الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي و البحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Master

Filière Électronique
Spécialité Électronique des Systèmes Embarqués

présenté par

Zouaoui Djazia Souheir

&

Slamani Nacira

Système automatique de reconnaissance vocale d'un locuteur basé sur la programmation dynamique (DWT)

Proposé par : Ykhlef Farid

Année Universitaire 2018-2019

Remerciements

On tient à remercier en premier lieu le Dieu Le tout puissant qui nous a accordé la volonté, la santé et le courage d'accomplir ce travail.

Nous remercions Très sincèrement YKHLEF FARID notre encadreur de ce travail, pour son suivi, ces conseils pertinents, ses soutiens moraux pour l'accomplissement de ce travail.

On remercie aussi Messieurs les membres du jury, qui ont accepté de m'honorer en acceptant d'examiner, et d'évaluer notre mémoire.

A toute personne qui a contribué de près ou de loin à l'élaboration de ce travail.

MERCI.

ملخص:

يحتوي التعرف التلقائي على السماعات (RAL) على مجموعة واسعة من التطبيقات في أنظمة الأمان. إنها عملية يتم فيها التعرف على شخص ما بناءً على إشاراته الصوتية. في هذه الرسالة ، يتم تقديم تطبيق نظام التعرف على الكلام على أساس البرمجة الديناميكية (DTW). يعتمد النموذج المرجعي للتعرف على خصائص نوع معامل Cepstral Mel-Frequency MFCC)). يتم عرض نتائج التعرف على الصوت للمتحدث في هذا المستند.

كلمات المفاتيح: DWT، MFCC,التعرف التلقائي على السماعات

Résumé :

La reconnaissance automatique du locuteur (RAL) a une large gamme d'applications dans les systèmes de sécurité. C'est un processus dans lequel une personne est reconnue sur la base de ses signaux vocaux. Dans ce mémoire, la mise en œuvre d'un système de reconnaissance vocale à base de la programmation dynamique (DTW) est présenté. Le modèle de référence pour la reconnaissance est basé sur les caractéristiques de type coefficients Cepstral Mel-Fréquence (MFCC). Des résultats d'identification vocale d'un locuteur sont présentés dans ce document.

Mots clés : Reconnaissance Automatique du Locuteur, MFCC, DWT.

Abstract :

Automatic Speaker Recognition (RAL) has a wide range of applications in security systems. It is a process in which a person is recognized on the basis of his or her voice signals. In this thesis, the implementation of a speech recognition system based on dynamic programming (DTW) is presented. The reference model for recognition is based on Cepstral Mel-Frequency (MFCC) coefficient as features. Voice identification results of a speaker are presented in this document.

Keywords : Automatic Speaker Recognition, MFCC, DWT.

Listes des acronymes et abréviations

RAL	Reconnaissance Automatique du locuteur
VAL	Vérification Automatique du locuteur
IAL	Identification Automatique du locuteur
DTW	Dynamic Time Warping
FFT	Fast Fourier Transform
MFCC	Mel Frequency Cepstrum Coefficient

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

La reconnaissance automatique du locuteur (RAL) a été largement étudiée car La motivation de ces études était d'apprendre comment l'homme reconnait les locuteurs. Le travail le plus important qui a stimulé la recherche sur la reconnaissance du locuteur par la machine a été réalisé par Kersta qui a introduit le spectrogramme (ou il l'a noté comme empreinte vocale) en tant que moyen d'identification personnelle.

Cependant le signal de parole est l'un des signaux les plus complexes à caractériser ce qui rend difficile la tâche d'un système RAL. Cette complexité du signal de parole provient de la combinaison de plusieurs facteurs, la redondance du signal acoustique, la grande variabilité inter et intra-locuteur, les effets de la coarticulation en parole continue et les conditions d'enregistrement. Pour surmonter ces difficultés, de nombreuses méthodes et modèles mathématiques ont été développés, parmi lesquels on peut citer : la comparaison dynamique, les réseaux de neurones, les machines à vecteurs supports (Support Vector Machine SVM), les modèles de Markov stochastiques et en particulier les modèles de Markov cachés (Hidden Markov Models HMM).

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

En ce qui concerne l'extraction des caractéristiques, les coefficients cepstraux incorporant le modèle auditif, connus sous le nom de Coefficients Cepstraux à Fréquence Mel (MFCC) et leurs coefficients dynamiques ont été les caractéristiques ou paramètres dominants.

Dans ce travail de projet de fin d'étude, nous allons étudier un système complet de reconnaissance du locuteur pour des applications de systèmes embarqués.

En plus de l'introduction générale, le manuscrit comprendra trois chapitres et une conclusion générale :

Des généralités sur la reconnaissance vocale sont introduites dans le premier chapitre. Le chapitre deux présentera sur la production de la parole et les outils utilisés pour la reconnaissance du locuteur. En troisième chapitre, quelques résultats et étapes de la reconnaissance vocale sont résumés dans le cadre d'une application d'identification de locuteur. A la fin, une conclusion générale clôturera le travail de ce rapport de master.

Chapitre 1 Généralités sur la reconnaissance vocale du locuteur

1.1 Introduction

La reconnaissance du locuteur est l'un des domaines de l'authentification biométrique, qui fait référence à la reconnaissance de l'identité d'une personne à l'aide de ses caractéristiques fondamentales. En plus de la voix, il existe de nombreux autres modèles physiques et comportementaux pour l'authentification biométrique, par exemple : l'iris, les réseaux veineux de la rétine, les réseaux veineux de la paume de la main, l'empreinte digitale, etc.

En pratique, le choix d'un modèle biométrique approprié devrait tenir compte au moins des considérations suivantes: robustesse, précision (accuracy), accessibilité et acceptabilité.

Par rapport à ces critères de sélection, l'authentification biométrique basée sur la reconnaissance du locuteur est probablement la solution la plus naturelle et la plus économique pour les systèmes de communication homme-machine. Cela se justifie par le faite que la collecte de données du signal de la parole est beaucoup plus pratique que les autres technologies. En plus, la parole est le mode dominant d'échange d'informations entre êtres humains et il tend à être le mode indispensable d'échange d'informations pour les systèmes de communication homme-machine.

Ce chapitre présentera les principaux outils de la reconnaissance automatique du locuteur [1].

Dans cette figure on va présenter les différentes étapes du traitement de la parole et l'utilisation de la reconnaissance du locuteur.

1.2 Reconnaissance vocale automatique du locuteur

La reconnaissance d'un locuteur est l'identification d'une personne à partir des caractéristiques vocales. Le terme « reconnaissance vocale » peut faire référence à « la reconnaissance du locuteur » ou « la reconnaissance de la parole ».

La reconnaissance du locuteur peut être utilisée pour authentifier (ou vérifier) l'identité d'un locuteur dans les systèmes de sécurité. Elle peut aussi simplifier la tâche de traduction d'une parole dans des systèmes à voix spécifiques [2].

1.3 Identification et vérification automatique de locuteur

La reconnaissance automatique du locuteur consiste à l'obtention de renseignements concernant l'identité d'une personne à partir de sa propre voix. Il existe deux principales tâches d'un système de reconnaissance automatique de locuteur, à savoir : vérification et identification.

1.3.1 Vérification Automatique de Locuteur (VAL)

Dans ce type d'application, il s'agit de choisir entre deux hypothèses, soit le locuteur est le locuteur autorisé, c'est-à-dire celui dont l'identité est revendiquée, soit nous nous adressons à un imposteur qui cherche à passer pour un locuteur autorisé [3].

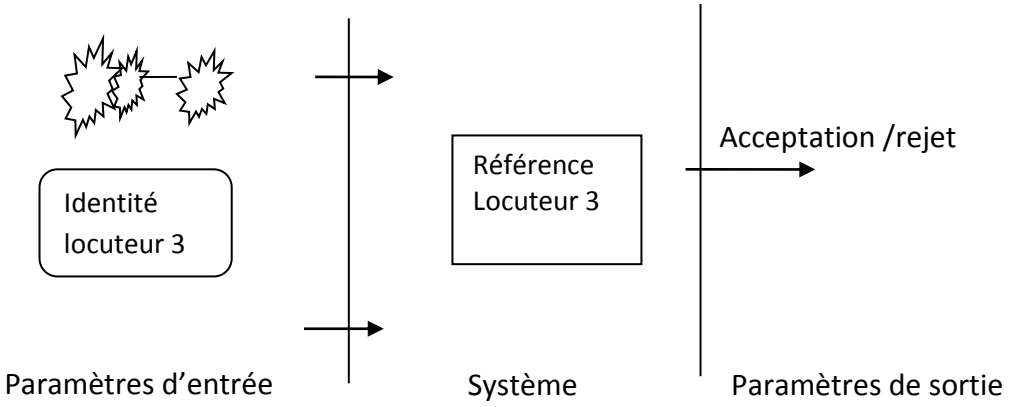


Figure 1.1 Vérification automatique du locuteur.

1.3.2 Identification Automatique de Locuteur (IAL)

L'identification automatique du locuteur (IAL) est le processus permettant de déterminer, parmi une population de locuteurs connus, qui a prononcé un message particulier. Schématique parlant, une séquence de parole est donnée en entrée du système d'IAL. Cette dernière est comparée à une base de données contenant la référence caractéristique du locuteur. L'identité du locuteur dont la référence est la plus proche aux références de la base est donnée en sortie du système d'IAL. Deux modes sont proposés en IAL [4] :

- identification en ensemble fermé pour lequel on suppose que la séquence de parole est effectivement prononcée par un locuteur connu du système.

- et l'identification en ensemble ouvert pour lequel le locuteur peut ne pas être connu.

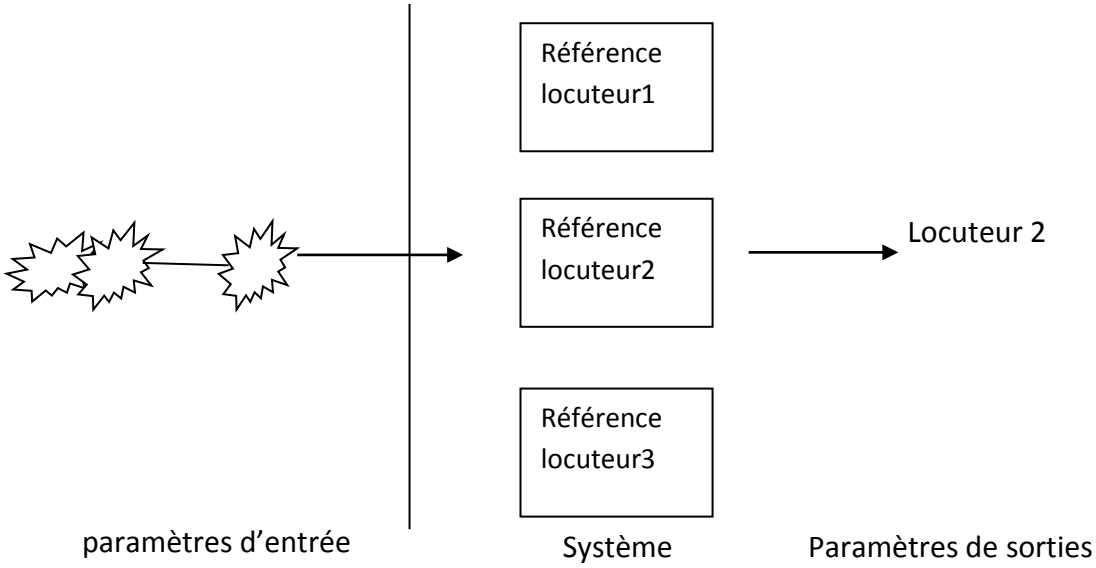


Figure 1.2 Identification automatique du locuteur.

1.4 Paramétrisation

Dans cette étape, les paramètres caractéristiques du signal de parole d'une personne donnée sont calculés. De nombreux travaux de recherche confirment que les paramètres basés sur la représentation spectrale de la parole sont les plus pertinents pour la RAL.

Cette confirmation revient du fait que ces paramètres sont corrélés à la forme du conduit vocal. Ajouter à cela, les paramètres prosodiques qui décrivent le style de la parole du locuteur sont aussi utilisés en pratique [5].

1.4.1 Paramètres de l'analyse spectrale

Les paramètres essentiels de l'analyse spectrale utilisée en RAL sont les coefficients de prédiction linéaire et leurs différentes transformations (LPC, LPCC, etc.) ; ainsi que les coefficients résultant de l'analyse en banc de filtres et de leurs différentes transformations (MFCC, ΔMFCC, etc.). Plusieurs références publiées ont comparées les différentes techniques de paramétrisation. L'objectif de ces travaux était de choisir les meilleurs paramètres représentant de façon efficace qui caractérisent proprement chaque locuteur. Les meilleurs résultats ont été obtenus par la méthode MFCC.

1.4.2 Paramètres prosodiques

Le terme "paramètres prosodiques" d'un signal de parole combine : énergie, durée et fréquence fondamentale (ou pitch). Ces paramètres caractérisent en grande partie

le style de parole d'un locuteur. L'énergie contient des informations liées au niveau acoustique moyen du signal de parole. Ces paramètres sont fragiles dans la pratique et ne permettent pas, en eux-mêmes, de discriminer de manière fiable les locuteurs. De ce fait, ils sont souvent associés aux paramètres de l'analyse spectrale [6].

