}$$

où r est un facteur de régulation indiquant la confiance accordée aux statistiques provenant des données d'apprentissage par rapport aux statistiques provenant des modèles du monde.

En pratique, seules les moyennes des distributions gaussiennes sont adaptées pour la reconnaissance des locuteurs.

2.5. Prise de décision

Durant la tâche de reconnaissance, chaque locuteur L_i est représenté par un modèle GMM λ_i. En identification, l'objective est de trouver le modèle qui possède la probabilité a posteriori maximale ayant la séquence de parole test X.

Ceci peut se formuler comme :

$$\hat{L} = arg \max_{1 \le i \le L} \log P(X/\lambda_i) \tag{2.22}$$

pour L locuteurs.

Comme développé en paragraphe 2.2.2.f, l'indépendance entre les observations donne :

$$\hat{L} = arg \max_{1 \le i \le L} \sum_{t=1}^{T} \log P(x_t/\lambda). \tag{2.23}$$

En vérification, l'objective est de déterminer si la séquence de test X a été prononcée par le locuteur L. Pour cela, nous avons besoin de deux hypothèses ;

- $H0$: X a été parlée par L

  et

- $H1$: X n'a pas été parlée par L.

Le rapport de vraisemblance (Likelihood ratio LR) entre les deux hypothèses H_0 et H_1 s'écrit par :

$$LR = \frac{P(H_0/L)}{P(H_1/L)}. \tag{2.24}$$

D'après le théorème de Bayes, on a :

$$P(H_0/L) = \frac{P(L/H_0)P(H_0)}{P(X)}. \tag{2.25}$$

En remplaçant $P(H_0/L)$ par sa valeur dans l'équation 2.24 nous obtenons :

$$LR = \frac{P(L/H_0)P(H_0)}{P(L/H_1)P(H_1)}. \tag{2.26}$$

Cette expression est utilisée en pratique car il est plus facile de calculer $P(L/H_0)$ que de calculer $P(H_0/L)$.

Dans la technique GMM-UBM, L'évaluation du rapport LR est réalisée en calculant la différence du logarithme de vraisemblance $\log P(X/\lambda_i) - \log P(X/\lambda_{UBM})$. Si cette différence est supérieure à un seuil donné, le locuteur est accepté, sinon il est rejeté.

Ceci peut se formuler comme :

$$\begin{cases} \log P(X/\lambda_i) - \log P(X/\lambda_{UBM}) \geq \theta, & Acceptation\ du\ locuteur \\ \log P(X/\lambda_i) - \log P(X/\lambda_{UBM}) < \theta, & rejet\ du\ locuteur \end{cases} \quad (2.27)$$

Notons, enfin, que le rapport des probabilités a priori $\frac{\mathrm{P(H_0)}}{\mathrm{P(H_1)}}$ est incorporé dans le seuil.

2.6. Evaluation

Les systèmes de reconnaissance des locuteurs utilisaient dans les années antérieures les courbes ROC (Receiver Operating Characteristic) empruntées de la théorie de détection de signaux. Les tâches effectuées dans la détection peuvent être vues comme un compromis de deux types d'erreurs : la fausse rejection, dans laquelle le système rejette des vrais locuteurs (authentiques), et la fausse acceptation (alarme), pour laquelle le système accepte des faux locuteurs (imposteurs). Martin et al. [39] montre l'intérêt d'utiliser la courbe DET (Detection Error Tradeoff) dans plusieurs applications comme la reconnaissance du locuteur ou du langage. Dans la courbe DET, on représente les taux d'erreurs de fausse alarme et de fausse rejection dans les axes horizontales (X) et verticales (Y), respectivement, avec une échelle logarithmique.

Une des mesures de performance les plus utilisées dans la reconnaissance du locuteur est le point vérifiant :

$$P_{FA} = P_{FR}\ , \quad (2.28)$$

où P_{FA} représente la probabilité de fausse alarme et P_{FR} la probabilité de fausse rejection.

Ce point est appelé taux d'erreur égale (Equal Error Rate EER).

Un autre point intéressant pour l'évaluation des systèmes de reconnaissance des locuteurs est appelé minimum de la fonction coût de détection (minimum detection cost fonction minDCF). Il est obtenu en minimisant l'équation suivante :

$$C_{DET} = C_{FR} \times P_{FR \mid Cible} \times P_{Cible} + C_{FA} \times P_{FA \mid Non\ Cible} \times (1 - P_{Cible}). \quad (2.29)$$

Les paramètres de cette fonction coûts sont les coûts relatives aux erreurs de détection C_{FR} et C_{FA}, et les probabilités a priori d'un locuteur cible spécifique P_{Cible}.

2.7. Conclusion

Dans ce chapitre nous avons décrit le système de reconnaissance du locuteur. On a vu que comme tout système de reconnaissance de forme, ce système peut être décomposée en deux phases, la phase d'apprentissage et la phase de reconnaissance (ou de test). On a commencé par décrire les étapes communes aux deux phases, qui sont :

Le prétraitement, qui regroupe la segmentation, le fenêtrage du signal parole et les méthodes de normalisation. Le but de cette étape était de rendre les données originales sous formes de trames de durées d'environs 20 ms, et de les transformer afin de les rendre plus convenables aux traitements suivants.

Dans l'étape d'extraction des paramètres, nous avons vu les méthodes les plus utilisées dans ce domaine, à savoir, la méthode de prédiction linéaire LPC, la méthode MFCC, et la méthode PLP.

Dans la phase d'apprentissage, l'étape de modélisation joue un rôle primordial dans le système de reconnaissance car c'est en se basant sur ces modèles qu'on jugera sur l'appartenance ou pas de la parole a un certain locuteur. Plusieurs techniques de modélisation ont été décrites comme la technique de quantification vectorielle, l'alignement temporel dynamique, et les réseaux de neurones. Les méthodes statistiques de mélange des Gaussiennes ont été étudiées avec plus de détails car ils sont les plus utilisées en reconnaissance des locuteurs.

La prise de décision dépend de la tache traitée. C'est ainsi qu'on a vu le fonctionnement de cette partie en vérification et en identification.

Enfin, nous avons terminé par décrire la façon avec laquelle le système de reconnaissance du locuteur est évalué. Pour cela on a défini le taux d'erreur égale (EER) et le point minDCF.

CHAPITRE 3

ETUDE DES PARAMETRES INVARIANTS POUR LA COMPENSATION DE LA VARIABILITE EN RECONNAISSANCE DU LOCUTEUR

3.1. Introduction

La parole est le mode de communication le plus naturel pour les êtres humains. Par conséquent, il a toutes les potentialités d'avoir la place préférée dans l'interaction homme - machine. Or, la communication parlée consiste à produire la parole, à la transmettre, et enfin à l'entendre. Chacune de ces étapes implique, forcement, des distorsions acoustiques pouvant affecter le bon déroulement d'un système de reconnaissance du locuteur.

Pour le cas des êtres humains, chaque individu normalement développé montre des capacités extrêmement robustes pour reconnaitre les personnes. Même les jeunes enfants peuvent reconnaitre leurs membres de famille dans des situations sévères comme lorsqu'ils se trouvent au milieu d'une foule ou dans un environnement bruité. En reconnaissance de la parole, nous pouvons comprendre les mots de quelqu'un qui appelle par téléphone portable, même si c'est son premier appelle.

Notre perception n'est pas seulement robuste à la parole, elle s'étale aussi à d'autre medias de communications. Une image visuelle qui change de forme d'un certain point de vue, est perçue de la même manière par les personnes. Comme pour la couleur, un oiseau dans la lumière du jour ne se présente pas de la même façon qu'en couché de soleil, mais l'être humain arrive à percevoir proprement l'équivalence entre les deux situations. Dans les systèmes de reconnaissance de

locuteur, le phénomène de variabilités est une cause majeure de la dégradation de performance.

3.2. Compensation de la variabilité

Il est clair que la réduction de la variabilité a pour conséquence de réduire les variances des modèles. Ce qui en résulte à une augmentation des performances du système de reconnaissance du locuteur.

L'idée de base des paramètres invariants est de décrire les objets par un ensemble mesurable de quantités appelées paramètres invariants. Ces derniers sont insensibles à des déformations particulières et qui assurent un pouvoir de discrimination suffisant pour faire la distinction entre les objets appartenant à différentes classes.

Les conditions non contrôlées où le types d'enregistrement, par exemple, d'un locuteur en apprentissage n'est pas le même que celui en phase de test, a pour effet d'élargir la variance des paramètres caractérisant le locuteur. Par Conséquent, le groupe contenant l'ensemble des observations d'un locuteur peut contenir des observations de paramètres issues d'un autre ensemble adjacent appartenant à un autre locuteur. Ce qui augmente les taux d'erreur de reconnaissance.

La figure suivante (3.1) illustre le phénomène de variabilité et son influence sur les paramètres caractéristiques du locuteur (l'illustration est faite par une représentation simple en deux dimensions). Deux cas sont présentés, le cas d'une petite variabilité et le cas d'une large variabilité.

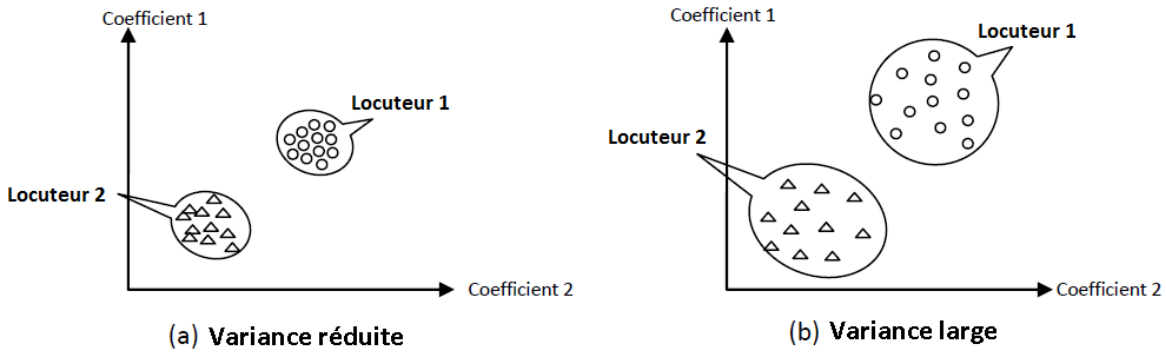


Figure 3.1: Illustration de l'effet de la variabilité sur les paramètres.

Pour compenser la variabilité, diverses méthodes ont été développées en agissant sur le module d'extraction des paramètres, la partie de modélisation, ou celle de la prise de décision.

3.2.1. Compensation de la variabilité dans le domaine des paramètres

Les méthodes de paramétrisation des signaux de parole sont appliquées au premier élément d'un système de reconnaissance du locuteur. Les prétraitements effectués comme la détection de la parole ou la suppression des bruits font aussi partie de cet élément. Leur rôle est d'améliorer la qualité de la parole et par voie de conséquence, les performances du système de reconnaissance. Diverses méthodes de compensation des variabilités existent à ce niveau.

- **Normalisation de la moyenne et la variance cepstrales (Cepstral Mean and Variance Normalization CMVN) :**

L'analyse cepstrale est un type de transformation homomorphique qui permet de transformer les produits de convolution en simples opérations additives. Une transformation homomorphique est une projection d'un signal sur un domaine où son comportement est linéaire.

