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

Université Amar Thelidji- Laghouat
FACULTE DE TECHNOLOGIE
DEPARTEMENT ELECTRONIQUE

MÉMOIRE DE MASTER

Domaine : Sciences et Technologies
Filière : Télécommunications
Spécialité : Télécommunications

OPTION : Réseaux et Télécommunications

Thème

Reconnaissance Automatique De La Parole Par Déformation Temporelle Dynamique

Présenté par :

OTHMANI Bouchra

KHALDI Feiza

Membres du jury :

Encadreur Dr. REGGAB Mourad

Président Dr. KORIBA Mustapha

Examinateur Dr. CHELALI safouane

Promotion : 2023

REMERCIEMENT

Tout d'abord on remercie **ALLAH** *puissant de la bonne santé, la volonté et de la patience qu'il nous a donnée tout au long de notre étude.*

Nous remercions très sincèrement notre encadreur de ce travail **Dr. Mourad REGGAB**, *pour ses conseils, et ses orientations judicieuses sa patience et diligence, et par ses suggestions a grandement facilité ce travail.*

Nous tenons à exprimer notre gratitude aux membres de jury qui ont bien voulu examiner ce travail.

Nos remerciements vont aussi à tous les enseignants de la faculté science technologie spécialement du département d'électronique branche télécommunications fil et sans fil.

Enfin, nous adresse notre sincère remerciement à tous ceux qui ont contribué, de près ou de loin, à la réalisation de ce travail.

Dédicaces

Je dédie ce travail :

Aux deux personnes les plus nobles, précieux et les plus chères au monde, Mon père qu'ALLAH repose son âme. A ma très chère mère, Mère exemplaire pour mes frères et pour moi-même ; tu as su donner l'éducation qu'il nous faut pour affronter les épreuves de la vie. Tu nous as comblés de ton amour maternel et tu répondais présente à chacune de nos sollicitations. Puisse le tout Puissant t'accorder longue vie afin de profiter des fruits de ce labeur.

À mes frères et mes sœurs, pour leurs aides, disponibilités et précieux conseils, que la vie vous apporte toute la joie et le bonheur.

À la femme de mon frère et à mes neveux, Muhammad et Maria.

À toute ma famille.

À tous mes amis sans exception. À tous les personnes qui m'ont souhaité le succès, la joie et le bonheur, et tous ceux qui m'aiment.

OTHMANI Bouchra

Dédicaces

Avant tout, je tiens à remercier l'ALLAH, et l'unique qui m'offre le courage et la volonté nécessaire pour affronter les différentes difficultés de la vie,

Je dédie ce modeste travail:

À ceux qui sont les plus chers du monde, mon père, et ma mère, à qui je n'arriverai jamais à exprimer ma gratitude et ma reconnaissance, pour ses amours ses soutiens tout au long de mes études.

À mes Sœurs : Raouda, Ayaa, Anfal. . À toute ma famille.

À mes amis: Lalia, Maroua, Karima, Iman, Hanan.

KHALDI Feiza

ملخص

الهدف من هذه المذكرة هو بناء نظام بإمكانه التعرف على أول عشرة أرقام عربية في وضع المتكلم الواحد، حيث اعتمدنا على خوارزمية DTW والترميز (المعالجة المسبقة) MFCC , ولقد حصلنا على معدل تعرف مقبول.

الكلمات المفتاحية: MFCC, DTW,RAP ,اللغة العربية, Mono locuteur.

Résume

L'objectif de ce mémoire est construction d'un système de RAP capable de reconnaître les dix premiers chiffres arabes en mode mono locuteur, basé sur la méthode DTW et le codage (prétraitement) MFCC, on a obtenu un taux de reconnaissance acceptable.

Mots-clés : RAP, DTW, MFCC, Arabe, Mono locuteur.

Abstract

The objective of this dissertation is the construction of an ASR system capable of recognizing the first ten Arabic digits in speaker dependant mode, based on the DTW method and the MFCC coding (preprocessing); we obtained an acceptable recognition rate.

Keywords: ASR, DTW, MFCC, Arabic, Speaker dépendant.

Sommaire

Liste des figures

Liste des tableaux

Liste des abréviations

ANN	automatic neutral network
ASR	Automatic Speech Recognition
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DTW	Dynamic Time Warping
FB	Filter bank
FFT	Fast Fourier Transform
HMM	Hidden Markov Models
IDCT	Inverse Discrete Cosine Transform
IHM	Interface homme machine
LPC	Linear predictive coding
LPCC	Linear predictive coding coefficient
MFCC	Mel-frequency cepstral coefficient
PLP	perceptual linear prediction
RAP	Reconnaissance Automatique de la parole
RASTA	RelAtive SpecTrAl
VAD	Voice activity detection

Introduction générale

Depuis presque trente ans, la reconnaissance automatique de la parole est un domaine qui a captivé le public ainsi que de nombreux chercheurs. À ses balbutiements, les projections sur ses applications étaient très optimistes : quoi de plus naturel que de parler à une machine, sans avoir à s'encombrer d'un clavier ? Malheureusement, malgré l'incroyable évolution des ordinateurs et des connaissances, la reconnaissance automatique de la parole n'en demeure pas moins un sujet de recherche toujours actif...et les résultats obtenus sont encore loin de l'idéal qu'on aurait pu en attendre, il y a vingt ans.