1.4.3 Paramètres dynamiques

Une information de type dynamique peut être un facteur d'amélioration des performances d'un système d'identification du locuteur. Une première approche, employée pour utiliser cette information au niveau des paramètres, consiste à utiliser une concaténation de plusieurs trames successives de parole (méthodes prédictives). Cependant, cette approche nécessite plus de paramètres dans les modèles et conduit à des problèmes d'estimation des modèles lors de l'apprentissage. La seconde possibilité consiste à calculer les dérivées du premier et du second ordre appelé aussi coefficient de (Δ) ou ($\Delta\Delta$) qui sont désormais très répandue en raison de leur simplicité de mise en œuvre.

1.5 Modélisation

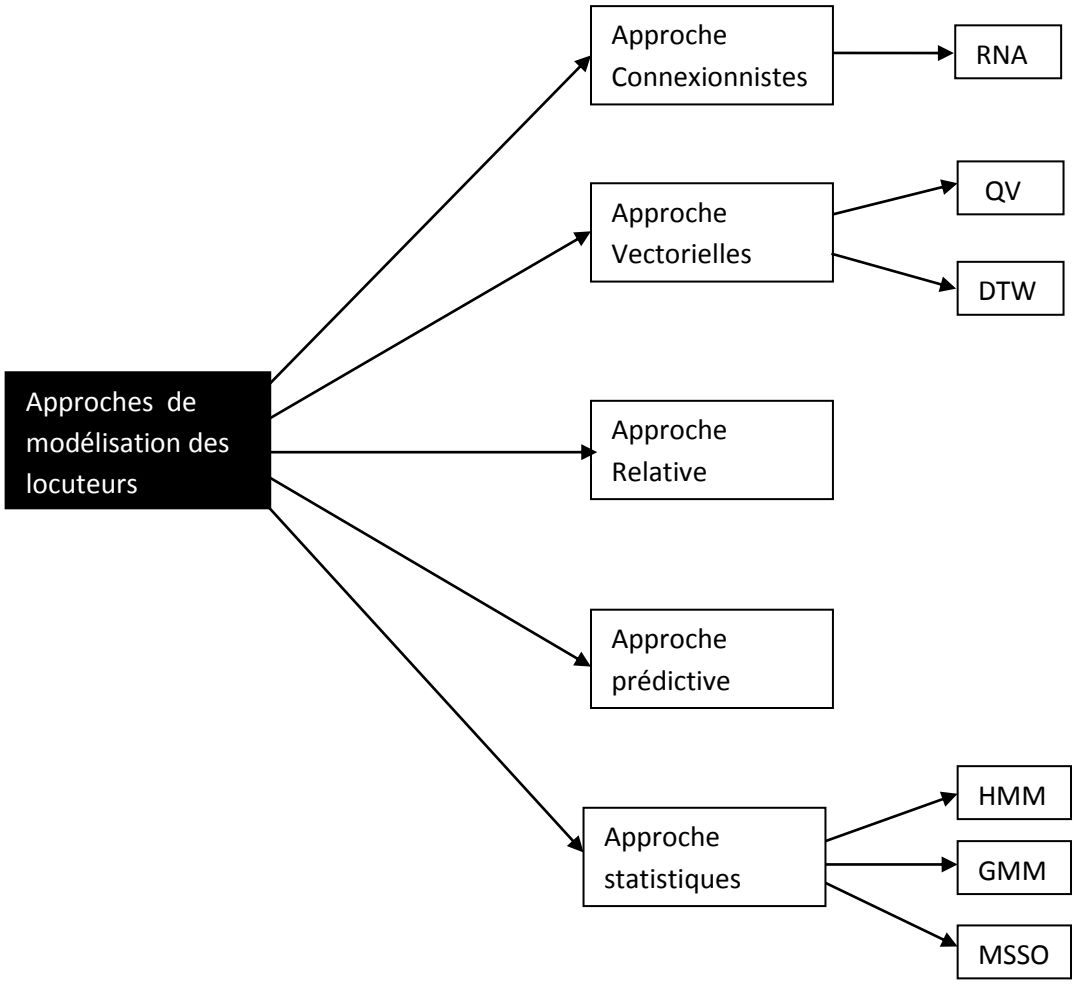


Figure1.3 Différentes approches de la modélisation.

1.5.1 Approche vectorielle

Dans l'approche vectorielle, les vecteurs paramétriques d'apprentissage et de test sont (directement ou indirectement) comparés, en supposant que les vecteurs de l'une des séquences soient une réalisation imparfaite des vecteurs de l'autre séquence. La distorsion entre les deux séquences représente leur degré de similitude. Cette approche consiste en deux techniques principales, l'alignement temporelle (Dynamic Time Warping) DTW et la quantification vectorielle (Vector Quantisation) VQ, qui ont été proposées pour des applications dépendantes et dépendantes du texte respectivement. La DTW aligne temporellement les suites d'observations, tandis que la VQ représente le locuteur par un dictionnaire de codes [7].

a Programmation dynamique

La reconnaissance par l'alignement temporel par programmation dynamique *DTW* (*Dynamique Time Warping)* est basée sur le principe que chaque mot est représenté par une prononciation de référence. Compte tenu des décalages temporels entre les différentes prononciations d'un même mot, l'algorithme met en correspondance des séquences de paramètres par distorsion temporelle (*Time warping)*. La programmation dynamique permet d'aligner temporellement une phrase de test avec une phrase d'apprentissage, ce qui signifie qu'il s'agit d'une technique exclusivement utilisée en mode dépendant du texte [9].

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

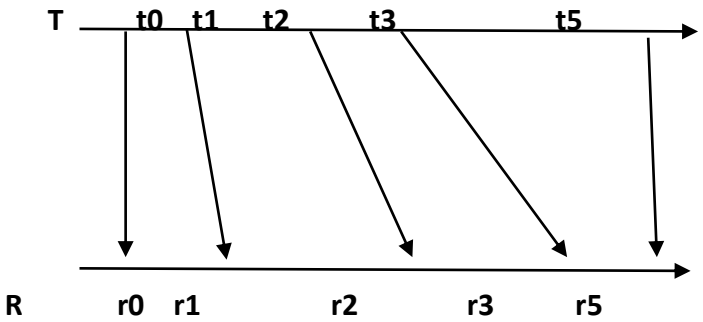


Figure 1.4 Principe de base de la DTW

L'alignement temporel, appelé DTW, est un procédé basé sur le principe de la comparaison d'un signal à analyser avec un ensemble de signaux stockés dans une

base de référence. Le signal à analyser est comparé à chacune des références et est classé en fonction de sa proximité avec l'une des références mémorisées (ou stockées).

La DTW est en réalité une application dans le domaine de la reconnaissance vocale de la méthode plus générale de la programmation dynamique. Cela peut donc être vu comme un problème de cheminement dans un graphe.

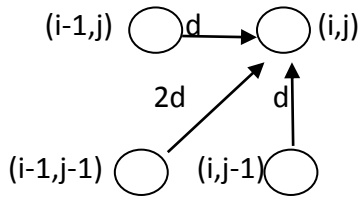
Ce type de méthode peut poser deux problèmes : la taille de la base de référence, qui doit être importante, et la fonction de calcul des distances, qui doit être choisie avec soin. La taille de la base contenant les signaux de référence est directement liée aux capacités et variables de reconnaissance du système d'alignement temporel. [10]

Soient deux images acoustiques (vecteurs MFCC) A et B de longueur I et J respectivement. Tous d'abord faut génère un chemin, les seuls chemins valides arrivants au point $(i\ j)$ sont ceux provenant des points $(i-1, j)$, $(i, j-1)$ et $(i-1, j-1)$. De plus on prend K tel que $C(K) = (I, J)$. On pose $C(1) = (1,1)$.

Cette méthode consiste à choisir le chemin qui passe par les distances $d(i,j)$ les plus petites, de sorte que la distance cumulée le long de ce chemin soit la plus petite possible. On définit ainsi $g(i,j)$ la distance cumulée au point (i,j) comme :

$$g(i,j) = \min \begin{cases} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2.d(i,d) \\ g(i,j-1) + d(i,j) \end{cases}$$

Ensuite, on remplit la matrice $I \times J$ (le plan du chemin) avec en $i^{ème}$ et $j^{ème}$ colonnes le résultat de $g(i,j)$.



Enfin on définit la distance normalisée entre deux prononciations du mot :

$$G = \frac{g(I,J)}{I+J}$$

On obtient une distance entre deux spectres. On effectue ce travail entre le mot à reconnaître et tous les mots stockés auparavant. On prend ensuite le mot de la base de données qui a la plus petite distance spectrale avec le mot à identifier.

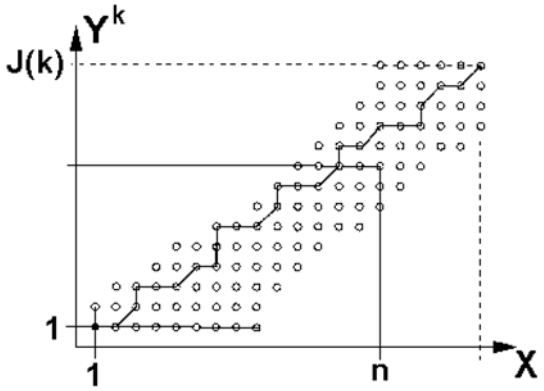


Figure 1.5 Représentation de la notion de chemin entre deux spectres.

On remarque dans la figure ci-haut que les différences entre les deux spectres «tordent». Le chemin idéal sera donc la diagonale.

b Quantification vectorielle

La quantification vectorielle est une méthode non paramétrique qui permet de décrire un ensemble de données à l'aide d'un petit nombre de vecteurs formant un dictionnaire associé aux données. Le dictionnaire est généralement calculé de manière à ce que la distance moyenne entre un vecteur des données et son voisin le plus proche dans le dictionnaire soit aussi petite que possible. La quantification vectorielle est une technique de regroupement d'autant plus appropriée que les données comportent naturellement des "points d'accumulation" autour desquels la densité de vecteurs à partir des données est importante.

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

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

1.5.2 Approche statistique

Cette approche consiste à représenter chaque locuteur par une densité de probabilités dans l'espace des paramètres acoustiques. Elle couvre les méthodes de modélisation utilisant des modèles de Markov cachés, des mélanges gaussiens et des mesures statistiques de second ordre.

a Méthodes statistiques du second ordre

Cette partie présente une famille de mesures de similarité entre les locuteurs. Ces mesures sont basées sur les caractéristiques de second ordre d'une séquence de vecteurs, c'est-à-dire sur le vecteur moyen et la matrice de covariance de cette séquence. Plusieurs mesures de distance ont été utilisées : le rapport de vraisemblance, la distance de Kullbak-Leibler, le maximum de vraisemblance, le test de sphéricité, la déviation absolue des valeurs propres. Ces mesures donnent des résultats très encourageants sur le signal de parole propre, et, naturellement, leurs performances se détériorent sur la parole téléphonique. En raison de leur simplicité relative, ces mesures peuvent également servir de référence pour évaluer la qualité d'une base de données [11].

b Mélange de gaussiennes

La reconnaissance du locuteur par mélanges de gaussiennes (*GMM : Gaussian Mixture Models)* consiste à modéliser un locuteur avec une somme pondérée de composants gaussiens. Ainsi, une large gamme de distribution peut être parfaitement représentée. Chaque composant gaussien est supposé modéliser un ensemble de classes acoustiques. L'utilisation de ce type de modèle semble très encourageante. Il semble modéliser les caractéristiques spectrales des voix des locuteurs et sa mise en œuvre est relativement simple. Nous pouvons assimiler un modèle GMM à un modèle HMM à un seul état, de sorte que nous ne modélisons pas les aspects temporels du signal. Cette méthode est la plus utilisée dans la reconnaissance du locuteur en mode indépendant du texte [12].

c Modèles de Markov cachés

Les modèles de Markov (*ou HMM pour Hidden Markov Models)* ont été initialement introduits dans la reconnaissance de la parole. Ensuite, leur utilisation s'est progressivement étendue au domaine de la reconnaissance du locuteur. Dans cette approche, il ne s'agit plus d'une mesure de distance d'une forme acoustique à une référence, mais de la probabilité que la forme acoustique ait été générée par le modèle de référence du locuteur. Le modèle d'un locuteur consiste en l'association d'une chaîne de Markov, d'une succession d'états à probabilités de transition d'un état à un autre et de lois de probabilités (probabilités d'observation d'un vecteur acoustique dans un état). Les propriétés statistiques des modèles de Markov cachés en font l'un des modèles les plus efficaces actuellement utilisés pour reconnaître le locuteur dépendant du texte. Les *HMM* sont utilisés pour modéliser des processus stochastiques variant dans le temps. Pour cela, ils combinent les propriétés des deux distributions de probabilité et d'une machine à états [13].

1.5.3 Approche prédictive

L'approche prédictive repose sur le principe selon lequel une trame de signal peut être prédite en n'observant que les trames précédentes. Selon ce concept, cette approche est considérée dans la littérature comme une approche dynamique, c'est-à-dire une approche tenant compte des informations dynamiques transmises par le signal de parole. Elle repose principalement sur l'estimation d'une fonction de prédiction, propre à chaque locuteur et apprise sur les signaux d'apprentissage. Pendant la reconnaissance, une erreur de prédiction peut être calculée entre une trame prédite (par la fonction de prédiction) et la trame réellement observée dans la séquence de test [8].