La méthode CMN appelée aussi soustraction de la moyenne cepstrale (Cepstral Mean Substraction CMS) est l'une des plus anciennes, simples, et plus réussies méthodes dans le domaine paramétrique. Elle est, donc très utilisée en reconnaissance de locuteur pour traiter le problème de variabilité.

Le développement fait dans le paragraphe 2.2.4 nous a permet de calculer des coefficients particulièrement intéressants, les cepstres, puisqu'ils donnaient une somme de plusieurs termes au lieu de produits de convolution.

Considérons l'exemple d'un signal de parole issu d'un canal de transmission, comme montré dans la figure suivante (figure 3.2).

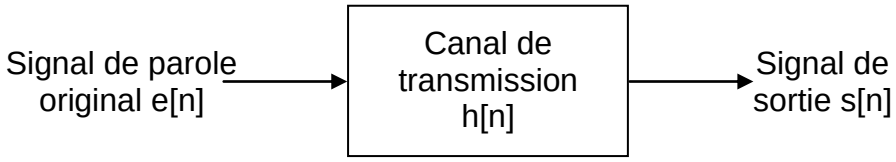


Figure 3.2: Système de communication de la parole.

La sortie peut s'écrire, dans ce cas, dans le domaine temporelle comme un produit de convolution par :

$$s[n] = e[n] * h[n]. \tag{3.1}$$

En passant au domaine fréquentiel, nous aurons :

$$S_{t,k} = E_{t,k}.H_{t,k}\ , \tag{3.2}$$

où t représente l'indice de la trame, et k l'indice de la fréquence. En forme cepstrale, cette équation (3.2) devient une somme.

La soustraction de la moyenne de chaque trame permet d'éliminer le terme correspondant au canal, en supposant que ce dernier ne varie pas durant le temps d'enregistrement du signal de parole.

Notons aussi, qu'en appliquant la CMN, la moyenne du signal d'entrée (signal parole d'une durée d'une trame) est aussi enlevée. Dans ce cas cette moyenne correspond à l'information moyenne du spectre, Büyük montre dans [40] que l'élimination de ce terme améliore les résultats de reconnaissance puisqu'il varie en fonction de l'effort de parole ou pour des raisons de santé.

La technique CVN transforme les paramètres pour permettre de compenser les distorsions non linéaires. Elle est souvent utilisée en conjonction avec la technique CMN.

Formellement, la technique CMN divise chaque vecteur de paramètre, correspondant à une trame, par l'écart type des vecteurs de paramètres de

l'enregistrement total du signal de parole. Le vecteur de paramètres résultant aura une variance égale à l'unité.

Cette normalisation se réalise selon la règle suivante :

$$c_{MV}(t) = \frac{c(t) - \mu_c}{\sigma_c}, \tag{3.3}$$

où c_{MV} est le cepstre résultant de la normalisation conjointe (en moyenne et en variance) du cepstre original c correspondant a la trame t. μ_c et σ_c sont la moyenne et l'écart type, respectivement.

- **Normalisation par Feature Mapping (FM) :**

  Introduite par Reynolds [41], la technique FM est parmi les premières méthodes utilisées pour résoudre le problème de variabilité du canal entre l'apprentissage et le test. Son principe repose sur la transformation des paramètres d'un l'espace dépendant du contexte à un espace neutre de tout contexte. Cette transformation non linéaire est définie par les différences entre l'ensemble des GMM représentant les conditions d'enregistrement d'intérêt, comme les différents types d'appareils téléphoniques, et le modèle GMM neutre. La normalisation projette les trames dépendants du canal vers un espace indépendant du canal en se basant sur l'indice de la gaussienne la plus vraisemblable dans le mode indépendant du canal.

  Pour une trame x(t), cette transformation s'effectue de la façon suivante :

$$x(t) = \frac{\sigma_{Gind}}{\sigma_{Gd}}(x(t) - \mu_{Gd}) + \mu_{Gind}, \tag{3.4}$$

où μ_{Gd}, μ_{Gind}, σ_{Gd}, σ_{Gind} sont, respectivement, les moyennes et les écarts types des GMM correspondants aux modèles dépendants et indépendants du canal.

- **Gaussianisation (Feature warping) :**

Cette technique est apparue en 2001 dans [42] pour construire une représentation plus robuste de chaque distribution de cepstre. Son élaboration consiste à modifier la répartition des coefficients cepstraux sur une fenêtre glissante en la faisant correspondre à une gaussienne de moyenne nulle et de variance unité.

La figure 3.3 montre la procédure de transformation de chaque paramètre afin d'obtenir une configuration plus robuste au phénomène de variabilité.

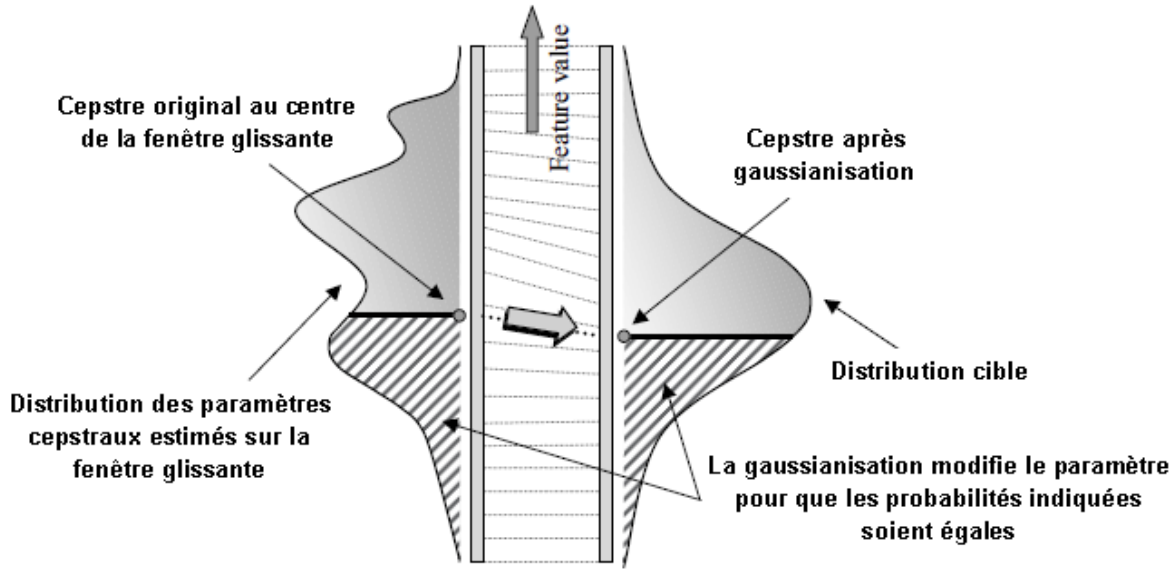


Figure 3.3 : Gaussianisation des paramètres selon la forme de la distribution cible [40].

Ce travail est justifié par le fait que les distorsions dues au canal n'affectent pas seulement la moyenne et la variance estimées sur le signal entier mais aussi la distribution marginale de chacun des coefficients.

La largeur de la fenêtre glissante, utilisée pour l'estimation de la distribution des paramètres cepstraux, est de 300 - 500 périodes trames, donc, environ 3 - 5 secondes.

3.2.2. Compensation de la variabilité dans le domaine de modélisation

Dans cette classe d'approches, on ne cherche pas à modifier les paramètres de test mais à adapter les modèles acoustiques. L'objective dans ce cas est de réduire les variabilités du canal en améliorant les modèles des cohortes et des locuteurs. Des méthodes comme l'analyse de facteur et la variabilité totale basées sur les GMM, ou la compensation NAP (Nuisance Attribute Projection) basée sur les SVM, sont souvent utilisées

- **Compensation NAP :**

  Cette technique a été originalement proposée dans [43]. Dans cette méthode, l'espace de paramètres est transformé en utilisant une projection orthogonale dans l'espace complémentaire de celui du canal, et qui ne dépend que du locuteur. La projection est effectuée en calculant une matrice de covariance interclasse.

  Pour l'application de la méthode NAP sur les paramètres acoustiques, on définit une matrice de projection d x d de rang k < d pour la j^{ème} gaussienne comme :

$$P_g = I - W_{[k]g} {W_{[k]g}}^T, \tag{3.5}$$

  où $W_{[k]g}$ est une matrice rectangulaire de faible rang dont les colonnes sont les k principaux vecteurs propres de la matrice S a partir de plusieurs enregistrements (sessions) par locuteur.

  Les matrices de covariances du modèle UBM transformées s'obtiennent par :

$$\Sigma_{tg} = P_g \Sigma_g P_g^T. \tag{3.6}$$

  La méthode NAP vise, donc, à éliminer des directions nuisibles de l'espace des paramètres dans chaque mélange.

- **Analyse jointe de facteur (Joint Factor Analysis JFA) :**

Le formalisme de la JFA constitue, ces dernières années, l'état de l'art dans la reconnaissance du locuteur [44,45]. Elle a été proposée pour modéliser les variabilités de la session et du locuteur dans l'espace des paramètres d'un modèle GMM. Ces variabilités sont déterminées par des sous espaces de l'ensemble des paramètres, appelées souvent hyper paramètres.

Dans le modèle JFA, chaque locuteur est représenté par les moyennes, la covariance, et les poids d'un mélange de C densités gaussiennes définies dans un espace de paramètres de dimensions F. Le modèle GMM d'un locuteur cible est obtenu en adaptant les moyennes du modèle du monde UBM. La supposition de base dans la méthode JFA est que le super-vecteur, composée des moyennes M dépendants de locuteur et de canal, peut être décomposé en une somme de deux super-vecteurs : le super-vecteur locuteur l et le super-vecteur canal c.

$$M = s + c, \tag{3.7}$$

où s et c suivent des distributions normales.

Le terme s de l'équation (3.7) est modéliser tel que :

$$s = m + Vy + Dz\ , \tag{3.8}$$

où m est le super-vecteur indépendant du locuteur et du canal obtenu a partir des moyennes UBM, D est une matrice diagonale, V est une matrice rectangulaire de faible rang et y et z sont des vecteurs aléatoires indépendants de distributions normales. y et z sont appelés les facteurs locuteur et commun, respectivement.

Le super-vecteur dépendant du canal c, qui représente l'effet du canal dans un enregistrement, est obtenu par :

$$c = Ux\ , \tag{3.9}$$

où U est une matrice rectangulaire de faible rang (connue par la matrice des canaux propres), x est un vecteur qui suit une distribution normale standard, ces composants sont appelés facteurs canal.

La tâche de l'analyse jointe de facteur est, par conséquent, d'apprendre les hyper paramètres U, V, et D a partir d'un large ensemble de données. On commence par estimer la matrice V en supposant que les deux autres matrices sont nulles. Étant donnée la matrice V, on estime ensuite la matrice U tout en continuant à considérer la matrice D nulle. Et on finit par estimer la matrice D, connaissant les deux autres matrices. Nous renvoyons le lecteur aux développements faits dans [46,47], pour une description détaillée de la procédure d'estimation des hyperparametres du formalisme JFA.

Le calcul de la vraisemblance de l'enregistrement test se fait par l'intégration sur les distributions a posteriori de y et z et la distribution a priori x.

- **Variabilité totale :**

La technique de modélisation JFA basée sur les facteurs locuteur et canal consiste à définir deux espaces : l'espace locuteur définit par les voix propres (eigenvoices) a travers la matrice V et l'espace canal définit par les canaux propres (eigenchannels) à travers la matrice U. L'approche de variabilité totale se définit sur un espace au lieu de deux. Cet espace contient, simultanément, le locuteur et le canal. Il est définit par la matrice de variabilité totale dont les vecteurs propres corresponds aux plus grandes valeurs propres de la matrice de covariance de la variabilité totale.

Pour un enregistrement donné, le super-vecteur GMM dépendant du locuteur et du canal peut s'écrire comme :

$$M = m + Tw\,, \tag{3.10}$$

où m est le super-vecteur indépendant du locuteur et du canal obtenu a partir des moyennes UBM, T est une matrice rectangulaire de faible rang, et w est une

variable aléatoire suivant une distribution normale N(0,1). Les composants du vecteur w sont les facteurs totaux. Ces vecteurs sont appelés vecteurs identité ou i-vecteurs.

En modélisation, M est supposée normalement distribuée de moyenne m et de matrice de covariance T. L'apprentissage de la matrice de variabilité totale T se fait de la même manière que la matrice V en JFA. Cependant, une importante différence entre les deux, en termes d'apprentissage, est qu'en voix propres en JFA, tous les enregistrements d'un locuteur donné proviennent du même locuteur. Dans le cas de la variabilité totale, l'ensemble des expressions parlées sont considérées produites par les locuteurs différents [48].

3.2.3. Compensation de la variabilité dans le domaine de décision

Dans un système GMM-UBM, la décision de rejeter ou d'accepter l'identité proclamée d'un locuteur se base sur des scores de vraisemblances du modèle locuteur et le modèle du monde. La variabilité interlocuteur peut causer la non coïncidence des deux distributions des scores imposteur et authentique d'un locuteur à un autre. Ce qui augmente les erreurs de reconnaissance. Des techniques de compensation au niveau des scores ont été proposées.

- **Normalisation zero (Z-norm):**

La technique Z-norm est une méthode populaire de normalisation [36], elle consiste à tester le modèle du locuteur proclamé contre les séquences d'un certain nombre d'imposteurs. Les moyennes et les écarts types des scores résultants seront ensuite calculés. La normalisation z s'exprime en fonction de ces paramètres calculés par :

$$\tilde{S}_{core}(x) = \frac{S_{core}(x) - \mu_z}{\sigma_z}\,, \tag{3.11}$$

Comme cette normalisation ne nécessite que les modèles des locuteurs, alors il est possible de procéder à l'estimation de ses paramètres durant l'étape d'apprentissage.

- **Normalisation test (T-norm):**

La technique T-norm se fait de la même manière que la Z-norm, à la différence que cette fois, c'est les séquences de test qui seront utilisés dans les modèles de locuteurs imposteurs pour estimer les paramètres moyennes et écarts type. Le choix des imposteurs, dans ce cas, doit se faire pour se rapprocher au maximum des locuteurs authentiques.

Afin de compenser les variabilités de l'apprentissage et de test, les deux techniques Z-norm et T-norm peuvent utiliser ensemble l'une après l'autre, ZT-norm ou TZ-norm.

3.3. Paramètres invariants proposés

Les nouveaux paramètres que nous proposons dans cette thèse visent à la compensation de la variabilité intersession dans la reconnaissance du locuteur. Cette variabilité mène à une différence entre les énoncés d'apprentissage et celles des tests. Ces paramètres se basent sur la théorie d'invariance qui nécessite l'introduction de quelques termes de base et des définitions.

3.3.1. Définitions

L'espace S de signaux est un sous-ensemble d'un espace V de vecteurs complexes de dimensions finies.

Les éléments de S sont appelés "formes", ils sont notées par des vecteurs (P.ex. **v**, **w**).

G est un groupe agissant sur V par un operateur linéaire g. L'action de G introduit une relation d'équivalence $\sim$ dans S.

Deux formes $\boldsymbol{v}, \boldsymbol{w}$ sont dites équivalentes, **v**~ **w**, alors :

$$g\boldsymbol{v} = g\boldsymbol{w}. \tag{3.12}$$

Dans notre problème de reconnaissance de locuteur, l'équation (3.12) implique que le changement du spectre causé par la variabilité pour un même locuteur, doit mener à des vecteurs de paramètres pareils. Nous appelons le sous-ensemble $O(\boldsymbol{v}) = \{gv \mid g \in G\}$, pour un vecteur donné $\boldsymbol{v}$, une orbite de G dans S. L'orbite contient toutes les formes possibles d'une classe équivalente. Deux orbites O_1 et O_2 sont soit identique, soit n'ont aucun point en commun.

La classification complète de cet ensemble d'orbites aboutit à une discrimination exacte entre les différentes classes. D'un autre cote, une classification incomplète peut aboutir aux mêmes paramètres de formes provenant de différentes classes. En pratique, on cherche à atteindre un "haut degré de complétude".

Une solution pratique pour ce problème est la construction d'un espace de paramètres complet F [49]. Cet espace est un sous-ensemble d'un certain espace vectoriel complexe avec l'application $T : S \rightarrow F$ ayant les propriétés suivantes :

Propriété 1 : $T(\boldsymbol{v}) = T(g\boldsymbol{v})$, pour tout $g \in G, \boldsymbol{v} \in S$.

Propriété 2 : $T(\boldsymbol{v}) = T(\boldsymbol{w})$, alors il existe $g \in G$ avec $\boldsymbol{v} = g\boldsymbol{v}$.

La condition de la propriété 1 garantie que les formes équivalentes transformées par l'application T en un seul point, alors que la propriété 2 assure les formes non équivalentes serons transformées par l'application T en des points distincts dans l'espace des paramètres.

3.3.2. Construction des espaces de paramètres pour la reconnaissance du locuteur

Soit $\boldsymbol{T}(v) = (f_1(v), f_2(v), \dots, f_n(v)), n \in \mathbb{N}$, les composantes de l'application f, où $f_i : S \rightarrow \mathbb{C}$.

Ces composantes f_i sont considérées polynomiales pour être implémentées sur machine.

Diverses méthodes existent pour la construction de tels paramètres invariants. Les approches de normalisation comme les moments font partie de ces méthodes. Dans cette approche, les formes sont transformées par rapport aux points extrêmes de l'orbite.

Dans les approches de différentiation, les paramètres invariants s'obtiennent en résolvant les équations différentielles partielles telles que les paramètres restent invariants pour des variations infinitésimales des actions du groupe G. Enfin, la dernière catégorie regroupe les approches de moyenne.

Dans cette catégorie les paramètres sont constants dans une orbite. Par conséquent, les propriétés décrites sont communes pour toutes les formes équivalentes. Cela suggère la construction des paramètres comme des moyennes appropriées.

Nous essayons de calculer la fonction de moyenne $A_f(v)$ en intégrant f sur l'orbite $O(v)$ par :

$$A_f(\boldsymbol{v}) = \int_{O(v)} f(g\boldsymbol{v})dg. \tag{3.13}$$

Nous appelons A_f la moyenne du groupe de f.

Cette approche a été appliquée en parole et en traitement d'image [1,4]. Ceci nous a motivé de l'examiner dans le contexte de la reconnaissance du locuteur.

Selon le théorème de Noether [49], on peut construire les bases polynomiales f de valeurs complexes définies sur l'espace S et dont les valeurs sont invariants sous l'action de G sur S, on les note par $\mathrm{f} \in \mathbb{C}[\mathrm{S}]^{\mathrm{G}}$.

La construction de ces valeurs se fait par le calcul de la moyenne de tous les monômes tel que :

$$v_0^{b_0}\ v_1^{b_1}\ \dots\ v_{n-1}^{b_{n-1}}\ , \tag{3.14}$$

avec

$$b_0\ +\ b_1 +\ \dots +\ b_{n-1} \leq\ |G|, \tag{3.15}$$

où $|G|$ est l'ordre du groupe G.

Supposons que l'espace de signaux S ne contient qu'un nombre fini de formes. Soit G, le groupe d'actions fini sur S. Alors, il existe un ensemble invariant $\mathrm{f} \in \mathbb{C}[\mathrm{S}]^{\mathrm{G}}$ vérifiant les propriétés 1 et 2 données dans le paragraphe 3.3.1.

3.3.3. Technique d'extraction des paramètres proposée

Dans le cas de la reconnaissance du locuteur, le signal est représenté par un vecteur temps fréquence (TF) $\mathrm{v_k(n)}$, où $1 \leq \mathrm{n} \leq \mathrm{N}$ est l'indice de temps et $1 \leq \mathrm{k} \leq \mathrm{K}$ est l'indice de la fréquence.

Le vecteur $\mathrm{v} = (\mathrm{v_1}, \mathrm{v_2}, \dots, \mathrm{v_K})$, qui contient toutes les valeurs spectrales pour un temps spécifique, appelé trame en parole, représente la forme dans notre cas de l'espace invariant.

Pour fixer l'action du groupe dans nos paramètres invariants, nous commençons par voir la façon par laquelle la variabilité affecte les composants du vecteur v.

Pour cela, considérons, par exemple, deux signaux de parole provenant d'un même locuteur. Un est enregistré par un certain type de microphone. Le second, est enregistré par un autre type de microphone.

Le premier enregistrement sera considéré provenir d'un canal représenté par une fonction de transfert $\mathrm{H_1[k]}$, alors que le deuxième enregistrement sera considéré provenir d'un canal représenté par une fonction de transfert $\mathrm{H_2[k]}$.

Les signaux de parole résultant de chaque enregistrement peuvent être exprimés dans le domaine fréquentiel par :

$$Y_1[k] = H_1[k]S[k] + B[k], \tag{3.16}$$

et

$$Y_2[k] = H_2[k]S[k] + B[k], \tag{3.17}$$

où $Y, H, S,$ et B sont, respectivement, le signal de parole, la fonction de transfert, le signal de sortie, et le bruit additif, du système d'enregistrement de la parole.

En utilisant la représentation vectorielle introduite dans cette section, nous associons un vecteur $\boldsymbol{v}$ au signal Y_1 et un vecteur $\boldsymbol{w}$ au signal Y_2.

Puisque ces deux vecteurs représentent le même locuteur, un choix adéquat de l'action du groupe doit mener à la deuxième propriété des paramètres invariants, c.à.d.;

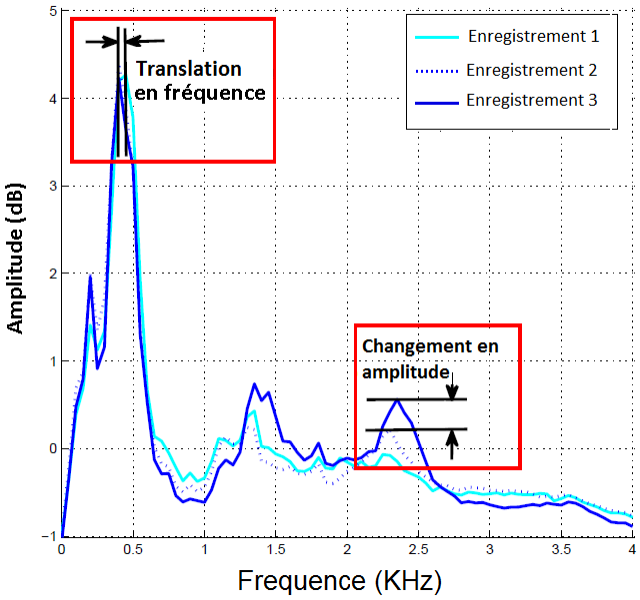
$T(\boldsymbol{v}) = T(\boldsymbol{w})$, alors il existe $g \in G$ avec $\boldsymbol{w} = g\boldsymbol{v}$.