1.5.4 Approche connexionniste

L'approche connexionniste est basée sur la discrimination entre les locuteurs. Elle consiste à fournir à un réseau de neurones et un ensemble de signaux de parole provenant d'une population de locuteurs afin que ce dernier puisse apprendre à distinguer un locuteur des autres. L'approche connexionniste est donc réduite à une tâche de classification [8].

Approches	Avantages	Inconvénients
DWT	• Très rapide. • Présente des performances relativement bonnes.	• Utilisée exclusivement en mode dépendant du texte. • Très sensible à la qualité d'alignement des vecteurs et au choix du point de départ.
QV	• S'applique en mode dépendant ou indépendant du texte.	• Sa rapidité et ses performances dépendent fortement de la taille du dictionnaire.
Approches Connexionnistes	• Bonne performance.	• Complexité d'apprentissage. • L'ajout d'un nouveau client nécessite le réapprentissage de tous les modèles
Approche Prédictive	•L'information dynamique transportée par le signal de parole est prise en considération.	• Les performances obtenues ne sont pas assez suffisantes pour un usage pratique.

Approche Relative	• La modélisation d'un nouveau Locuteur ne se fait plus de façon absolue mais relativement à un ensemble de Locuteurs bien appris.	• Le taux d'identification dépend de la quantité de données d'apprentissage pour la construction des Locuteurs de référence.
HMM	• Prend en considération l'aspect temporel du signal de parole • Excellents résultats en mode dépendant du texte.	• Utilisée uniquement en mode dépendant du texte.
GMM	• Très bonnes performances en mode indépendant du texte.	• Quantité importante de signaux d'apprentissage requise pour une bonne évaluation des paramètres du modèle.
MSSO	•Simplicité de mise en oeuvre • Performante sur de courtes durées.	• Ne capture que les caractéristiques stables le long du signal de parole. • les variations locales ne sont pas prises en compte.

Tableau 1.1 Avantages et inconvénients des approches.

1.6 Décision et mesures de performances

1.6.1 Identification automatique de locuteur

Elle consiste à reconnaître un locuteur parmi un ensemble de locuteurs en comparant son identité vocale à des références connues. La performance du système d'identification est donnée en termes de taux d'identification correct I_c ou incorrectement I_i [14] :

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

et

$$Ii = \frac{\text{Nombre de testes Mal identifiées}}{\text{Nombre totales de tentatives}}$$

Avec

$$I_c + I_i = 100\%$$

1.6.2 Vérification automatique de locuteur

Elle consiste à vérifier l'adéquation du message vocal avec la référence acoustique du locuteur qu'il prétend être. C'est une décision en tout ou rien. Les performances de vérification du locuteur sont exprimées en termes de faux rejets F_R et de fausses acceptations F_A.

Faux rejet : c'est l'erreur commise lorsque le système rejette de manière incorrecte un locuteur légitime (c'est-à-dire une erreur commise lors d'un test de locuteur) :

$$F_R = \frac{\text{Nombre de tentatives d'abonnés rejetés}}{\text{Nombre total de tentatives d'abonnés}}$$

Fausse acceptation : c'est l'erreur commise lorsqu'un imposteur est accepté par inadvertance en tant qu'utilisateur légitime (c'est-à-dire une erreur commise dans un test d'imposteur)

$$F_A = \frac{\text{Nombre de tentatives d'imposteurs acceptés}}{\text{Nombre total de tentatives d'imposteurs}}$$

1.7 Conclusion

Dans ce chapitre nous avons présenté les différents types et les différentes approches de modélisation utilisés dans un système de reconnaissance automatique du locuteur. Nous avons aussi abordé une approche globale de la reconnaissance de la parole. Ainsi, il existe une littérature riche qui présente des études comparatives sur les modélisations utilisées dans ce domaine la reconnaissance de la parole. Dans notre travail, nous nous somme penché sur la DWT. Cet algorithme est capable de comparer deux spectres audio avec des durées et une intensité de la voix différents, et cela d'une manière optimale en recherchant le meilleur chemin pour passer d'un spectre à l'autre. Cependant, d'autres méthodes existent, telles que les modèles de Markov cachés (HMM), par exemple, qui sont beaucoup plus puissants que l'algorithme DTW mais beaucoup plus complexes.

Chapitre 2 Production de la parole et outils de reconnaissance du locuteur

2.1 Introduction

Le signal de parole est un signal continu, d'énergie finie, non stationnaire, qui contient des composantes complexes et variables en fonction des sons émis. Les informations contenues dans ce signal sont redondantes et codent à la fois le message utile et les caractéristiques du locuteur qui a envoyé le message. Le traitement automatique de la parole nécessite une paramétrisation de ce signal vocal afin d'obtenir une représentation plus concise et moins redondante, et d'autre part d'extraire des paramètres spécifiques au signal de parole de l'application choisie : reconnaissance de la parole, du locuteur ou de la langue, transmission ou synthèse de la parole.

Pour cela, les méthodes d'analyse du signal vocal sont basées sur les techniques traditionnelles de traitement du signal, telles que la transformée de Fourier ou des transformations aux laquelle s'ajoutent des méthodes plus élaborées tenant compte des modèles de la production ou de la perception de la parole. Dans ce chapitre, nous introduisons certaines de ces méthodes d'analyse conduisant à l'extraction des paramètres (caractéristiques).

2.2 Production de la parole

La production de la parole implique différents organes. La source de la parole provient des poumons qui émettent un flux d'air. Ce flux d'air traversera le ***larynx*** pour faire vibrer ou non les **cordes vocales**. Il traversera ensuite le ***conduit vocal*** (cavité nasale et buccale) et les articulateurs tels que les ***lèvres*** et la ***langue*** (Figure 2.1). Le contenu fréquentiel du signal acoustique de parole produit par un locuteur dépend fortement des caractéristiques morphologiques de son appareil de parole. Celui-ci peut être divisé en quatre parties: le générateur, le vibreur, le résonateur et les modulateurs.

- **Générateur :** l'air expulsé des poumons passe à travers l'appareil vocal en tant qu'instrument à vent et crée la pression nécessaire pour générer un signal acoustique.
- **Vibrateur :** l'air expulsé des poumons traverse la trachée pour arriver dans le larynx où se trouvent les cordes vocales. Les cordes vocales sont une paire de muscles dont la longueur moyenne se situe entre 20 et 25 millimètres. Cette longueur varie cependant d'un individu à l'autre. L'air traversant le larynx met en vibration les cordes vocales. La fréquence de vibration des cordes vocales est modulée en fonction de leur degré de contraction. Le locuteur peut ainsi moduler la hauteur des sons qu'il émet.
- **Résonateur** : Les vibrations des cordes vocales sont modifiées par le passage de l'air dans les différentes cavités qui composent le pharynx et dans les fosses nasales, la bouche et le larynx avec lesquels il communique. Ces résonateurs influent sur le son en atténuant certaines fréquences et en amplifiant d'autres. La forme et le volume de ces cavités, spécifiques au locuteur, modifient fortement le son produit.
- **Modulateurs** : Enfin les organes modulateurs que sont la langue, les lèvres et la mâchoire sculptent le son pour produire les phonèmes qui composent la parole. La position de ces différents organes est le mécanisme final qui permet la production de parole articulée [15].

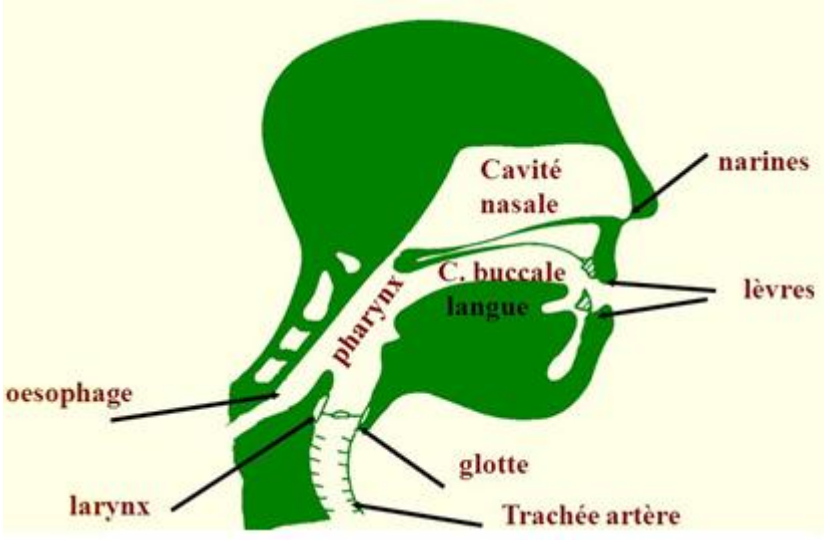


***Figure 2.1** Vue schématique de l'appareil vocal.*

2.2.1 Mécanisme de production de la parole

La production de la parole est un processus de nature linguistique (message à transmettre) qui évolue vers une exécution motrice (séquence de contractions musculaires) impliquant plusieurs composantes de l'anatomie humaine et aboutissant à un signal de parole. Ce processus peut être décomposé en trois étapes:

1. Conceptualisation (ou préparation conceptuelle) : à cette étape, l'intention de produire la parole génère les concepts souhaités correspondant au message à transmettre.

2. Formulation : la forme linguistique requise pour l'expression du message souhaité est produite. La formulation comprend le codage grammatical (mots choisis et forme syntaxique appropriée), le codage morpho phonologique (division des mots en syllabes), la syllabification et le codage phonétique.

3. Articulation et l'exécution motrice de la parole : elle consiste en l'exécution de la séquence articulatoire correspondant au message. Dans cette étape, le locuteur exécute une série de signaux neuromusculaires qui servent de commandes et contrôlent les cordes vocales, les lèvres, la mâchoire, la langue et le vélum (voile du palais), produisant ainsi la séquence sonore souhaitée à la sortie [15].

2.2.2 Propriétés acoustiques du conduit vocal

a Continuité

La production d'un son est fortement influencée par les sons qui le précédent et le suivie à cause de l'anticipation du geste articulatoire. L'identification correcte d'un segment de parole isolé de son contexte est parfois impossible. Certainement, il est plus facile de reconnaître des mots isolés bien séparés par des périodes de silence que de reconnaître la séquence de mots constituant une phrase. En fait, dans ce dernier cas, non seulement la frontière entre les mots n'est plus connue, mais, en outre, les mots deviennent fortement articulés [16].

b Variabilité de la parole

Les informations transmises par le signal de parole sont multiples. La variabilité du signal de parole entre les locuteurs est principalement utilisée dans la reconnaissance du locuteur pour reconnaître les individus. C'est la variabilité inter-locuteurs. La capacité des systèmes RAL à identifier une personne repose particulièrement sur la capacité de discriminer des personnes en fonction de cette variabilité. Mais d'autres facteurs de variation modifient la parole. Le signal de parole

est par exemple considéré comme non reproductible par son locuteur. Il existe une variabilité spécifique au locuteur en fonction de son état physique mais également psychologique. C'est la variabilité intra-locuteur. De plus, les conditions environnementales influencent l'onde acoustique du signal de parole. Les bruits ambiants additifs ou de convolution causés par l'enregistrement sonore modifient aussi le signal de parole [16].

c Redondance

La redondance naturelle du signal de parole permet de réduire très fortement le débit binaire dans une très large mesure, au prix d'un traitement plus ou moins complexes et du risque d'une certaine dégradation de la qualité de la représentation [16].

2.3 Pré-traitement du signal de parole

2.3.1 Conversion Analogique/Numérique

Les traitements effectués sur le signal de parole sont aujourd'hui réalisés dans le domaine numérique. Au-dessus de 8 kHz, les informations vocales sont négligeables, la bande de fréquence généralement utilisée est [0-8] kHz. Un échantillonnage du signal de parole à 16 kHz convient pour conserver la quasi-totalité des informations (théorème d'échantillonnage de Shannon). L'amplitude est quantifiée, généralement sur 16 bits, afin d'obtenir une bonne qualité. Pour le codage à faible débit, l'échantillonnage est réalisé à 8 kHz, ce qui permet de conserver la bande téléphonique (300-3400Hz). Le signal est représenté dans le domaine fréquentiel par l'utilisation de transformées de Fourier ou sous une forme permettant de regrouper les informations temporelles et fréquentielles: le spectrogramme (figure 2.2) [10].

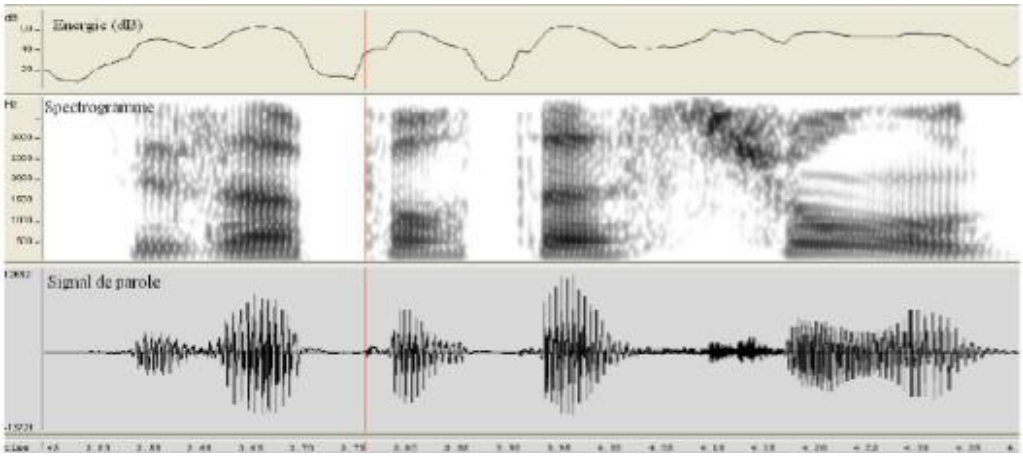


Figure 2.2.Représentation d'un signal de parole, de son spectrogramme et de son énergie.

Du fait que le signal de parole est quasi-stationnaire sur de courtes périodes, il est souvent analysé sur des trames découpées et pondérées par des fenêtres de 20 à 30 ms avec un taux de recouvrement (chevauchement) de 50% à 75%, puis représentées dans le domaine spectral (Figure 2.3.b). Dans le cas d'un signal échantillonné à 8 kHz, une fenêtre d'analyse en 256 points correspond à une longueur de 32 ms (Figure 2.3.a). Le plus souvent, une fenêtre de Hamming est utilisée.

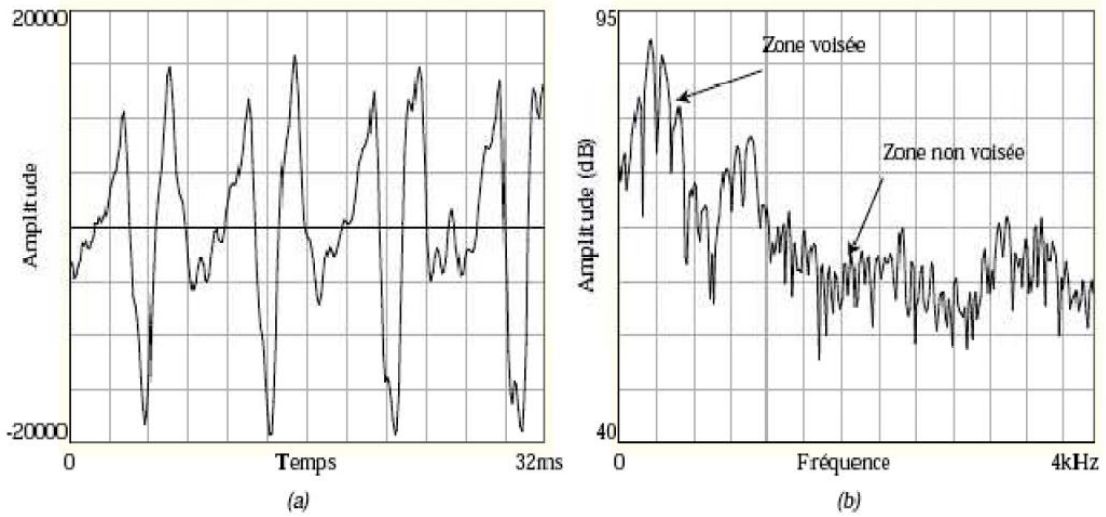


Figure 2.3 Représentation temporelle (a) et spectrale (b) d'un signal de parole voisé et non voisé.

2.3.2 Traitement court-terme du signal vocal

a Segmentation et tramage

Les méthodes de traitement du signal utilisées dans l'analyse du signal vocal fonctionnent sur des signaux stationnaires, tandis que le signal vocal est un signal non stationnaire. Pour remédier à ce problème, l'analyse de ce signal est réalisée sur des trames de parole successives, de durée relativement courte, sur lesquelles le signal peut en général être considéré comme quasi stationnaire. Dans cette étape de segmentation, le signal pré-accentué est ainsi découpé en trames de N échantillons du signal de parole. En général, N est défini de telle sorte que chaque trame correspond à environ 20 à 30 ms. Deux trames successives sont séparées par M échantillons correspondant à une période de l'ordre du centi-seconde.

b Fenêtrage

La segmentation du signal en trames produit des discontinuités aux limites des trames. Dans le domaine spectral, ces discontinuités se manifestent par les lobes secondaires. Ces effets sont réduits en multipliant les échantillons de la trame par une fenêtre de pondération telle que la fenêtre de Hamming. Ces fenêtres d'analyse sont juxtaposées de manière à permettre l'apparition d'un chevauchement. Ainsi, dans la

phase de récupération du signal, la nécessité de lisser les signaux synthétisés doit être prise en compte [17] [19].

c Energie court-terme

L'un des outils permettant de fournir une représentation fidèle des variations de l'amplitude du signal vocal $x\ (n)$ dans le temps est l'énergie à court terme (figure 2.4). En général, il est défini par la formule :

$$\mathrm{E}(n) = \sum_{m=-\infty}^{\infty} [X(m)w(n-m)]^2$$

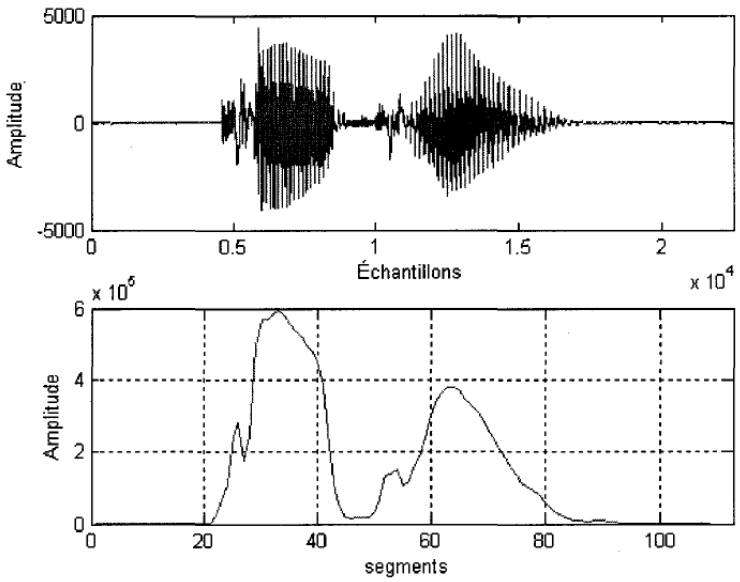


Figure 2.4 Énergie à court terme d'un signal vocal.

En absence de bruit de mesure, l'énergie est un outil efficace pour séparer la parole du silence [18].

d Transformée de Fourier court-terme

La transformée de Fourier à court terme est obtenue en extrayant de l'audiogramme une trentaine de ms du signal vocal, en multipliant ces échantillons par une fenêtre de pondération (souvent une fenêtre de Hamming) et en effectuant une transformation de Fourier sur ces échantillons [18].

e Pré-accentuation

La préaccentuation est un processus par lequel les hautes fréquences sont amplifiées pour leur donner une amplitude supérieure à celle du bruit. La

préaccentuation du signal qui consiste en un passage dans un filtre de transmittance, $(1 - \mu\, z^{-1})$ avec μ compris entre 0,9 et 1 (0,95 pour notre cas). Son but est d'accentuer la haute fréquence du spectre (Figure 2.5) [19].

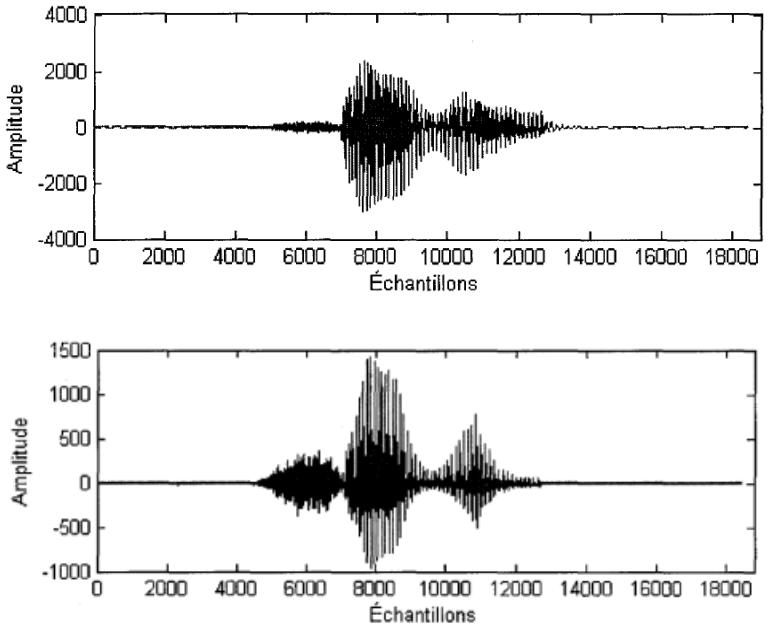


Figure 2.5 Augmentation des hautes fréquences

Remarque : Le rôle principal de la préaccentuation consiste à augmenter les hautes fréquences pour se caler sur notre perception des aiguës. Il serait dommage de négliger ces zones qui contiennent beaucoup d'informations (c'est là que se trouve l'énergie des fricatives).

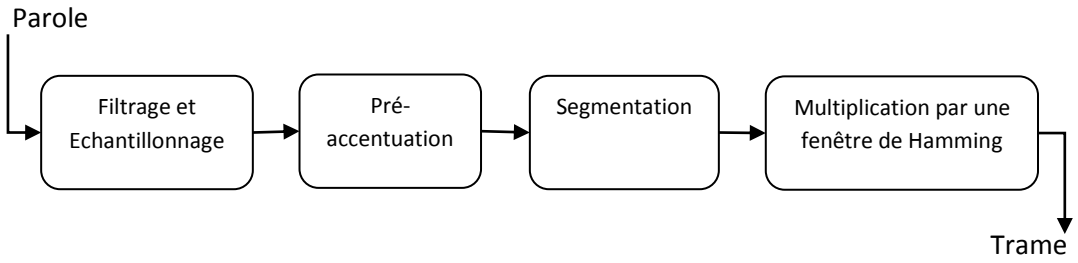


Figure 2.6 Chaîne de prétraitement du signal parole

2.3.3 Détection d'activité vocale VAD

La détection d'activité vocale est généralement définie comme le processus de distinction entre les parties contenant la voix et les parties de celles qui n'en contiennent pas. L'acronyme VAD signifie " Voice Activity Detector" en anglais ou Détecteur d'activité vocale". En principe, le VAD émet un signal binaire ('0': absence de

la voix et '1': présence de la voix) lors de la détection de la voix à partir d'un signal appliqué à son entrée. Le VAD s'applique dans de nombreux systèmes de communication vocale actuels. Ceux-ci incluent la téléphonie mobile, la téléphonie Internet, les systèmes de communication sans fil et la reconnaissance vocale. C'est un problème non trivial qui explique l'existence d'une multitude de méthodes pour développer un VAD. Chaque méthode a ses propres qualités et limites, et les recherches dans ce domaine fournissent régulièrement des résultats plus nombreux et meilleurs [18].

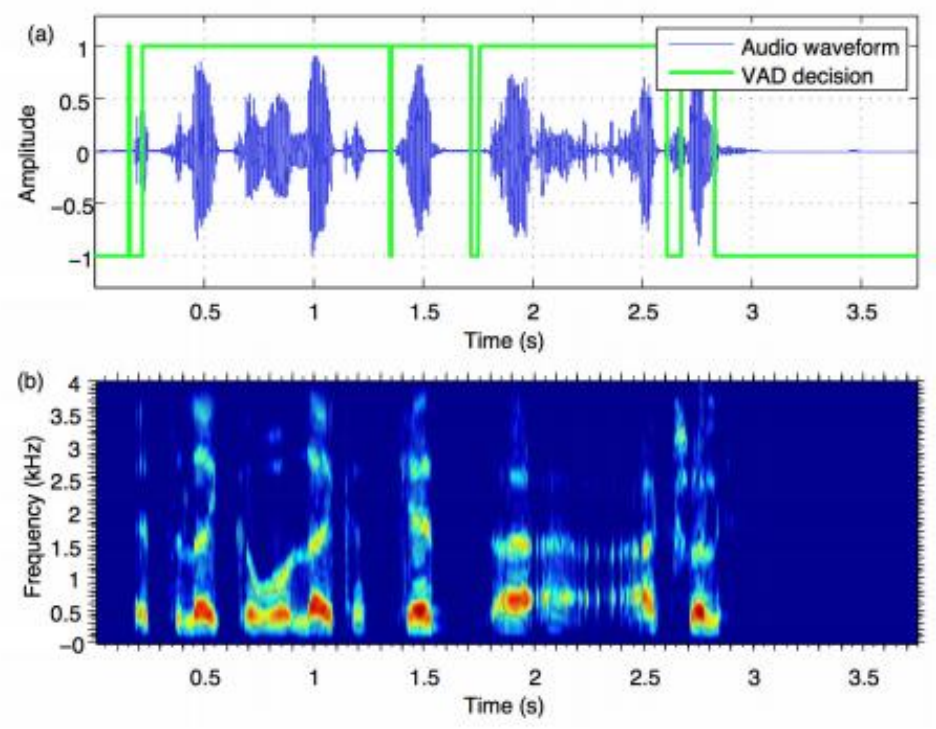


Figure 2.7 La forme d'onde d'un signal de parole avec la VAD superposée. Une valeur de 1 et -1 indique respectivement la parole et la non-parole (silence).