Cette action dépend de la nature des effets des variabilités considérées.

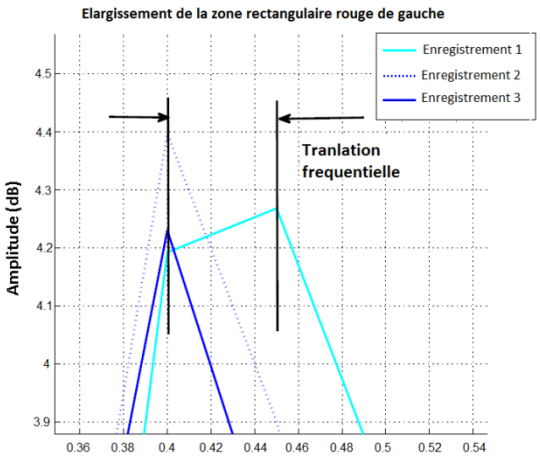
Dans notre travail, la base de données que nous allons utiliser pour la reconnaissance du locuteur nous permet d'étudier la variabilité du canal, puisque pour le même locuteur, des enregistrements sont effectuées sur des téléphones portables, d'autres (enregistrements) sont effectuées sur laptops (des détails sur ces enregistrements seront données dans le chapitre suivant).

Pour avoir une idée plus claire sur la différence entre plusieurs enregistrements correspondant à un même locuteur, nous représentons dans la figure 3.4 les spectres relatifs à un seul et même locuteur enregistré trois fois dans des conditions différentes.

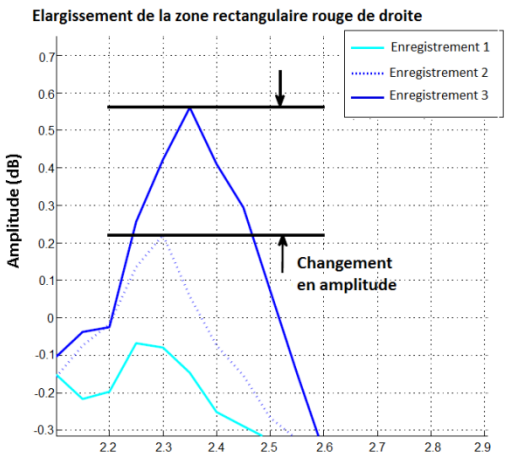
La figure 3.4 illustre les effets de la variabilité en termes de changement de la fréquence (Figure 3.4.b) ainsi que l'amplitude (Figure 3.4.c).



(a)



(b)



(c)

Figure 3.4 : Spectres d'amplitudes correspondant à trois enregistrements d'un seul locuteur. (a) Spectres complets. (b) Illustration des effets de changement de fréquence (c) Illustration des effets de changement d'amplitude.

Nous pouvons, voir a partir de ces figures, que même s'il y'a certaines similarités entre les spectres d'amplitude, des différences de fréquences et d'amplitudes sont observables. Ainsi, l'effet de changement de fréquence entre le premier enregistrement et le deuxième est clairement représenté dans la figure 3.4.b. Alors que le changement d'amplitude entre le deuxième enregistrement et le troisième est représenté dans la figure 3.4.c.

Par conséquent, ces effets peuvent être attribués à l'action du groupe G. Cette action correspond à l'opération de translation de la fréquence et l'opération de multiplication pour l'amplitude.

En termes d'expression, le spectre S d'un même locuteur dans deux différentes conditions d'enregistrement A et B peut s'écrire comme :

$$S_A(\omega) = C(w).S_B(\omega + \emptyset(\omega)) , \tag{3.18}$$

où $\emptyset$ représente la translation de la fréquence et C le changement d'amplitude des composants du spectre.

Soit $v_k(n)$ le vecteur qui dénote la représentation TF d'un signal de parole. Comme mentionné dans [4], les conditions aux limites périodiques doivent être utilisées car ils sont requises dans les applications pour la transformation des invariants.

Les conditions aux limites suivantes seront adoptées dans ce travail :

$$\begin{cases} v_k(n) = v_1(n) \quad , \quad pour\ k < 1 \\ v_k(n) = v_K(n) \quad , \quad pour\ k > K \end{cases} , \tag{3.19}$$

et

$$\begin{cases} Cv_k(n) = 0 \quad , \quad pour \quad Cv_k(n) < 0 \\ Cv_k(n) = 1 \quad , \quad pour \quad Cv_k(n) > 1 \end{cases} , \tag{3.20}$$

où $0 \leq v_k(n) \leq 1$.

Supposons que l'ensemble fini composé d'actions de translations et des multiplications, appliquées tout au long de l'espace des composantes spectrales, décrit les effets de la variabilité des locuteurs. Alors, un groupe fini G d'ordre | G | peut être défini.

La moyenne de groupe, dans ce cas, est donnée par :

$$A_f(\boldsymbol{v}) = \frac{1}{|G|} \sum_{g \in G} f(g\boldsymbol{v}). \tag{3.21}$$

Dans notre cas, de reconnaissance du locuteur, f transforme le spectre original $\mathrm{v_k}$ à $\mathrm{Cv_{k+i}}$ tel que montré par l'équation (3.18).

Les paramètres invariants sont définis comme la moyenne du groupe, en se basant sur les monômes m, tel que :

$$A(\boldsymbol{v}) = \frac{1}{2W+1} \sum_{i=-W}^{+W} m(v,i). \tag{3.22}$$

où $W \in \mathbb{N}_0$ représente la taille de la fenêtre. L'ensemble des monômes m est donné par :

$$m(v,i) = \prod_{k=1}^{K} C\, v_{k+i}^{b_k}, \tag{3.23}$$

où $b_k \in \mathbb{N}_0, i \in Z, et\ C \in \mathcal{R}$. L'ordre du monôme est défini par le terme :

$$\sum_{k=1}^{K} b_k. \tag{3.24}$$

Si on considère un monôme m d'ordre un, par exemple, alors :

$$b_k = \begin{cases} 0 & pour\ tout\ \ k \in \{1,2,\dots,K\} \setminus k_1 \\ 1 & pour \qquad k = k_1 \end{cases} . \tag{3.25}$$

La figure 3.5 résume l'opération complète d'extraction de paramètres invariants.

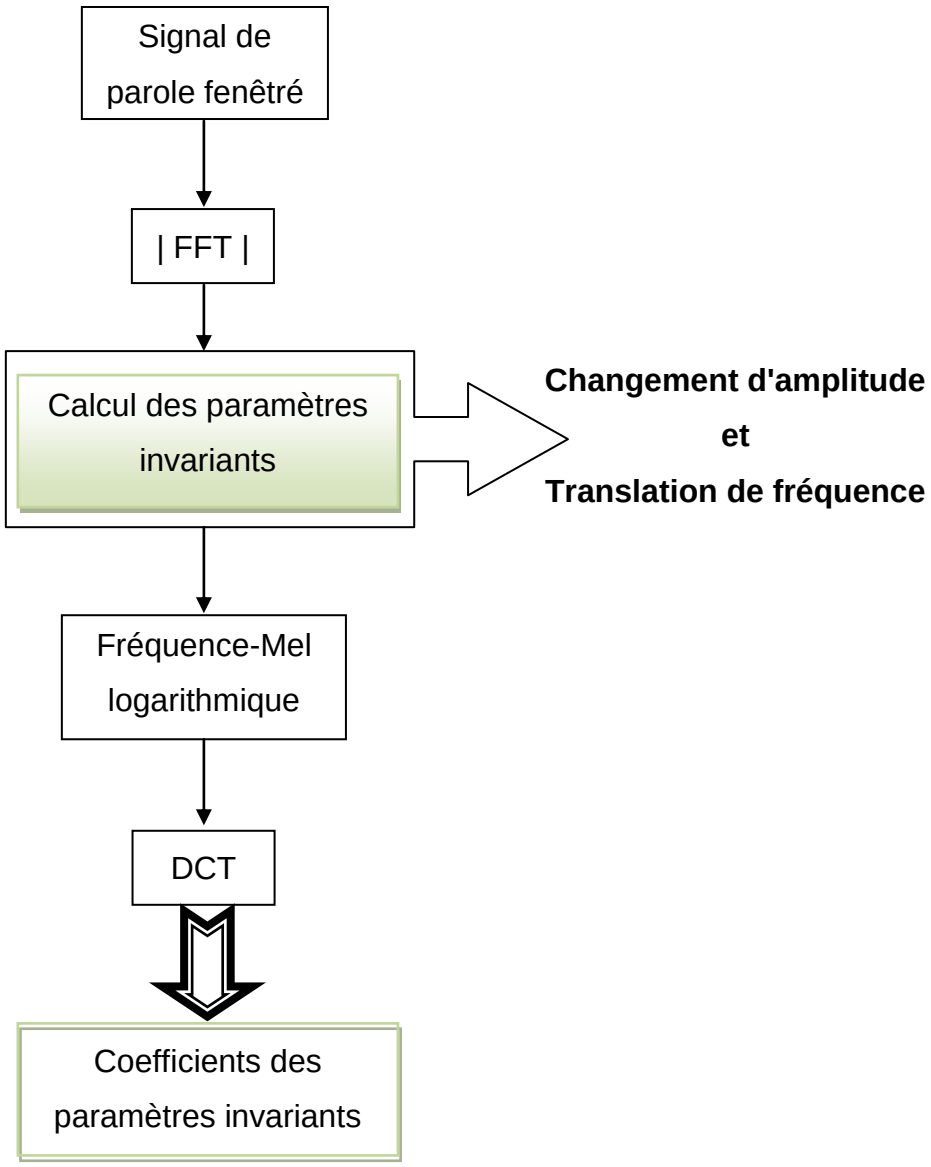


Figure 3.5 : Etapes de calcul des coefficients des paramètres invariants.

Comme expliqué dans les paragraphes précédents, des actions de changements en amplitude et des translations en fréquence sont appliquées au spectre pour obtenir l'ensemble des paramètres invariants.

Les monômes d'ordre un sont utilisés conformément a l'équation (3.22). Ceci est équivalent au calcul de la moyenne du spectre à certaines fréquences k. Nous calculons ensuite le logarithme de l'amplitude du spectre en fréquence mel. La transformation DCT est finalement appliquée pour obtenir des vecteurs non corrélés convenables à la technique de modélisation GMM-UBM. Les paramètres résultants seront appelés "paramètres invariants".

3.4. Conclusion

Ce chapitre a été consacré au traitement du problème de la variabilité. Par conséquent, nous l'avons commencé par définir les différentes sources de variabilité, puis les méthodes de compensation dans les systèmes de reconnaissance du locuteur. Ces méthodes ont été présentées selon le module du système de reconnaissance, c'est ainsi qu'on a vu la compensation au niveau de l'étage de prétraitement, puis au niveau de la modélisation et enfin au niveau de la décision. Par la suite, on a défini un nouveau type de paramètres basés sur la théorie de l'invariance. Nous avons modifié l'action d'invariance pour qu'elle soit appliquée en fréquence et en amplitude. Les paramètres invariants proposés ont été adaptés afin de pouvoir les utiliser dans le système de reconnaissance GMM-UBM pour compenser la variabilité due à la différence d'enregistrement entre l'apprentissage et le test.