2.4 Extraction des paramètres d'un signal vocal

La résolution de la complexité de la parole est possible par le calcul de coefficients représentatifs du signal traité. Ces coefficients sont calculés dans des intervalles de temps réguliers. Pour cela, le signal de parole est transformé en une série de vecteurs contenant des coefficients. Ces derniers devraient représenter au mieux le signal qu'ils sont censés le modéliser et extraire le maximum d'informations utiles pour la reconnaissance. Un système de paramétrisation du signal vocal est divisé en deux blocs, le premier formatage et l'autre calcul de coefficient.

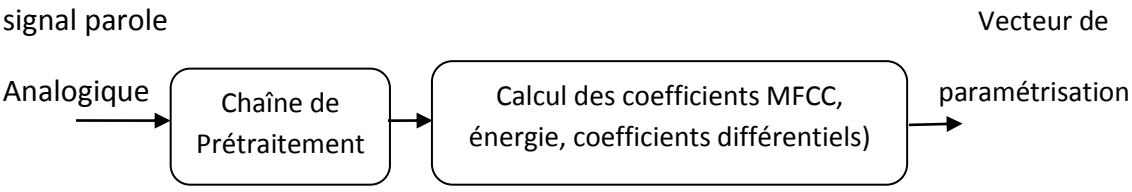


Figure 2.8 Phase de paramétrisation acoustique.

2.4.1 Analyse spectrale

L'analyse spectrale est le passage du domaine temporel au domaine spectral. Une estimation du spectre de chaque trame est calculée à l'aide de la transformée de Fourier discrète (DFT).

2.4.2 Traitement par banc de filtres

Les techniques de filtrage sont des transformations permettant d'obtenir une nouvelle représentation du signal dans un nouvel espace dans lequel le traitement est effectué. Cela revient à multiplier le signal d'origine dans ce nouvel espace par une fonction de transfert. La transformation inverse permet d'observer le résultat de l'opération. Les filtres permettent de sélectionner des fréquences spécifiques [10].

Un banc de filtres est un ensemble de filtres conçus pour diviser le spectre d'un signal en bandes de fréquences, appelées "bandes critiques". Ces bandes se chevauchent et dont les fréquences centrales ont la plus grande amplitude. Chaque bande critique correspond à la séparation de fréquence nécessaire pour que deux harmoniques soient discriminés. Cette analyse est basée sur le système de perception humaine. Leur mise en scène de fréquence imite la distribution et la forme des filtres cochlée.

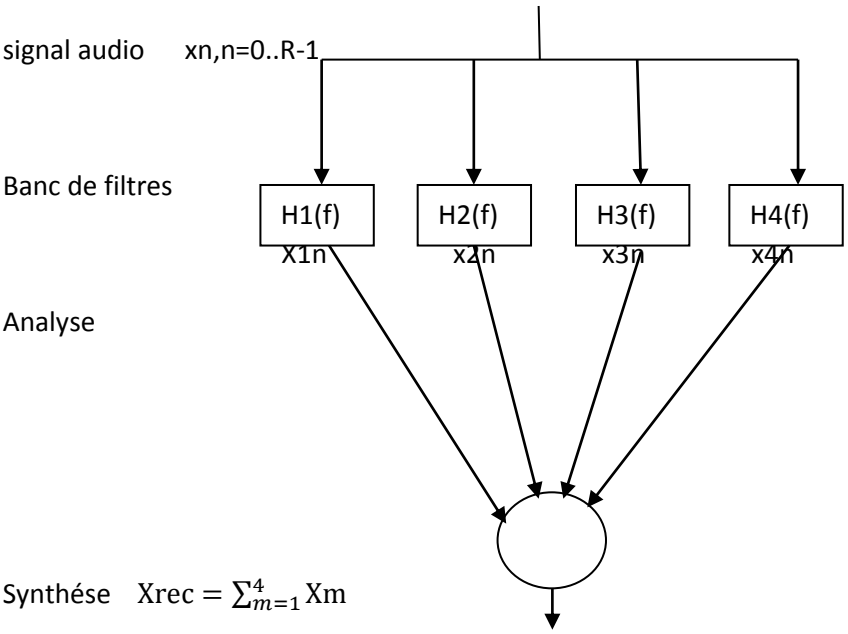


Figure 2.9 Schéma d'un banc de filtre.

La distribution de fréquence des filtres est différente selon les échelles choisies, soit linéaire, soit logarithmique. Il existe plusieurs implémentations de bancs de filtres telles que l'échelle de MEL ou l'échelle de Bark.

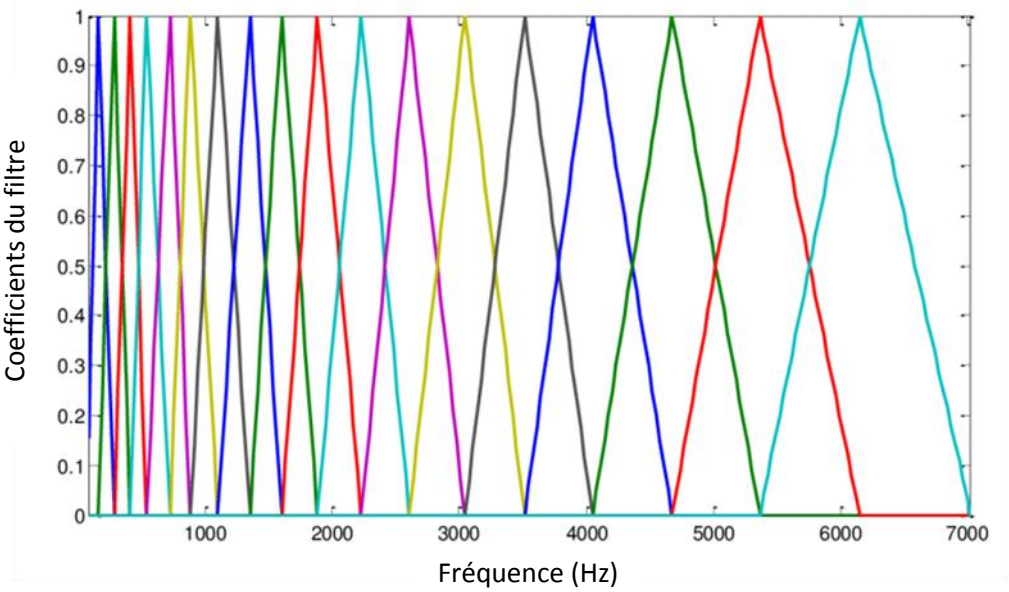


Figure 2.10 Implémentation de bancs de filtres selon l'échelle de MEL.

2.4.3 Transformée de Fourier discrète inverse

La transformée de Fourier est une généralisation de la série de Fourier appliquée aux signaux non périodiques. Soit $x(t)$ un signal non périodique défini et intégrable dans l'intervalle $(-T/2\,,+T/2)$: la transformée de Fourier de la fonction $x(t)$ dénotée TF est donné par l'équation suivante [17] :

$$X(f) = \int_{-\infty}^{+\infty} x(t)\exp(-j\omega t)\,dt \qquad (1)$$

La transformée de Fourier inverse est donnée par :

$$x(t) = \int_{-\infty}^{+\infty} X(f)\exp(j\omega t)\,df \qquad (2)$$

$\omega = 2.\pi.f$: pulsation.

La transformée de Fourier est en générale une fonction complexe pouvant se mettre sous la forme suivante.

$$F(f) = Re\{F(f)\} + jIm\{X(f)\} \qquad (3)$$

Re : partie réelle

Im : partie imaginaire

Aussi, elle révèle une autre écriture :

$$X(f) = |X(f)| \exp\big(j\theta(f)\big)$$

$$|F(f)| = \sqrt{(Re\{F(f)\})^2 + (Im\{X(f)\})^2}$$

$$\theta(f) = \text{artg}\left(\frac{Im\{X(f)\}}{Re\{F(f)\}}\right) \qquad (4)$$

2.4.4 Cepstre

Le cepstre d'un signal $x(t)$ est une transformation de ce signal du domaine temporel en un autre domaine analogue au domaine temporel. Pour rappeler le fait que l'on effectue une transformation inverse à partir du domaine fréquentiel, les dénominations des notions sont des anagrammes de celles utilisées en fréquence. Ainsi, le spectre devient le cepstre, la fréquence une quéfrence, un filtrage un liftrage [21].

$$\mathrm{C}(T) = \mathrm{C}\big(x(t)\big) = TF^{-1}\big(\ln\big(\big|TF\big(x(t)\big)\big|\big)\big)$$

2.4.5 Coefficients Cepstraux

Le signal vocal résulte de la convolution de la source par le conduit vocal. Dans le domaine spectral, cette convolution devient un produit rendant difficile la séparation de la contribution de la source et de celle du conduit. Ce problème peut être résolu par l'analyse cepstrale par passage dans le domaine log-spectral. En pratique, le cepstre réel d'un signal numérique $x(n)$ estimé sur une fenêtre d'analyse à N échantillons est obtenu comme suit:

x(n) $\xrightarrow{\text{FFT}}$ X(f) $\xrightarrow{\log|.|}$ $\log|X(f)|$ $\xrightarrow{FFT^{-1}}$ Cepstre

Les coefficients cepstraux sont donnés par :

$$c(n) = \frac{1}{N}\sum_{i=0} \log(|X(i)|)\, e^{\frac{2ijn}{N}} \qquad \text{pour n=0,1,.....N-1}$$

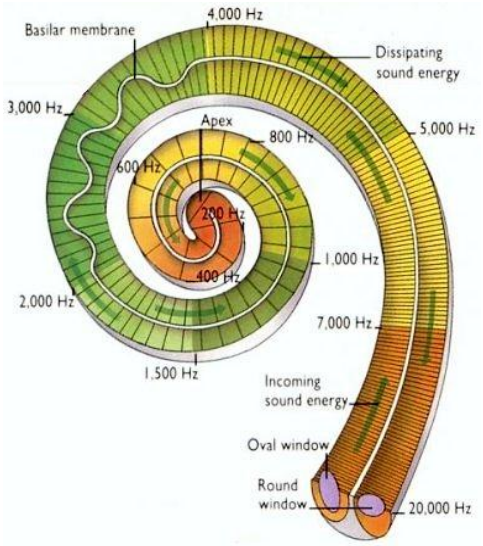
Ce type de calcul des coefficients cepstraux n'est pas utilisé dans la reconnaissance vocale en raison du calcul important de la FFT et de la FFT inverse. D'autre part, les coefficients cepstraux utilisés peuvent être obtenus à partir des coefficients de la prédiction linéaire ou des énergies d'un banc de filtres. Ainsi, les paramètres LPCC (Linear Prediction Cepstral Coefficients) sont calculés à partir d'une analyse par prédiction linéaire.

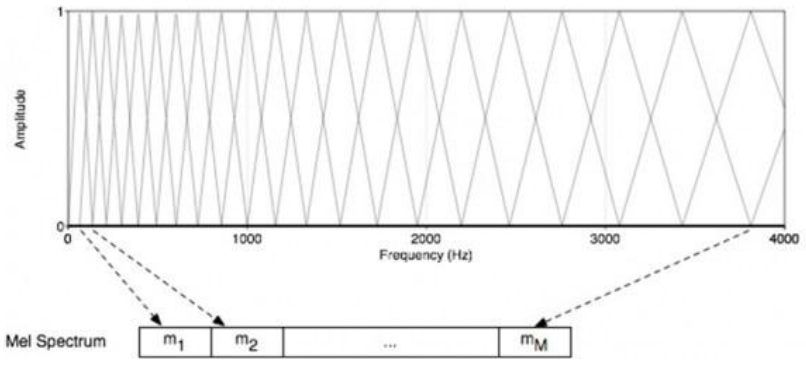
Les coefficients MFCC (Mel Frequency Cepstral Coefficients) sont les paramètres les plus couramment utilisés dans les systèmes de reconnaissance vocale. L'analyse MFCC consiste à exploiter les propriétés du système auditif humain en transformant l'échelle

linéaire des fréquences en l'échelle de Mel. Cette dernière échelle est codée par un banc de 15 à 24 filtres triangulaires espacés linéairement jusqu'à 1 kHz, puis logarithmiquement jusqu'aux fréquences maximales.

2.4.6 Bancs de filtres Mels

Il s'agit d'aborder le fonctionnement de l'oreille humaine. Nous n'entendons pas de la même manière les fréquences contiguës selon qu'elles sont graves ou aiguës. Les basses fréquences (graves) sont beaucoup mieux distinguées par la cochlée qui agit comme une série de filtres contigus de différentes largeurs. Nous devons simplement faire la même chose avec notre spectre. Pour cela, nous allons appliquer un banc de filtres dont les largeurs de bande et l'espacement correspondent à l'échelle de Mel [22]. En calculant la somme des énergies contenues dans chaque filtre nous obtenons autant de valeurs représentatives que de filtres.





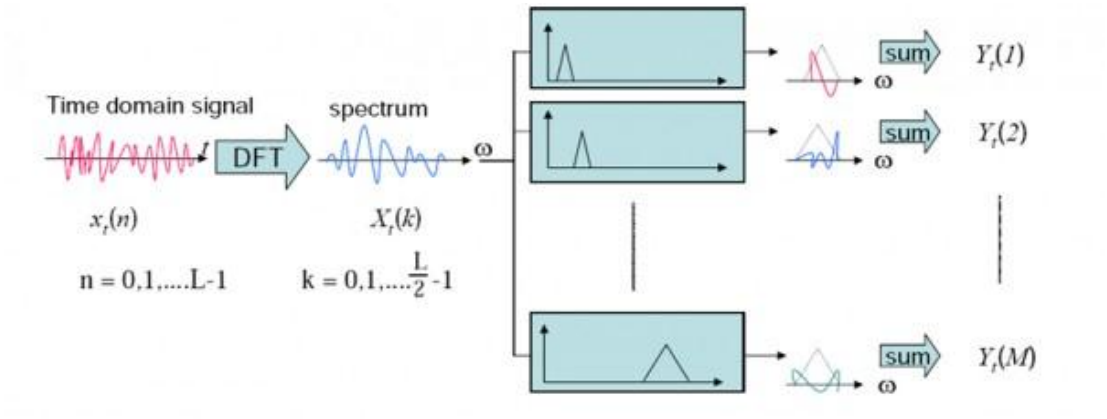


Figure 2.11 Cochlée (haut), échelle de MEL (milieux) et Implémentation du banc de filtres (bas).

2.4.7 Extraction des paramètres par les MFCCs

L'extraction des paramètres basée sur les MFCC est une méthode d'extraction de paramètres en fonction de l'échelle de la MEL. En effet, la perception de la parole par le système auditif humain repose sur une échelle de fréquence similaire à celle de la MEL. Cette échelle est linéaire aux basses fréquences et logarithmique aux hautes fréquences et est donnée selon l'équation suivante:

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

f représente la fréquence.

Après une préaccentuation et une subdivision de différents segments avec un signal à libération conditionnelle, on applique la méthode MFCC qui consiste à calculer la transformation de Fourier de chaque segment. Ensuite, utiliser les filtres triangulaires, espacés suivant l'échelle de MEL pour filtrer cette transformation et obtenir les énergies à partir des modules carrés de la transformée de Fourier [23].

Enfin, le calcul de la transformée en cosinus discrète (DCT) des logarithmes des énergies obtenues afin d'extraire les coefficients MFCC utilisés pour la reconnaissance. Ces coefficients sont donnés par l'équation suivante :

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} \log_{10}(E_j) \cos\left(\frac{\pi i}{N}(j + 0.5)\right)$$

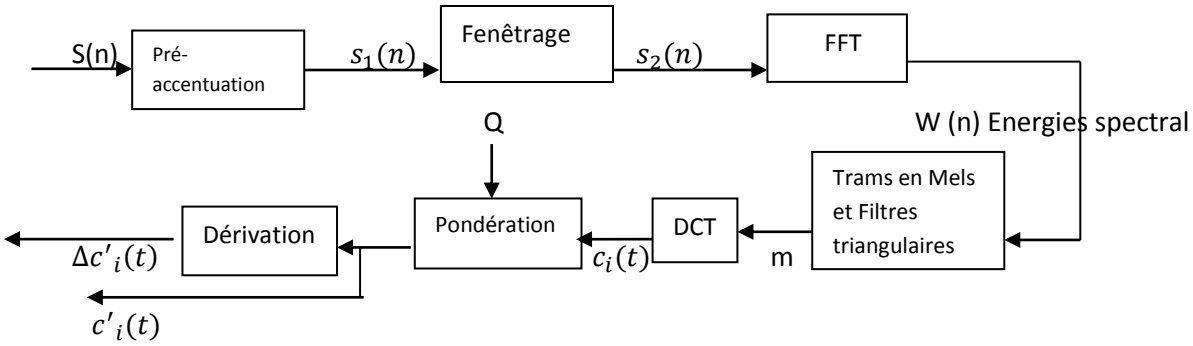


Figure 2.12 Étapes de calcul des coefficients MFCC.

2.5 Conclusion

Dans ce chapitre, nous avons présenté quelques outils de base du traitement du signal vocal. L'analyse à court terme a fait objet d'étude dans ce chapitre et cela en référence à l'utilisation de segments de parole de courte durée durant lesquels la parole est quasi stationnaire. La parole est un signal redondant, ce qui lui confère une meilleure résistance au bruit. Cependant, les informations qu'il véhicule ne sont pas toutes pertinentes pour la reconnaissance vocale, donc dans la pratique pour réduire le nombre de données à traiter, il est préférable d'extraire que les paramètres essentiels. Dans cette optique, les coefficients MFCC sont les paramètres les plus utilisés dans les systèmes RAL. Ces coefficients sont généralement utilisés avec leurs paramètres dynamiques pour améliorer les performances. Néanmoins, l'ajout de paramètres nécessite, dans certaines applications RAL, de réduire le nombre de paramètres acoustiques en sélectionnant les plus pertinents.

Chapitre 3 Résultats

3.1 Introduction

La reconnaissance automatique du locuteur appartient à un sujet beaucoup plus vaste dans le domaine de la reconnaissance des formes. Le but de la reconnaissance de formes est de classer les objets d'intérêt dans des catégories ou classes. Les objets d'intérêt sont appelés de manière générique des caractéristiques et, dans notre cas, sont des séquences de vecteurs acoustiques extraites de la parole en entrée en utilisant les techniques décrites dans le chapitre précédent. Les classes ici se réfèrent à des locuteurs individuels.

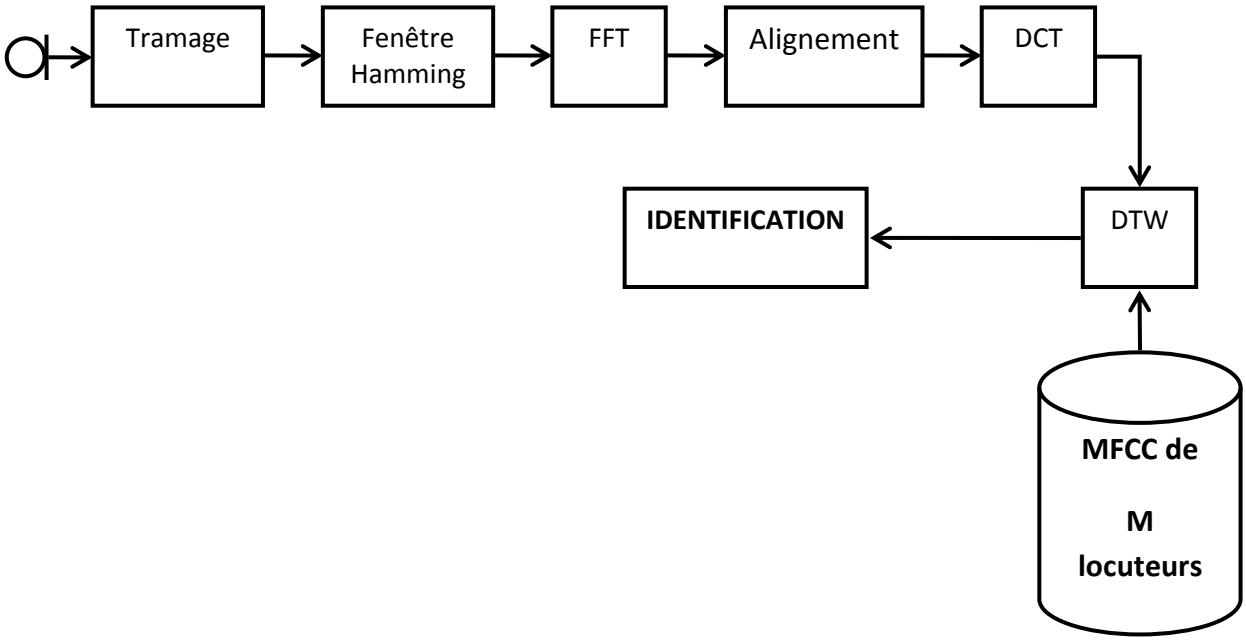
Il existe différentes manières d'extraire des caractéristiques d'un signal vocal. Les propriétés les plus importantes des méthodes d'extraction de caractéristiques utiles devraient être :

- Variation élevée entre les locuteurs
- Faible variation intra-locuteur
- Facile à mesurer
- Robuste contre la distorsion et le bruit
- Caractéristiques totalement indépendantes en interne

L'un des moyens les plus populaires et les plus efficaces consiste à utiliser les MFCC.

Dans ce chapitre, on s'intéresse à l'application d'identification du locuteur.

3.2 Processus d'identification



3.2.1 Extraction de caractéristiques

Après acquisition du signal vocal, le processus d'extraction de caractéristiques par MFCC comprendra cinq étapes principales, exécutées dans l'ordre suivant :

a Tramage

La première étape est le tramage. Le signal de parole est divisé en trames d'une longueur de 10 à 30 millisecondes. La longueur de trame est importante car si elle est trop longue, elle ne pourra pas capturer les propriétés spectrales locales et si elle est trop courte, la résolution de fréquence se dégradera. Les trames se chevauchent généralement de 25% à 75% de leur propre longueur. Le chevauchement permet de s'assurer que chaque son de la parole est approximativement centré sur une trame.

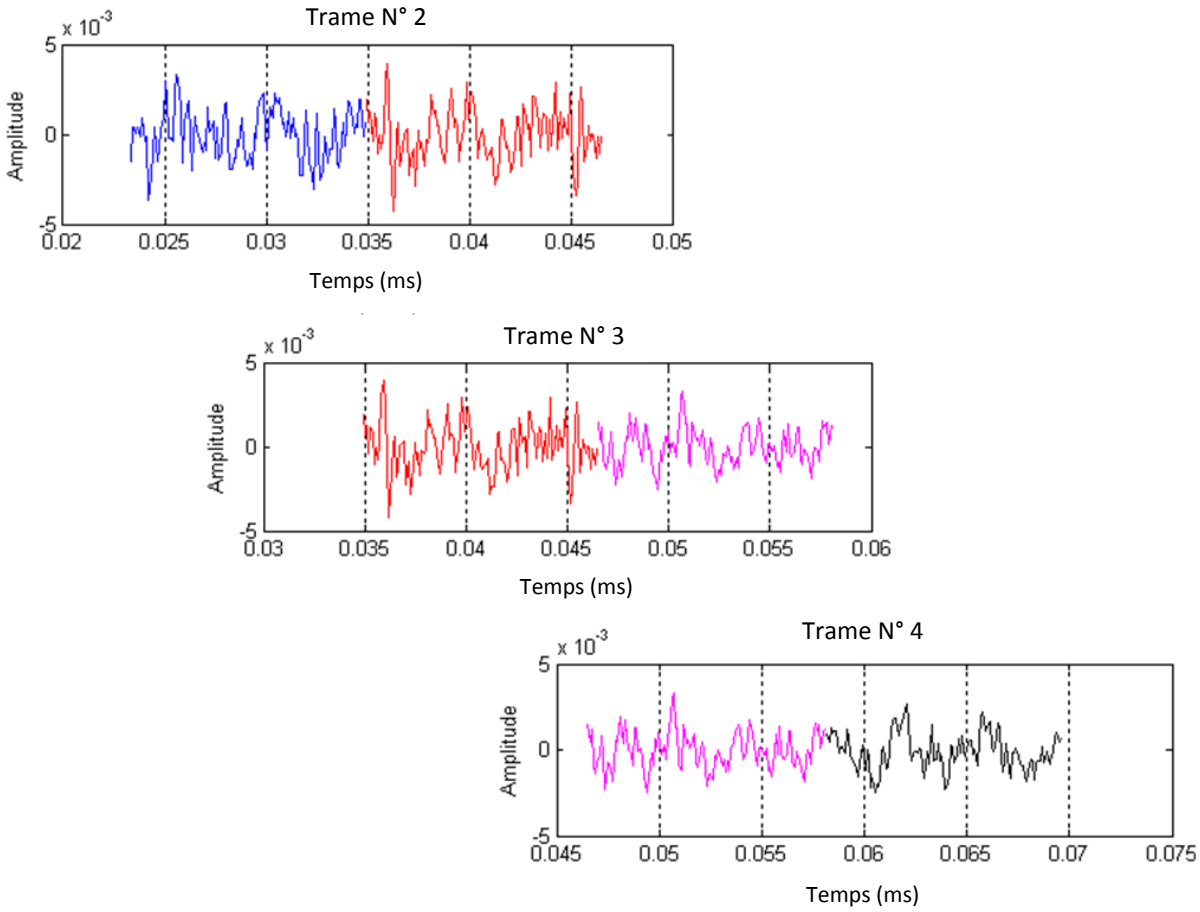


Figure 3.1 Chevauchement de trames. Longueur de trame de 20 ms avec un chevauchement de 50%.

b Fenêtrage

Une fois le signal découpé en trames, chaque trame est multipliée par une fonction fenêtre. Une réduction progressive des bords est souhaitée pour minimiser les discontinuités. La fenêtre la plus commune et utilisée dans le traitement de la parole est la fenêtre de Hamming. La fenêtre de Hamming est définie comme :

$$w[n] = \begin{cases} 0.54 - 0.46.\cos\left(\frac{2\pi n}{N}\right), 0 \leq n \leq N-1 \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad sinon \end{cases}$$

La forme d'onde et la réponse en fréquence sont illustrées dans la figure ci-dessous.