CHAPITRE 4

EXPERIMENTATIONS ET RESULTATS

4.1. Introduction

Afin de pouvoir mieux comprendre les aspects théoriques, déjà vus dans les chapitres précédents, relatifs au système de reconnaissance du locuteur, nous effectuons dans ce chapitre plusieurs expériences où nous nous basons, à chaque fois, sur un paramètre pour voir son influence sur le système complet. Ces expériences vont être réalisées grâce à l'utilisation de la base de données MOBIO spécialement dédiée à la reconnaissance des personnes à travers la modalité de la parole ou la modalité du visage.

Les expériences vont, aussi, nous permettre de pouvoir juger l'applicabilité et l'efficacité de nos paramètres invariants proposés ainsi que l'idée de fusion de la méthode MFCC et celle des invariants.

Les évaluations des résultats vont être exprimées en termes du taux d'erreur égale (EER) et le minimum de la fonction minDCF pour chaque expérience réalisée.

4.2. Description de la base de données et du protocole d'expérimentation

Nos expériences ont été effectuées sur la base de données MOBIO [50,51]. Cette base de données est constituée de données de parole collectées à partir de 152 personnes (100 males, 52 femelles). L'enregistrement des fichiers de parole a été réalisé par des téléphones cellulaires et d'ordinateurs portables (laptops).

Les téléphones cellulaires utilisés pour la capture des enregistrements sont de marque NOKIA N93i, les laptops sont de marque MacBook standard 2008.

Pour chaque appareil les données sont enregistrées en format haute qualité avec une fréquence d'échantillonnage de 48 KHz. L'enregistrement de ces données s'est

effectué entre aout 2008 et juillet 2010 au sein de six différentes institutions (universités et laboratoires de recherche) appartenant à cinq différents pays.

En plus du fait que les sons de cette base de données sont bruités, la base de données a été enregistrée avec les défis suivants :

- **Variabilité de session :** Une seule session par client (locuteur) peut être utilisée pour apprendre les modèles des locuteurs cibles. Le test doit se faire en utilisant les cinq sessions restantes.
- **Variabilité lexicale :** la parole utilisée en apprentissage et en test est composée de différents lexiques (appropriée pour la reconnaissance du locuteur indépendamment du texte).
- **Différence de type de parole :** L'apprentissage se base sur des enregistrements de locuteur lisant un texte (parole lue), alors qu'en phase de test le locuteur est libre de dire ce qu'il veut (parole libre).
- **Différence d'endroits des locuteurs :** Les données relatives aux imposteurs permettant de construire le modèle du monde en apprentissage proviennent de deux sites (institutions universitaires). Les donnes utilisées en test proviennent des quatre sites restants.
- **Différence d'appareils d'enregistrement :** Le même locuteur peut être appris en utilisant des paroles issues de la session laptop ou la session téléphone mobile donnant deux types de modèles, un dépendant du laptop et l'autre du téléphone mobile.

En phase d'expérimentation, nous avons suivi le même protocole que celui donné en [50,51], par conséquent, nous avons divisé la base de données en trois ensembles distincts : un ensemble d'apprentissage, un ensemble de développement, et un ensemble de test. Les trois ensembles sont complètement séparés en termes de locuteurs et de sites de données collectés.

L'objectif de l'ensemble de développement est de dériver les meilleurs paramètres pour le système de reconnaissance du locuteur. Ces paramètres seront ensuite

utilisés dans l'ensemble de test pour l'évaluation des performances de notre système.

Les protocoles d'apprentissage et de test sont les mêmes pour les deux ensembles d'évaluation et de test.

Cinq ensembles de questions - réponses issues d'une seule session peuvent être utilisés pour la modélisation des locuteurs clients.

La reconnaissance (le test) est effectuée sur tous les 15 fichiers, de parole libre, issues de chacune des 5 sessions restantes, donnant 75 enregistrements de test par client.

Les scores des imposteurs se calculent en considérant tous les clients qui restent comme des imposteurs.

Les performances sont calculées en termes de taux d'erreur égale (Equal Error Rate EER) sur l'ensemble test et la fonction coût de détection (minimum detection cost function minDCF).

4.3. Résultats et discussion

Dans toutes nos expériences nous utilisons un système GMM - UBM dépendant du sexe, ainsi, deux modèles UBM, un pour les males et un autre pour les femelles, sont appris en utilisant toutes les données collectées des deux sessions (une session des enregistrements effectuées sur laptops et une session des enregistrements sur téléphones mobiles) provenant de deux sites.

Les modèles des locuteurs sont obtenus en utilisant le reste des sites comme expliqué dans le paragraphe précédent. Ce choix est justifié par le fait que la variabilité due à la différence de sexe entre les données d'apprentissage et de test est sans intérêt dans notre cas, où nous considérons la variabilité due à la différence de type d'enregistrements ou de canal.

Nous commençons nos expériences par l'utilisation des modèles relatifs aux locuteurs de sexe male pour étudier nos paramètres proposés dans différentes situations. Nous consacrons la dernière expérience pour montrer les performances

du système de reconnaissance du locuteur pour des essais en utilisant les locuteurs par rapport aux essais en utilisant des locutrices.

Les modules de prétraitement et d'extraction des paramètres sont implémentés sur le logiciel Matlab (Matrix Laboratory), quand aux restes des modules (modélisation et prise de décision), ils sont réalisés en utilisant la plate forme ALIZE [52] sous Linux pour l'obtention du modèle du monde UBM, des modèles des locuteurs, et le calcul des rapports de vraisemblances logarithmiques (scores).

5 itérations sont utilisées dans l'algorithme EM (Expectation - Maximisation) pour modéliser un mélange de 512 Gaussiennes du modèle UBM.

L'adaptation des modèles des locuteurs est effectuée via une seule itération dans l'algorithme MAP (Maximum A Posteriori). Seules les moyennes des Gaussiennes sont considérées durant l'étape d'adaptation.

Pour l'évaluation des performances, nous traçons la courbe DET (Detection Error Tradeoff), représentant la probabilité de fausse rejection (FR) en fonction de la probabilité de fausse alarme (ou fausse acceptation FA). Le point de la courbe pour lequel la probabilité de FR est égale à la probabilité de FA est appelé EER (Equal Error Rate) est un très puissant indicateur de performance des systèmes de reconnaissance du locuteur. Aussi, un autre point très utile à savoir le minDCF. Ces points sont adoptés dans les expériences de cette thèse.

Actuellement, les paramètres MFCC sont les plus utilisés pour la caractérisation de la parole dans les systèmes de reconnaissance des locuteurs [52,53,54]. Sa popularité vient du fait que ses paramètres sont faciles à extraire, et peuvent réalisés de meilleurs taux de reconnaissance par rapport aux autres méthodes conventionnelles. Nous allons par la suite montrer expérimentalement ce fait, qui était la raison principale de notre choix de comparer la méthode des invariants proposée avec la méthode MFCC.

Notre système GMM-UBM de base utilise 19 coefficients des paramètres MFCC. Il est basé sur un banc de 24 filtres calculés sur des trames de fenêtres Hamming de durées 20 ms. Les signaux de parole sont sous échantillonnés à 8 KHz (pour être en

adéquation avec les signaux issus des téléphones). Un taux de chevauchement de 10 ms entre deux trames successives est adopté.

Les paramètres MFCC sont augmentés par 19 coefficients delta (dérivée première), un coefficient delta log-énergie, et 11 coefficients double delta (dérivée deuxième). Le vecteur total est, par conséquent, composé de 50 coefficients [56].

Les paramètres invariants proposés utilisent des monômes d'ordre 1. Le choix de cet ordre a été expliqué dans [4] et les paramètres issus a partir de ce choix sont ceux qui donnent les meilleures performances de reconnaissance. Similairement aux paramètres MFCC, nous utilisons 19 coefficients des paramètres invariants. On leur rajoute 19 coefficients delta, un coefficient delta log-énergie, et 11 coefficients double delta.

Dans notre travail, la translation de fréquence est effectuée sur tout le spectre, alors que l'action de multiplication, par le facteur C, est appliquée sur les maximums d'amplitude avec un changement de 10 %. Nous concentrons notre effet de changement d'amplitude sur les maximums car les formants contiennent une information considérable sur l'identité des locuteurs.

Tout au long de nos expériences, nous représentons chaque locuteur par deux modèles, un modèle en utilisant le téléphone mobile et l'autre en utilisant le laptop. Nous testons, ensuite, les locuteurs dans le cas de correspondance (match case) et dans le cas de non correspondance (mismatch case). Dans le premier cas, nous comparons la parole enregistrée avec le téléphone mobile contre le modèle du locuteur enregistré, toujours, avec le téléphone mobile. Dans le deuxième cas, les expériences sont réalisées en comparant la parole enregistrée en utilisant le téléphone mobile et le modèle du locuteur enregistré par laptop.

1ère expérience :

Dans cette expérience, nous avons deux systèmes GMM-UBM tels que décrits précédemment, l'un est basé sur les paramètres MFCC, l'autres est basé sur les paramètres PLP. La même configuration, décrite précédemment pour les paramètres MFCC, a été reprise pour les paramètres PLP.

Tableau 4.1 : Performances du système de reconnaissance en utilisant les paramètres MFCC et PLP.

		Même	Différent
EER (%)	**MFCC**	17.96	21.39
	PLP	20.94	25.28
minDCF (x100)	**MFCC**	7.42	8.03
	PLP	7.82	8.52

Pour la suite de cette thèse, nous appellerons "même" le cas de correspondance où les enregistrements utilisés dans l'apprentissage et le test sont effectués sur le même type d'appareil (soit en utilisant le téléphone mobile pour les deux soit le laptop). Et nous appellerons "différent" le cas de non correspondance où les enregistrements utilisés dans l'apprentissage et le test ne sont pas effectués sur le même type d'appareil (si le téléphone mobile est utilisé pour l'enregistrement en phase d'apprentissage le laptop enregistre les fichiers de test et vice versa).

Les résultats montrés dans le tableau 4.1 permettent de confirmer la supériorité des paramètres MFCC sur les paramètres PLP en termes de reconnaissance, par conséquent, une nette réduction des erreurs EER et minDCF est obtenue avec l'utilisation des MFCCs. Par la suite de notre travail, ces paramètres MFCC seront utilisés comme référence de comparaison pour les différentes expériences.

2ème expérience :

Cette expérience vise à étudier les différents aspects de l'étape d'extraction des paramètres et son influence sur les performances de la reconnaissance du locuteur.

Le tableau 4.2 donne les résultats d'évaluation relatifs au système de reconnaissance pour ces différentes configurations.

Tableau 4.2 : Illustration des performances du système de reconnaissance du locuteur en trames de EER et minDCF pour les paramètres MFCC et invariants dans le cas de correspondance (même) et non correspondance (différent), avec normalisation et sans normalisation

.

	Sans normalisation CMN		Avec normalisation CMN	
	Même	Différent	Même	Différent
EER (%) **MFCC**	19.21	25.75	17.96	21.39
EER (%) **Paramètres Invariants** (Sans action d'amplitude)	19.08	25.23	17.37	22.18
EER (%) **Paramètres Invariants** (Avec action d'amplitude)	18.36	25.06	16.94	20.59

Les différents points considérés dans cette expérience sont détaillés comme suit :

a. La comparaison entre les résultats correspondants aux colonnes de mêmes types d'appareils entre la modélisation et le test, notées par "même", et l'utilisation de différents types d'appareils, notées par "différent", montre l'effet de la variabilité du canal sur la performance du système de reconnaissance. C'est ainsi qu'on constate une grande dégradation des résultats par l'augmentation de l'erreur EER pour tous les paramètres utilisés (MFCC et paramètres invariants).
   Cet effet de dégradation est illustré avec plus de détails par la courbe DET représentée sur la figure 4.1. A partir de cette courbe, nous observons que les taux d'erreurs de fausses alarmes et de fausses rejections augmentent significativement du cas de correspondance (Match) au cas de non correspondance (Mismatch).