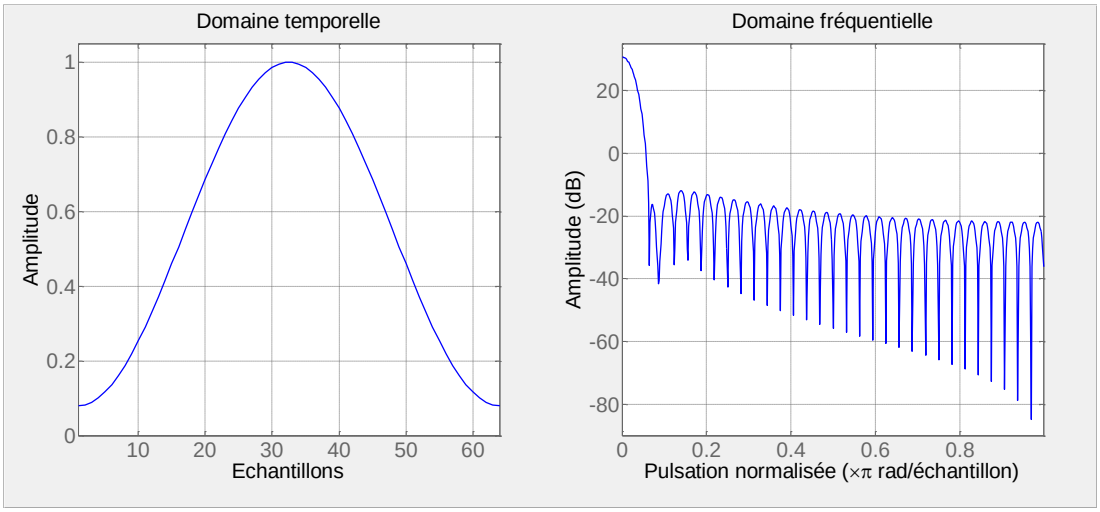


Figure 3.2 Fenêtre de Hamming.

Ci-dessous, nous pouvons voir une trame avant et après l'application de la fenêtre de Hamming.

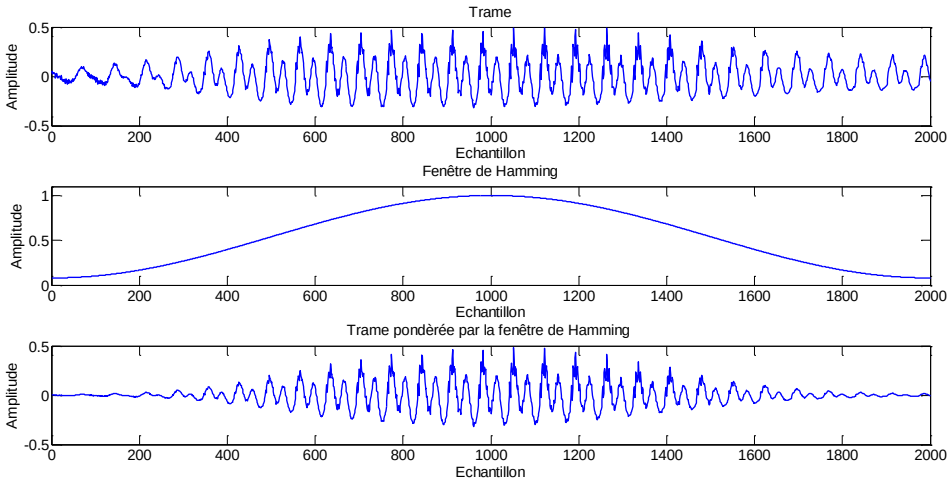


Figure 3.3 Avant et après l'application d'une fenêtre de Hamming à un cadre.

c Transformation de Fourier Discrète (TFD)

La troisième étape consiste à appliquer la transformation de Fourier discrète sur chaque trame. La transformation de Fourier permet de calculer le spectre d'amplitude de la trame.

$$X_t(k) = \sum_{n=0}^{N-1} x_t(n) e^{-j2\pi \frac{k}{N} n}, k = 0,1, \dots, N-1$$

où t indice trame et N longueur de la trame.

La manière rapide pour calculer la TFD est d'utiliser l'algorithme rapide FFT.

d Alignement de l'échelle de Mel

L'échelle MEL est basée sur la façon dont l'audition humaine perçoit les fréquences. Elle a été définie en définissant 1000 Mels égal à 1000 Hz comme point de référence.

Pour transformer une fréquence linéaire en une fréquence Mel, on utilise la formule de transformation en Mel (voir chapitre précédant).

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

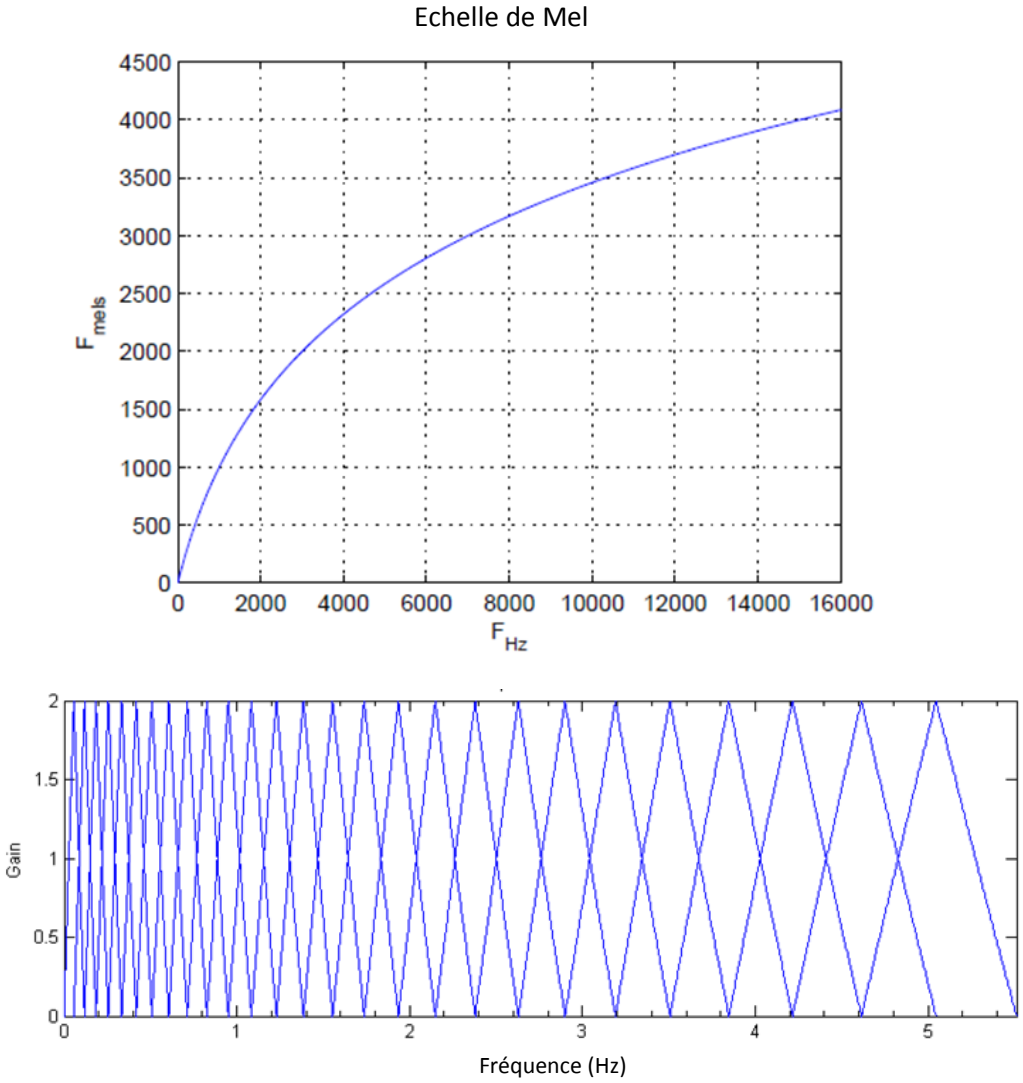


Figure 3.4 Relation entre l'échelle de fréquence et celle du Mel (haut), banc de filtres Mel (bas).

La distorsion de la fréquence Mel lors du calcul des MFCC est obtenue par l'utilisation d'un banc de filtres triangulaire espacée. Il se compose de plusieurs filtres triangulaires et espacés de Mel, et leurs sorties sont décrites par:

$$y(i) = \sum_{j=1}^{N} S_J . H_{i\ j}$$

où S_j est le spectre d'amplitude N-point et $H_{i\ j}$ la réponse d'amplitude échantillonnée d'un groupe de filtres à canal M. Le groupe de filtres de fréquence mel est appliqué dans le domaine fréquentielle.

e Transformation en cosinus discret (DCT)

La cinquième étape consiste à appliquer la compression en utilisant un logarithme sur les sorties de filtre, puis à appliquer la transformation en cosinus discrète qui donne les MFCC selon la formule suivante :

$$c[n] = \sum_{i=1}^{M} \log\big(Y(i)\big) . \cos\left(\frac{\pi n}{M}\left(i - \frac{1}{2}\right)\right)$$

f Apprentissage

Dans cette phase, le système de reconnaissance du locuteur enregistre la phrase prononcée par le locuteur, extraire les MFCC et les stocker sous forme d'empreinte vocale. L'apprentissage est utilisé pour améliorer la reconnaissance des performances des systèmes de reconnaissance du locuteur en enregistrant plusieurs prononciations de la phrase au lieu d'une seule.

3.2.2 DTW

Lors de l'utilisation de DTW, il existe un compromis entre précision de reconnaissance et l'efficacité de calcul. La raison en est que la création de plusieurs modèles pour une phrase augmentera le nombre de chemins DTW devant être calculés chaque fois qu'une décision est prise. La DTW calcule la distance entre les spectres et on prend ensuite la plus petite distance spectrale.

Dans le processus de reconnaissance automatique du locuteur, il est nécessaire de comparer les modèles de locuteur pour déterminer leur similarité.

DTW est utilisé pour mesurer la similarité de deux séries temporelles pouvant varier dans le temps.

Le but de DTW est de trouver un chemin $\{(p_1, q_1), (p_2, q_2), \ldots\ldots, (p_s, q_s)\}$ qui minimise

$\sum_{i=1}^{k} | t(p_i) - r(q_i)|$ Où t et r sont des vecteurs de longueur m et n, respectivement. Les contraintes suivantes doivent être appliquées:

Conditions aux limites: $(p_1, q_1) = (1,1) et\ (p_s, q_s) = (m, n)$

Contrainte locale: pour tout nœud donné (i, j) dans le chemin, les nœuds d'éventail possibles sont limités à (i − 1,), (i, j – 1) et (i − 1, j – 1). Il garantit un chemin monotone non décroissant. Pour tout élément donné dans t, il devrait y avoir au moins un élément correspondant dans r, et inversement.

Le chemin optimal est ensuite calculé par récursivité:

$$D(i,j) = | t(i) - r(j) | + \min\{D(i-1, j), D(i-1, j-1), D(i, j-1)\}$$

Avec la condition de départ D(1,1) = |t (1) − r (1) |

La distance DTW totale sera donnée par D (m, n) une fois que toute la matrice sera remplie à l'aide de la formule récursive pour D (i, j) ci-dessus

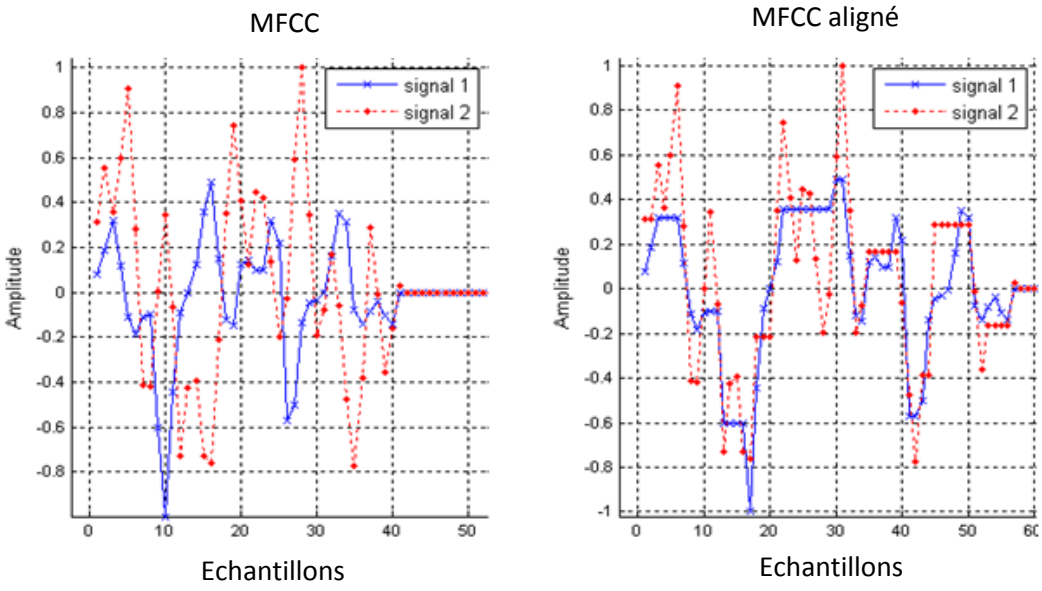


Figure 3.5 MFCC avant et après un alignement dynamique dans le temps.