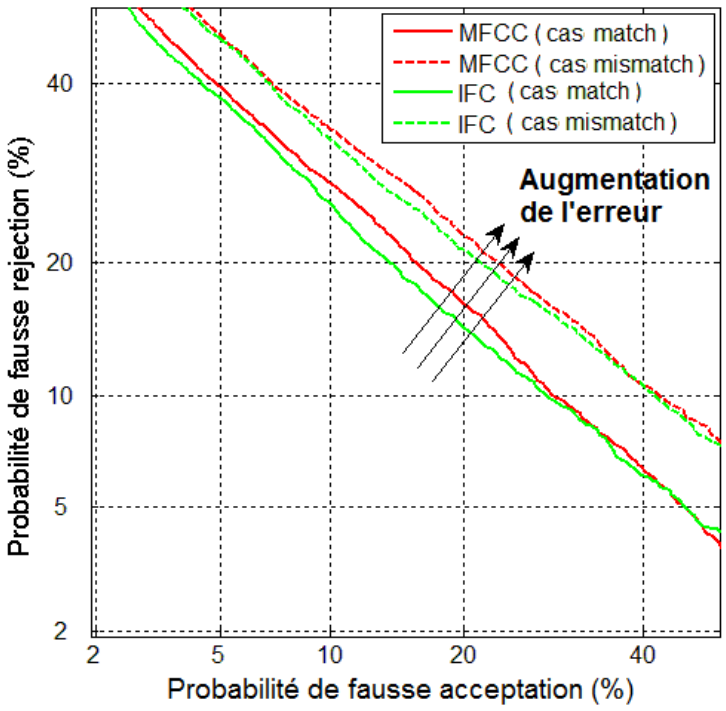


Figure 4.1 : Courbe DET montrant l'effet de dégradation dû au cas de non correspondance (mismatch) (lignes en pointillés) par rapport au cas de correspondance (match) (lignes continues).

b. Pour montrer l'effet d'utiliser l'invariance en amplitude en plus de l'invariance en fréquence sur les performances du système de reconnaissance, nous avons comparé les paramètres invariants lors de l'application, seule, de l'action de translation de fréquence puis en combinant les deux actions, la translation de la fréquence et le changement d'amplitude. Les résultats obtenus dans le tableau 4.1 montrent clairement la contribution positive due à l'ajout de l'action du changement d'amplitude par rapport aux performances obtenues avec l'action de translation de fréquence seule (sans changement d'amplitude). Une réduction en termes d'erreur EER est ainsi constatée. Cette amélioration est plus significative dans le cas de différence des types d'enregistrements entre l'apprentissage et le test.

c. Un autre point étudié dans cette expérience concerne l'apport de l'utilisation d'une normalisation de prétraitement sur la reconnaissance du locuteur et la variabilité. Pour cela, nous avons choisi la méthode de soustraction de la moyenne (CMN) connue par sa simplicité et son efficacité à éliminer les distorsions temporelles introduites par les canaux de transmissions. Cette normalisation a été appliquée sur les deux paramètres MFCC et invariants. A travers les résultats représentés dans le tableau 4.1, nous avons confirmé les conclusions théoriques. Une amélioration notable est réalisée en normalisant les paramètres de parole par la méthode CMN.
   Nous allons appliquer cette méthode de normalisation pour le reste de nos expériences.

d. Enfin, la comparaison entre les performances de reconnaissance pour le système utilisant les paramètres invariants et les paramètres MFCC, montre l'amélioration apportée par les paramètres invariants dans les deux cas (match et mismatch). Ce résultat permet de confirmer l'efficacité d'utiliser le concept d'invariance dans la reconnaissance du locuteur.

3ème expérience :

Dans l'adaptation des modèles GMM par l'algorithme MAP, le facteur de régulation contrôle la quantité de données qui doit être observée dans le mélange avant que les nouveaux paramètres commencent à remplacer les anciens. Cette nouvelle expérience nous montre l'effet de changer ce facteur sur les performances du système de reconnaissance du locuteur.

Dans nos expériences précédentes nous avions un facteur de régulation égal à 14, conformément aux conclusions de [36] et le travail de bimbot et al. [57].

Ici, nous allons changer les valeurs du facteur de régulation, et pour chaque valeur nous allons effectuer une nouvelle modélisation des locuteurs, et sur cette nouvelle base nous recalculerons les nouveaux rapports de vraisemblance (scores) permettant de calculer les erreurs EER.

Le tableau 4.3 donne les valeurs des erreurs en EER pour les deux paramètres, MFCC et invariants pour des valeurs de facteur de régulation variant entre 2 et 22.

Tableau 4.3 : Effet du changement de la valeur du facteur de régulation sur les performances du système de reconnaissance du locuteur.

	Relevance factor	**EER (%) (MFCC)**	**EER (IFC) (%)**
Même	2	17.10	16.65
	6	17.23	16.67
	10	17.51	16.83
	14	17.96	16.96
	18	18.45	17.17
	22	18.63	17.24
Différent	2	20.74	20.04
	6	20.88	20.22
	10	21.08	20.25
	14	21.39	20.59
	18	21.91	21.12
	22	22.05	21.12

Les valeurs du tableau 4.3 montrent que les meilleurs résultats sont réalisés pour des facteurs de régulation compris entre 2 et 10. Pour la suite de notre travail, nous fixons la valeur du facteur de régulation à 6.

4ème expérience :

Pour profiter des avantages des paramètres MFCC en termes de puissance de discrimination et des paramètres invariants en termes de compensation de la

variabilité, nous proposons la fusion de ces deux méthodes dans un seul système de reconnaissance.

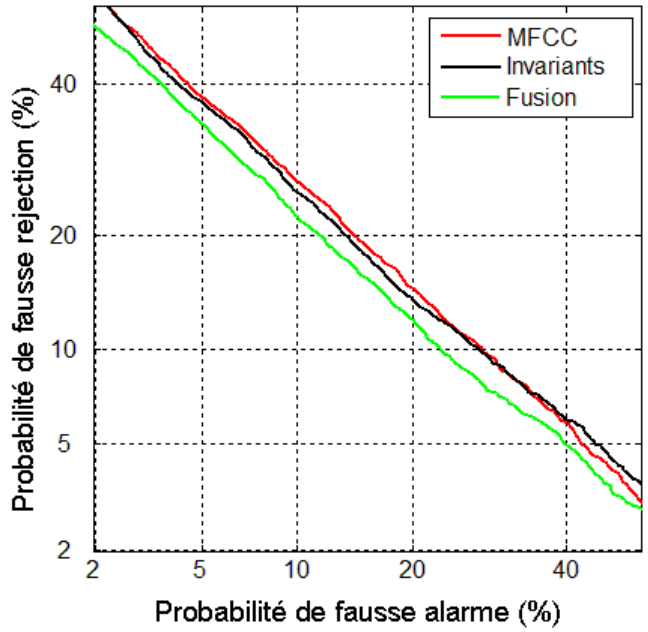
Une combinaison linéaire pondérée des scores relatifs aux paramètres MFCC et ceux relatifs au paramètres invariants est décrite dans [58]. Les nouveaux scores du système de reconnaissance aura la forme suivante :

$$scores = \beta \times scores\,(MFCC) + (1-\beta) \times scores\,(invariants). \qquad (4.1)$$

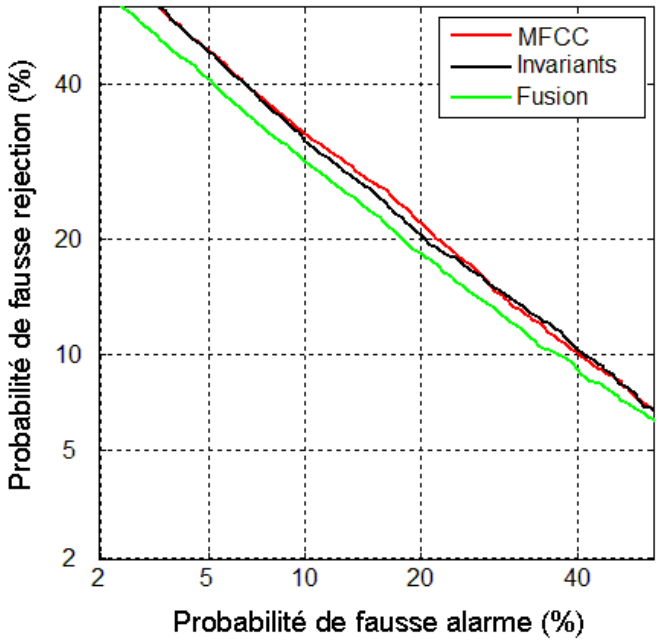
Dans le contexte de la reconnaissance du locuteur basée sur le système GMM-UBM, les scores réfèrent aux fonctions des rapports de vraisemblances logarithmiques $\mathrm{P}(\mathrm{X}|\lambda)$, où X représente le vecteur des paramètres extrait par la méthode MFCC pour les scores MFCC, alors que pour les scores invariants, X représente le vecteur des paramètres extrait par la méthode proposée d'invariance.

Nous remarquons que la méthode de fusion des deux types de paramètres dépend du paramètre β. Alors, nous avons varié, expérimentalement, ce paramètre entre les valeurs 0 et 1, afin de trouver le meilleur résultat (correspondant au minimum d'erreur). Les résultats trouvés montrent que la valeur $\beta = 0.5$ permet d'atteindre cet objectif, puisque les erreurs EER et minCDF sont minimales pour cette valeur.

La figure 4.2 montre les courbes DET correspondant aux trois systèmes de reconnaissance du locuteur basés sur les méthodes : MFCC, d'invariance, et la fusion des deux. La figure 4.2.a représente le cas de correspondance (match) et la figure 4.2.b représente le cas de non correspondance (mismatch).



(a)



(b)

Figure 4.2 : Courbes DET dans le cas de correspondance (match) (a) et non correspondance (mismatch) (b).

Le tableau 4.4 suivant donne les valeurs des erreurs EER et minDCF pour la technique proposée de fusion des paramètres MFCC et invariants.

Tableau 4.4 : Performances de la technique de fusion MFCC-invariants.

	Même	**Différent**
EER (%) MFCC - invariants	15.62	19.03
minDCF (x100) MFCC - invariants	6.79	7.41

Comme prévu, une amélioration additionnelle est obtenue par la technique de fusion dans les deux cas (match et mismatch).

La dégradation causée par l'utilisation de différentes conditions d'enregistrement (téléphone mobile ou laptop) est compensée par une amélioration de prés de 10% pour la technique proposée de fusion des paramètres MFCC-invariants comparativement à la méthode de parametrisation MFCC. Encore, en termes d'erreur minDCF la technique de fusion proposée fonctionne mieux que les paramètres MFCC, avec une amélioration de plus de 6%.

5ème expérience :

L'objectif de cette expérience est de tester le système de reconnaissance du locuteur en utilisant les trois paramètres : MFCC, invariants, et fusionnés pour les deux sexes de locuteurs de la base de données MOBIO.