3.2.3 Décision

La dernière étape est le processus de décision, il est basé sur la distance DTW calculée. Il s'agit de trouver la distance minimale entre les références MFCC des M locuteurs enregistrés.

3.3 Résultats

3.3.1 Entrainement

Comme il a été mentionné dans l'introduction du chapitre, l'identification vocale d'un locuteur correspond à une procédure de classification. Les classes se réfèrent dans ce cas à des locuteurs individuels de référence. L'obtention des locuteurs référentiels se réalise par une procédure d'entrainement, se qui nous mène à un modèle de référence (base de données).

Lors de la séance d'entraînement, nous identifions chaque locuteur en lui associons une étiquette. Dans notre expérience, nous avons utilisé onze locuteurs (S1 à S11). Ces locuteurs sont des segments de parole échantillonnés à 12500 Hz.

La figure (3.6) présente l'évolution temporelle et les MFCCs du locuteur S1.

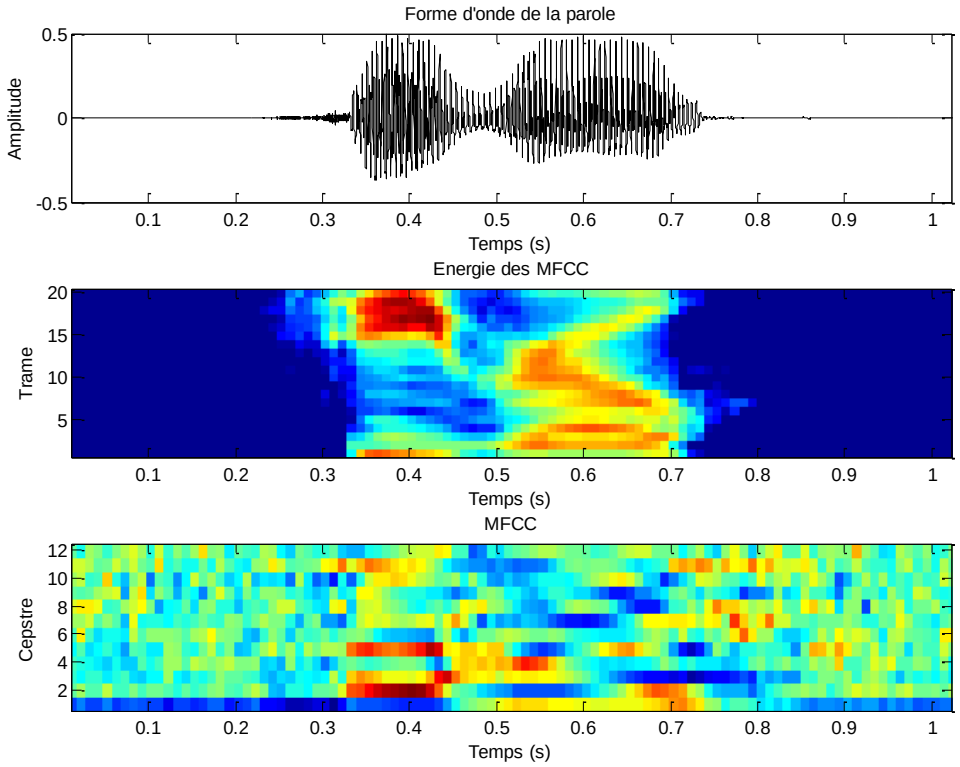


Figure 3.6 Évolution temporelle et MFCCs du locuteur S1.

3.3.1 Test

Pour cette phase, onze locuteurs (S1t à S11t) échantillonnés à 12500 Hz sont utilisés pour le test. L'évolution temporelle ainsi que les MFCCs du locuteur de test S1t et S4t est représenté dans la figure (3.7).

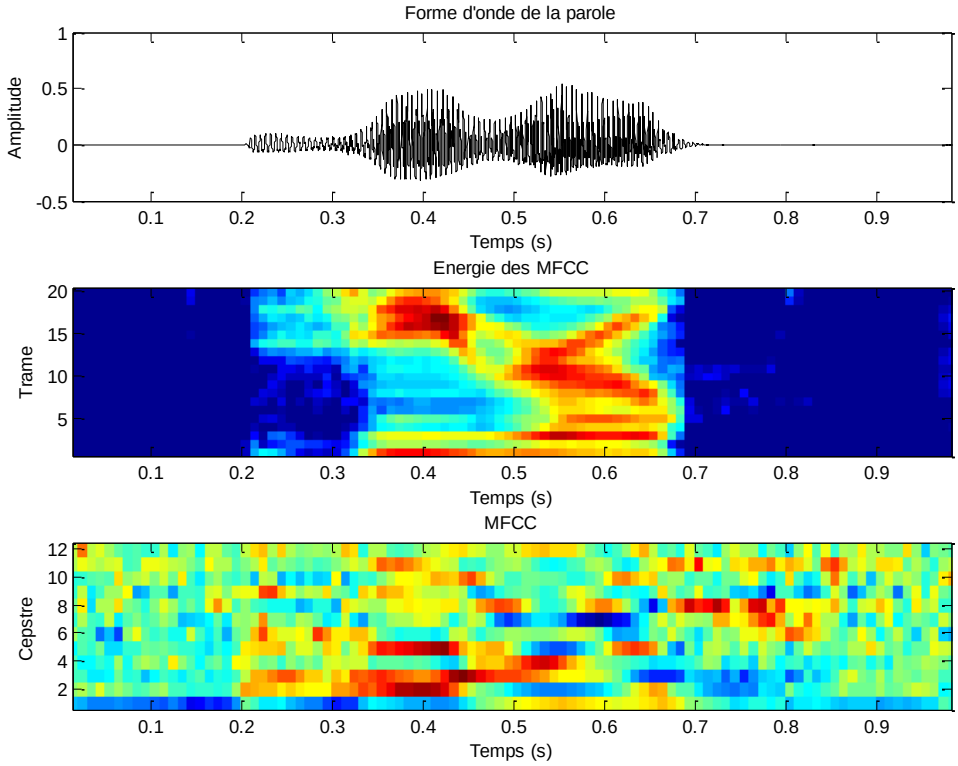


Figure 3.7 Évolution temporelle et MFCCs du locuteur de test S1t.

3.3.2 Identification

Dans cette étape, un locuteur test est comparé à la base de données par la DTW afin de trouver d'identité du locuteur à l'entrée.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
S1t	**2.36**	3.41	2.94	2.73	3.30	2.74	3.32	3.09	5.73	5.46	5.53
S2t	2.96	**1.78**	3.49	3.98	2.50	2.97	2.90	3.61	5.84	5.57	5.43
S3t	3.59	3.48	**0.00**	3.15	3.39	2.51	3.01	3.21	6.74	6.35	6.30
S4t	2.53	3.76	3.67	**2.24**	3.34	3.39	3.25	2.96	5.72	5.41	5.36
S5t	3.09	2.72	2.98	3.57	**2.13**	2.91	2.61	3.36	6.56	6.25	6.38
S6t	3.28	3.77	2.39	3.21	3.30	**1.87**	3.01	2.79	6.86	6.51	6.31
S7t	3.10	2.68	2.68	3.34	2.66	2.25	**2.49**	3.03	6.22	6.11	6.09
S8t	2.66	3.24	2.97	2.96	3.00	2.35	2.77	**2.34**	6.14	5.95	5.91
S9t	5.64	6.03	6.74	6.08	5.83	6.08	6.19	6.48	**0.00**	2.43	2.62
S10t	5.21	5.85	6.35	5.55	5.45	5.74	5.97	6.22	2.43	**0.00**	2.44
S11t	5.08	5.72	6.30	5.55	5.33	5.74	5.77	5.75	2.62	2.44	**0.00**

Le tableau ci-dessus présente le résultat de calcul de la distance par la DTW entre les locuteurs de test par rapport à ceux de la base de données. On peut clairement noter que la distance minimale se trouve dans la diagonale. Ce résultat prouve que l'identification des 11 locuteurs a été accomplie par la DTW.

3.4 Conclusion

Des études montrent que les paramètres MFCC semblent être plus efficaces que les fonctions basées sur le spectre de puissance lors de la représentation de la parole.

Il existe plusieurs techniques de classification populaires HMM, GMM, DTW, VQ. Les comparaisons entre VQ et DTW montrent que DTW donne beaucoup moins de taux d'erreur sur les petits échantillons.

DTW convient donc à une haute précision, en raison de son faible taux de réception erronée et de son utilisation avec des phrases courtes.

Conclusion générale

La reconnaissance automatique du locuteur est un sujet vaste et d'un grand intérêt. De nombreuses études ont été passées en revue pour aboutir à des conclusions sur la manière de mettre en place un programme de reconnaissance fonctionnel des locuteurs. Les résultats des études examinées sur la reconnaissance du locuteur ont permis de conclure que MFCC et DTW travaillent bien ensemble pour la reconnaissance du locuteur avec moins d'erreurs. Même si l'extraction de caractéristiques est une fonction essentielle d'un système de reconnaissance du locuteur, il restait une quantité surprenante de problèmes latéraux à comprendre et à résoudre. La partie traitement de signal de la reconnaissance vocale, par exemple, ne doit pas être sous-estimée, il faut donc chercher de nouvelles approches dans ce domaine pour mieux améliorer la reconnaissance.

L'algorithme DTW est un très bon outil capable de comparer deux spectres audio ayant des durées différentes, un débit, une intensité de la voix différente et cela de façon optimale en recherchant le meilleur chemin pour passer d'un spectre à l'autre. Néanmoins d'autres méthodes existent, comme les modèles de Markov cachés (HMM) par exemple bien plus puissant que l'algorithme DTW mais bien plus complexe.

Bibliographie

[1]Julien ALLEGRE, APPROCHE DE LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE,2003.

[2] https://interstices.info/de-la-reconnaissance-automatique-du-locuteur-a-la-signature-vocale/ Publié le : 19/03/2007, : Gilles Gonon & Frédéric Bimbo

[3] ZRIBI BOUJELBENE, Dorra BEN AYED MEZGHANI, et Noureddine ELLOUZE, Identification du Locuteur par Système Hybride GMM –SMO, SETIT 2009

5th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications – TUNISIA

[4] Houda KADI, La reconnaissance automatique du locuteur par la voix IP,Master science et technique système intelligeant et réseaux ,2014.

[5]Gaël RICHARD Master IAD Module PS Février 2008 Reconnaissance de la parole, Février 2008 .

[6] Corine fredouille, Approche statique pour la Reconnaissance Automatique du locuteur : Informations Dynamiques et Normalisation Bayesienne des Vraisemblances ,2000.

[7] Mohamed Fezari, mouldi Bedda, Ali Al Dahoud , IDENTIFICATION DU LOCUTEUR PAR UNE METHODE HYBRIDE ,Mohamed Fezari, mouldi Bedda, Ali Al Dahoud

[8]Bachiri Salah Eddine, REALISATION D'UNE APPLICATION EN VUE DE LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE TRANSCODEE SPEEX ET G.729, *MEMOIRE DE MASTER,2016.*

[9] R. McMunn, Speed, Distance and Time Questions (Testing Series), 2015.

[10] Mlle.BELGHITRI KARIMA. Pour l'obtention du diplôme de Master en Réseaux et Système distribué. Thème Système sécurisé à base vocale.2015.

[11]Mme Feriel DEBBECHE-GUERID, CONCEPTION D'UN SYSTEME ACOUSTICO-ANATOMIQUE POUR L'IDENTIFICATION DU LOCUTEUR : ARCHITECTURE ET PARAMETRISATION, Présenté en vue de l'obtention du diplôme de Magistère,2008.

[12] Douglas A. Reynolds Automatic Speaker Recognition Using Gaussian Mixture Speaker Models ,1995.

[13] Speech Recognition Berlin Chen ,Department of Computer Science & Information Engineering, National Taiwan Normal University.

[14]Yassine. Mami. Reconnaissance De Locuteur Par Localistion Dans un Espace de Locuteurs de Référence. Thése de Doctorat de l'École National de Télécommunications, Paris, Octobre 2003.

[15]Mr. Abdelghani HARRAG ,Thème Extraction des données d'une base: Application à l'extraction des traits du locuteur, Doctorat Sciences,2011.

[16] Mr. Hacine Gharbi Abdenour, Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole, l'obtention du diplôme de Doctorat en Sciences, décembre 2012.

[17] BENCHENIEF Abderezek, RECONNAISSANCE VOCALE BASÉE SUR LES SVM , l'obtention du diplôme de magister en Electronique,2011.

[18]ALEXANDRU CRACIUN ,IMPLÉMENTATION D'UNE MÉTHODE ROBUSTE DE DÉTECTION D'ACTIVITÉ VOCALE SUR LE PROCESSEUR DE SIGNAL TMS320C6711,2004.

[19] M.DIDICHE1 & A.DEBILOU MODELISATION NEURO-PREDICTIVE POUR LA CLASSIFICATION PHONETIQUE DE LA LANGUE ARABE, Décembre 2013, pp.87-91.

[20] René Boite,Hervé Bourlard,Thierry Dutoit,Joel Hncq et Henri Leich,traitemant de la parole,2000.

[21] https://www.cepstral.com/

[22] gaël richard, télécom paris Tech ,juin 2012 .

[23] Minh N. Do, How to Build an Automatic Speaker Recognition System.