Pour avoir un meilleur aperçu sur les performances relatives à chaque locuteur, nous calculons les deux valeurs d'évaluation qui sont l'erreur EER et la minDCF. Ces valeurs correspondant à chaque locuteur vont nous servir pour le calcul de la moyenne EER et minCDF de tous les locuteurs.

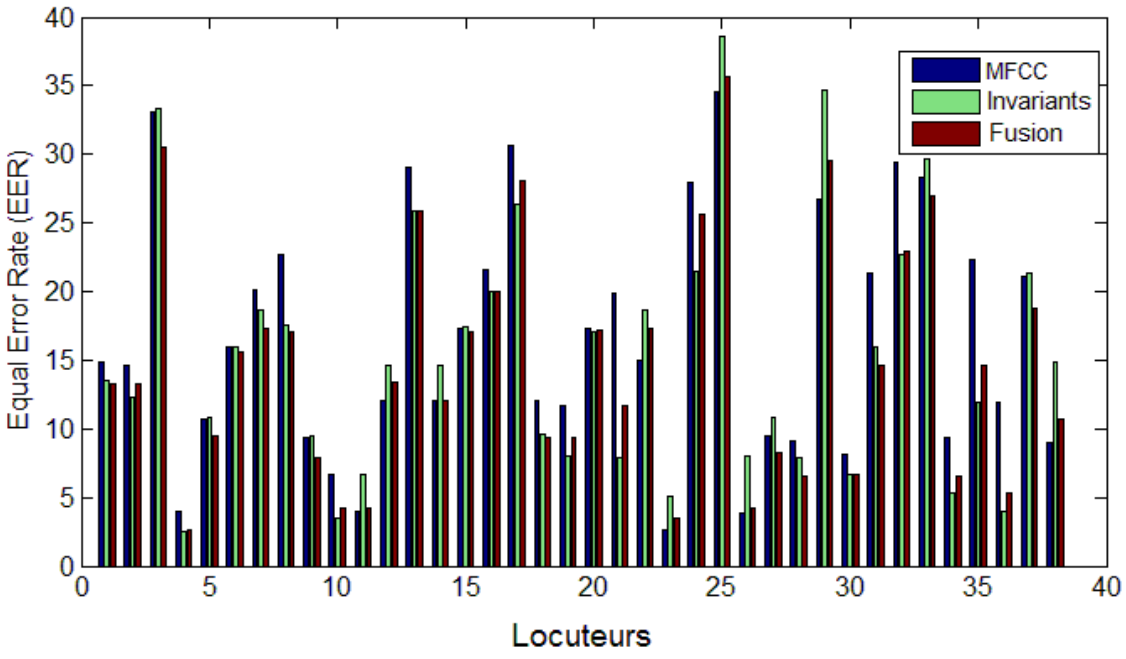
Le tableau 4.5 montre les résultats obtenus pour les deux sexes en utilisant les trois types de parametrisation. Ces résultats restent en adéquation avec toutes les conclusions tirées jusque là. Particulièrement, la supériorité de performance de la technique proposée par rapport à la méthode MFCC pour le cas des locutrices comme pour le cas des locuteurs.

Tableau 4.5 : Moyenne en EER et minDCF de tous les locuteurs de test calculée à partir des EER et minDCF individuels.

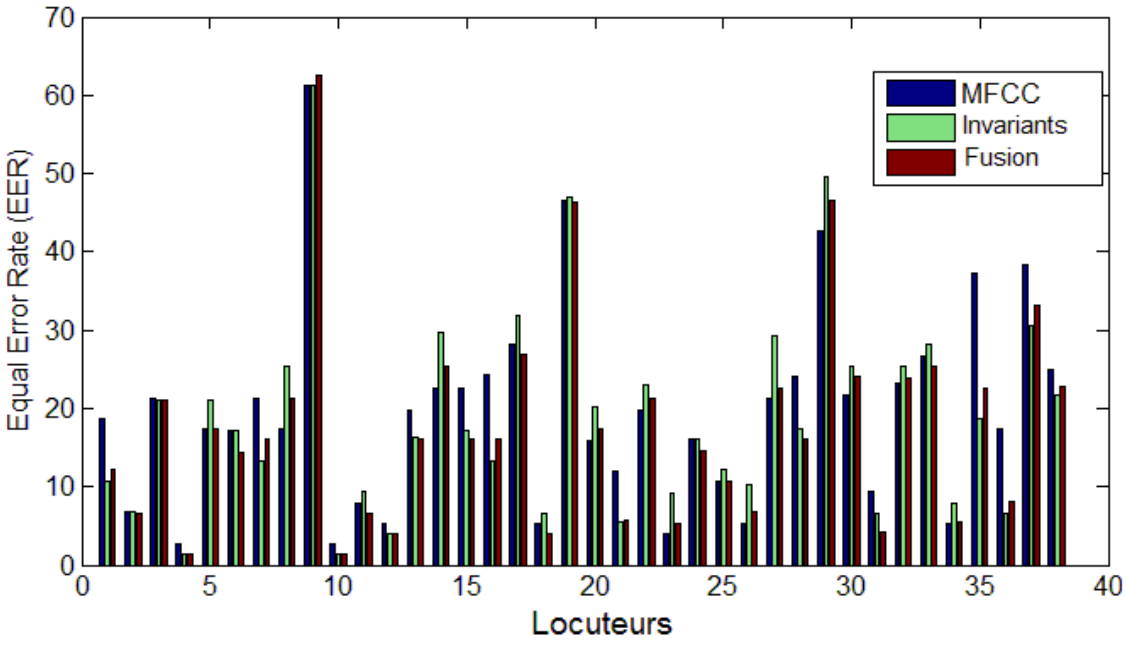
.

	Femelles		**Males**	
	Même	Différent	Même	Différent
EER (%) **MFCC**	17.40	22.11	16.05	18.94
Min DCF (x100) **MFCC**	6.77	8.08	6.71	7.26
EER (%) **Paramètres Invariants**	16.57	21.49	15.16	18.63
minDCF (x100) **Paramètres Invariants**	6.65	7.97	6.55	7.21
EER (%) **Fusion MFCC-Invariants**	16.36	20.85	14.19	17.33
minDCF (%) **Fusion MFCC-Invariants**	6.40	7.84	6.29	6.80

Les représentations montrées sur la figure 4.3 permettent de voir avec détails les valeurs des paramètres d'évaluations (EER et minDCF) pour chacun des locuteurs de la base de données et dans les deux cas (match et mismatch).



(a)



(b)

Figure 4.3 : L'erreur EER pour chaque locuteur dans le cas de correspondance (match) (a) et non correspondance (mismatch) (b).

A partir des figures 4.3.a et 4.3.b nous pouvant remarquer que dans la majorité des cas une amélioration des résultats de reconnaissance est réalisée.

6ème expérience :

Pour voir l'influence du nombre de coefficients sur les performances du système de reconnaissance du locuteur, une deuxième configuration est considérée dans cette partie. Le nouveau système GMM-UBM utilise 29 coefficients, qui seront augmentés par 29 coefficients delta (dérivée première), un coefficient delta log-énergie, et 11 coefficients double delta (dérivée deuxième). Le vecteur total est, par conséquent, composé de 70 coefficients. Cette configuration sera adoptée par les trois types de paramétrisation déjà étudiés, a savoir, les paramètres MFCC, invariants et la fusion MFCC-invariants. Nous garderons les mêmes paramètres de configuration pour les autres étapes (prétraitement, modélisation). Le tableau 4.6 donne les résultats de performance en termes des erreurs EER et minDCF.

Tableau 4.6 : Influence du nombre de paramètres sur les performances du système de reconnaissance du locuteur.

	Même	**Différent**
EER (MFCC) (%)	15.51	18.20
minDCF (MFCC) (x100)	6.70	7.20
EER (IFC) (%)	14.86	16.21
minDCF (IFC) (x100)	6.39	6.86
EER (MFCC-IFC) (%)	13.76	15.65
minDCF (MFCC-IFC) (x100)	6.11	6.62

A partir de ces résultats, on constate qu'une amélioration de performances est obtenue par l'utilisation de 70 coefficients au lieu 50 coefficients (voir Tableau 4.5), où une réduction des EER et de minDCF est produite. Il est à noter que les même constats tracés dans les expériences précédentes sont valables ici, puisque les paramètres invariants proposés sont toujours meilleurs que les parametres MFCC en termes des erreurs EER et minDCF dans les deux cas, de correspondance

(match) et non correspondance (mismatch). Aussi, la fusion des deux méthodes de paramétrisation MFCC et invariants permet de mieux compenser la variabilité

4.4. Conclusion

Ce chapitre à été consacré a l'expérimentation du système de reconnaissance du locuteur. Les objectifs étaient de vérifier les notions théoriques acquises, et la compensation de la variabilité par l'utilisation des paramètres invariants proposées.

Pour atteindre nos objectifs, plusieurs expériences ont été réalisées. La variation des paramètres à tester nous permettait d'étudier leur influence sur les performances du système de reconnaissance. La discussion des résultats nous a permis de vérifier, l'efficacité des paramètres invariants comparées aux paramètres MFCC, puis l'avantage de fusionner les deux scores des paramètres MFCC et invariants pour obtenir plus d'amélioration surtout dans le cas de non correspondance (mismatch).

CONCLUSION GENERALE

Dans cette thèse, nous avons mené un travail de recherche dans le domaine de la reconnaissance du locuteur afin d'améliorer la robustesse des systèmes vis à vis au problème de la variabilité.

Dans un premier temps, on a commencé par étudier l'influence de la variabilité sur le système de reconnaissance du locuteur, en considérant le cas de non correspondance (mismatch) entre l'apprentissage et le test. Puis, sur cette base on a opté pour le choix de compenser cette variabilité au niveau de l'étage d'extraction des paramètres.

Dans cet objectif, nous avons étudié les paramètres invariants, connus par leur efficacité à traiter certains types de variabilités environnementales dans les domaines d'analyse d'images ou de la reconnaissance de la parole, puis nous les avons adaptés pour les appliquer dans la reconnaissance des locuteurs. Par rapport aux paramètres invariants existants, nous avons ajouté, l'action de changement d'amplitude. Ces paramètres ont été intégrés dans la partie d'extraction de paramètres du système globale de reconnaissance du locuteur.

Pour valider nos connaissances et tester l'efficacité de notre choix de paramètres en termes de performance de reconnaissance de locuteurs en général et en présence de variabilité en particulier, des expériences ont été effectuées.

Dans ces expériences, on a commencé par montrer l'intérêt de comparer nos paramètres avec la méthode la plus utilisée en reconnaissance du locuteur, qui est la méthode MFCC. Puisque expérimentalement, les paramètres MFCC sont plus performants que les paramètres PLP, sur la base de données MOBIO. Nous avons ensuite, montré la dégradation des performances dans le cas d'utilisation de différents types d'enregistrements entre le test et l'apprentissage (cas mismatch) par rapport au cas de même type d'enregistrement (cas match). Les paramètres invariants ont, par la suite, été validés dans le contexte de reconnaissance du locuteur dans les deux cas (match et mismatch). Quoique les résultats obtenus,

donnaient l'avantage aux paramètres invariants par rapport aux paramètres MFCC, nous avons pensé à combiner les scores de ces deux paramètres pour bénéficier du pouvoir de discrimination des paramètres MFCC et la compensation de la variabilité des paramètres invariants. Comme prévu, les résultats obtenus permettaient d'améliorer encore les performances et on a constaté une réduction des erreurs EER et minDCF. D'autres expériences ont été effectuées afin de clarifier les résultats ou de montrer l'amélioration que peuvent apporter certaines méthodes comme la normalisation CMN au niveau du prétraitement.

Les résultats encourageants obtenus ouvrent d'avantages de perspectives pour le développement d'autres types de paramètres invariants pour le traitement de la variabilité en reconnaissance du locuteur. L'étude d'autres sources de variabilités et leurs effets sur la parole originale peut être incorporée dans ces paramètres pour pouvoir les compenser.

BIBLIOGRAPHIE

1. Brown, M., Lowe, D.G. "Recognising panoramas". In Proceedings of the International Conference on Computer Vision (ICCV 2003), Nice, France, 14–17, pp. 1218–1225, (October 2003).

2. Minematsu, N., Asakawa, S., Suzuki, M., and Qiao, Y. "Speech Structure and Its Application to Robust Speech Processing". New Generation Comput, 28(3): 299-319, (2010).

3. Qiao, Y., and Minematsu, N. "A study on invariance of f-divergence and its application to speech recognition". IEEE Transactions on Signal Processing, 58(7): 3884-3890, (2010).

4. Müller F., and Mertins, A. "Contextual invariant-integration features for improved speaker independent speech recognition". Speech Communication, vol. 53, no. 6, pp. 830 - 841, (2011).

5. Alimohad, A., Bouridane, A., Guessoum, A. " Efficient Invariant Features for Sensor Variability Compensation in Speaker Recognition ", *Sensors*, *14*, 19007-19022, (2014).

6. Fazel, A., Chakrabartty, S. " An Overview of Statistical Pattern Recognition Techniques for Speaker Verification ", IEEE Circuits and Systems Magazine (MCAS), (2011).

7. Li, S. Z. and Jain, A. K., "Encyclopedia of Biometrics", Springer, (2009).

8. Petrovska-Delacrétaz, Dijana., Chollet, Gérard., Dorizzi, Bernadette., " Guide to Biometric Reference Systems and Performance Evaluation ", Springer, (2009)

9. Zhang, D. Y., Jain, A. K. (Eds.), " Biometric Authentication ",First International Conference, ICBA 2004, Hong Kong, China, July 15-17, (2004).

10. Cutting, J. E., and Kozlowski, L. T., " Recognition of friends by their walk ", Bulletin of the Psychonomic Society, 9:353–356, (1977).

11. Grassi, S., " Optimized Implementation of Speech Processing Algorithms ", PhD Thesis, Université de Neuchâtel, (February 1998).

12. Fant, G. " Acoustic Theory of Speech Production ", Mouton De Gruyter, (1970).

13. Nezzari, H., " introduction d'une fenetre de ponderation fractionnaire et son utilisation en filtrage numerique RIF ", Mémoire de magister en électronique (Université de Constantine), (2005).

14. Rabiner, L., and Juang, B.H., " Fundamentals of Speech Recognition ", Englewood Cliffs, NJ: Prentice-Hall, (1993).

15. Furui, S. " Digital Speech Processing, Synthesis and Recognition ", 2nd Edition, Marcel Dekker Inc., New York, (2001).
16. Huang, X., Acero, A., Hon, H., and Reddy, R., " Spoken Language Processing: A Guide to Theory, Algorithm and System Development ". Prentice-Hall, 2001.

17. Margot, P., " Un changement de nom dans la continuité ", Revue internationale de criminologie et de police technique et scientifique, 52(1): 6 - 8, (1999).

18. Tounsi, b., " Inférence d'identité dans le domaine forensique en utilisant un système de reconnaissance automatique du locuteur adapté au dialecte Algérien ", Mémoire de magister en informatique (INI), (2007).

19. Srinivasan, K. and Gersho, A., " Voice activity detection for cellular networks ", in Proc. IEEE Speech Coding Workshop, pp. 85-86, Oct. (1993).

20. Alam, J., Kinnunen, T., Kenny, P., Ouellet, P., and O Shaugnessy, D., " Mutlitaper MFCC and PLP Features for Speaker Verification Using I-Vectors ", Speech Communication, 55(2), pp. 237-251, (February 2013).

21. Shriberg, E., " Higher-Level Features in Speaker Recognition ", Lecture Notes in Artificial Intelligence, Springer, vol. 4343, (2007).

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

23. Kinnunen, T., and Li, H., " An Overview of Text-Independent Speaker Recognition: from Features to Supervectors ", Speech Communication 52(1): 12-40, (January 2010).

24. Harrington, J., and Cassidy, S., " Techniques in Speech Acoustics ", Kluwer Academic Publishers, Dordrecht, (1999).

25. Besacier, L., Bonastre, J., and Fredouille, C., " Localization and selection of speaker specific information with statistical modeling ", Speech Communication 31, 89-106, (June 2000).

26. Besacier, L., Bonastre, J., " Subband architecture for automatic speaker recognition ", Signal Processing 80, 1245-1259, (July 2000).

27. Hermansky, H., " Perceptual Linear Prediction (PLP) analysis of speech ", J. Acoust. Soc. America, Vol. 87, 1738-1753, (April 1990).

28. Furui, S., " Speaker-independent isolated word recognition based on emphasized spectral dynamics ", Proc. ICASSP, (1986).

29. Jourani, R., " Reconnaissance automatique du locuteur par des GMM à grande marge ", thèse de doctorat, Université de Rabat, (2012).

30. Meuwly, D., "Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique ", PhD dissertation, University of Lausanne, Lausanne, Switzerland, (2001).

31. Rudasi, L., and Zahorian, S. A., " Text- independent talker identification with neural netowkrs ", Proc. IEEE ICASSP, pp. 389-392, (May 1991).

32. Cristianini, N., and Shawe-Taylor., J, " An Introduction to Support Vector Machines and Other Kernel-based Learning Methods ", Cambridge University Press, (2000).

33. DJEFFAL, A., " Utilisation des méthodes Support Vector Machine (SVM) dans l'analyse des bases de données ", Thèse de Docteur en science Spécialité Informatique, Université Mohamed Khider – Biskra, (2011/2012).

34. Baum, L.E, and Eagon, J.A., " An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology ", Bulletin of American Mathematical Society, 73, pp. 360-363, (1967).

35. Reynolds, D.A. and Rose, R.C., " Robust Text independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. SAP, pp.72-83, (January 1995).

36. Reynolds, D.A., Quatieri, T.F., and Dunn, R.B., " Speaker Verification Using Adapted Gaussian Mixture Models ", Digital Signal Processing, vol.lO, pp.19-41, (2000).

37. Dempster, A.P., Laird, N.M., and Rubin, D.B., " Maximum Likelihood from Incomplete Data via the EM Algorithm ", Journal of the Royal Statistical Society Series, Vo1.39, pp.1-38, (1977).

38. Gauvain, J.-L., and Chin-Hui, " Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains ", IEEE Transactions on Speech and Audio Processing, 2(2):291-298, (1994).

39. Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M., "The DET curve in assessment of detection task performance", In Proceedings of the European Conference on Speech Communication and Technology, pp. 1895-898, (1997).

40. Büyük, O., "Telephone-Based Text-Dependent Speaker Verification", PhD dissertation, Boğazici University, (2011).

41. Reynolds, D., " Channel robust speaker verification via feature mapping ", Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2003), Hong Kong, China, Volume 2. 31, (2003).

42. Pelecanos, J., and Sridharan, S., " Feature warping for robust speaker verification ", Speaker Odyssey, The Speaker Recognition Workshop, Chania, Greece, Chania, Crete, pp. 213–218. 18, (2001).
43. Solomonoff, A., Campbell, W., and Boardman, I., " Advances in channel compensation for SVM speaker recognition," in Proc. ICASSP, vol. 1, pp. 629-632, (2005).

44. Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P. " Factor analysis simplified ", In Proc. of ICASSP, volume 1, pages 637-640, (2005).

45. Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P. " Improvements in factor analysis based speaker verification ", In Proc. of ICASSP, volume 1, pages 113-116, (2006).

46. Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P., " Joint Factor Analysis Versus Eigenchannels in Speaker Recognition ", IEEE Transactions on Audio, Speech and Language Processing, 15(4) :1435-1447, (2007).

47. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P. " A study of interspeaker variability in speaker verification ", IEEE Transactions on Audio, Speech and Language Processing, 16(5) :980-988, (2008).

48. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., " Front-End Factor Analysis for Speaker Verification," Audio, Speech, and Language Processing ", IEEE Transactions on , vol.19, no.4, pp.788,798, (May 2011).

49. Schulz-Mirbach, H., " On the existence of complete invariant feature spaces in pattern recognition ", InProc. Int. Conf. Pattern Recognition, Vol. 2. Hague, Netherlands, pp. 178-182, (1992).

50. Marcel, S., McCool, C., Matejka, P., Ahonen, T., and Cernocky, J., " On the results of the first mobile biometry (mobio) face and speaker verification evaluation ", In Proc. of ICPR, pages 210-225, (2010).

51. McCool, C., Marcel, S., Hadid, A., Pietikäinen, M., Matejka, P., Cernocky, J., Poh, N., Kittler, J., Larcher, A., Lévy, C., Matrouf, D., Bonastre, J-F., Tresadern, P., and Cootes, T., " Bi-Modal Person Recognition on a Mobile Phone: using mobile phone data ", in IEEE ICME Workshop on Hot Topics in Mobile Mutlimedia, (2012).

52. Bonastre, J-F., Scheffer, N., Matrouf, D., Fredouille, C., Larcher, A., Preti, A., Pouchoulin, G., Evans, N., Fauve, B., and Mason J., " ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition. Speaker Odyssey, (2008).

53. Manas A., Pathak, Bhiksha Raj., " Privacy-Preserving Speaker Verification and Identification Using Gaussian Mixture Models ", IEEE Transactions on Audio, Speech & Language Processing, 21(2): 397-406, (2013).

54. Nemala, Sridhar Krishna., Patil, Kailash., Elhilali, Mounya., " A Multistream Feature Framework Based on Bandpass Modulation Filtering for Robust Speech Recognition ", IEEE Transactions on Audio, Speech & Language Processing, 21(2): 416-426, (2013).

55. McLaren, Mitchell., van Leeuwen, David A., " Source-Normalized LDA for Robust Speaker Recognition Using i-Vectors From Multiple Speech Sources ", IEEE Transactions on Audio, Speech & Language Processing, 20(3): 755-766, (2012).

56. Fauve, B., Matrouf, D., Scheffer, N., Bonastre, J.F., Mason, J., " State-of-the-art performance in text-independent speaker verification through open-source software ", IEEE Transactions on Audio, Speech and Language Processing 15, 1960-1968, (2007).

57. Bimbot, F., Bonastre, J-F, Fredouille, C., Gravier, C., MagrinChagnolleau, I., Meigner, I., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., and Reynolds, D A., " A Tutorial on Text-independent Speaker Veriifcation ", EURASIP Journal on Applied Signal Processing, 10, 430-451, (April 2004).

58. Campbell, W. M., Reynolds, D. A., and Campbell, J. P., " Fusing Discriminative and Generative Methods for Speaker Recognition: Experiments on Switchboard and NFI/TNO Field Data ", In Proc. Odyssey: The Speaker and Language Recognition Workshop in Toledo, Spain, ISCA, 41-44, (31 May - 3 June 2004).