Cependant, si le système de reconnaissance idéal n'existe pas encore, des applications concrètes émergent petit à petit. La reconnaissance automatique de la parole commence à équiper certains téléphones ou GPS qui, en identifiant certains mots clefs, permettent d'effectuer les tâches demandées. Les systèmes de reconnaissance sont également utilisés pour indexer de grandes bases de données audiovisuelles, pour rechercher des termes dans des flux audio ou encore comme interface de dialogue homme-machine. Dans la pratique, quand les conditions d'utilisation sont correctes, ces systèmes s'avèrent efficaces. Néanmoins, les principales limites des systèmes actuels sont relatives à leur robustesse : les conditions d'utilisation doivent être similaires à celles utilisées pour entraîner le système, l'environnement sonore peu bruité, les locuteurs ne peuvent pas parler simultanément. Souvent, l'utilisateur a dû s'adapter pour utiliser les logiciels.

De plus en plus, la technologie de la reconnaissance de la parole fait son chemin vers des applications réelles. Actuellement un changement qualitatif dans l'état de l'art est apparu promettant d'apporter des capacités de reconnaissance et de les mettre à la portée de chaque personne.

La reconnaissance de la parole continue pour un vocabulaire moyen (quelques milliers de mots) est actuellement possible dans un logiciel de reconnaissance de la parole. La reconnaissance de la parole humaine se situe à l'intersection de nombreux domaines tels que l'acoustique, l'électronique, la phonétique...Pour atteindre un haut niveau, un système de reconnaissance de la parole doit s'inspirer des travaux d'une vaste gamme de disciplines scientifiques : Mathématique, informatique, technologie,....

Le signal de la parole est l'un des signaux les plus complexes, il n'est pas facile de le caractériser par un modèle simple. L'un des problèmes de la reconnaissance de la parole est la variabilité du signal. Les précurseurs dans le domaine du traitement de la parole, pensaient

que la parole était une simple juxtaposition d'éléments appelés "caractéristiques distinctives" ou invariants (phonèmes), et qu'à partir d'échantillons représentant les invariants d'une langue, on pouvait reconstituer ou reconnaître n'importe quelle phrase. Cette idée théorique n'est pas toujours vraie dans la reconnaissance de la parole continue à cause du problème de coarticulation. Pour surmonter ces difficultés, de nombreuses méthodes et modèles mathématiques originaux ou adaptés d'autres domaines ont été développés, parmi lesquelles on peut citer : la comparaison dynamique, les systèmes experts, les réseaux de neurones, les modèles stochastiques et en particulier les modèles de Markov cachés, etc.

Notre travail s'inscrit dans le cadre général de la RAP. Elle consiste à réaliser un système de reconnaissance automatique de la parole par la méthode DTW. Ce mémoire s'articule autour de quatre chapitres :

- Le premier chapitre est une introduction générale au domaine de la reconnaissance de la parole, ses difficultés, les principales méthodes utilisées en la reconnaissance.
- Le deuxième chapitre illustre le signal de la parole dans l'état de la production et perception (naturelle + modèle mathématique), enfin techniques de prétraitement du signal parole.
- Le troisième chapitre présente la technique et l'algorithme de DTW.
- Le dernier chapitre présente les résultats et les interprétations.

CHAPITRE I :

GÉNÉRALITÉS SUR LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

I.1 Introduction

La parole est un moyen de communication très efficace et naturel utilisé par l'humain. Depuis long temps, il rêve de pouvoir s'adresser par ce même moyen à des machines ce qui les rendre plus intelligentes. Cependant, malgré les énormes efforts de recherches consacrés dans l'essai de créer une telle machine intelligente qui peut reconnaître le mot parlé et comprend sa signification, on est loin d'atteindre le but désiré d'une machine qui peut comprendre le discours parlé par tous les locuteurs dans tous les environnements, ce qui est due aux limites du système de reconnaissance de la parole. Alors, qu'est-ce que l'on entend par la reconnaissance automatique de la parole (RAP) ? La reconnaissance automatique de la parole est l'un des deux domaines du traitement automatique de la parole, l'autre étant la synthèse vocale. La reconnaissance automatique de la parole permet à la machine de comprendre et de traiter des informations fournies oralement par un utilisateur humain. Elle consiste à employer des techniques d'appariement afin de comparer une onde sonore à un ensemble d'échantillons, composés généralement de mots mais aussi, plus récemment, de phonèmes.

Dans ce chapitre, on va essayer de donner une idée générale sur la RAP, de discuter la difficulté de cette dernière, ainsi que les technique utilisées pour elle.

I.2 Définition

La reconnaissance automatique de la parole (souvent improprement appelée reconnaissance vocale) est une technique informatique qui permet d'analyser la voix humaine captée au moyen d'un microphone pour la transcrire sous la forme d'un texte exploitable par une machine.

La reconnaissance de la parole, ainsi que la synthèse de la parole, l'identification du locuteur ou la vérification du locuteur, font partie des techniques de traitement de la parole. Ces techniques permettent notamment de réaliser des interfaces homme-machine (IHM) où une partie de l'interaction se fait à la voix : « interfaces vocales ».

Parmi les nombreuses applications, on peut citer les applications de dictée vocale sur ordinateur où la difficulté tient à la taille du vocabulaire et à la longueur des phrases, mais aussi les applications téléphoniques de type serveur vocal interactif, où la difficulté tient

plutôt à la nécessité de reconnaître n'importe quelle voix dans des conditions acoustiques variables et souvent bruyantes (téléphones mobiles dans des lieux publics).

Dans Parole et dialogue homme-machine, W. Minker et S. Bennacef expliquent que la reconnaissance automatique de la parole est un domaine complexe, car il existe une différence importante entre le langage formel, qui est compris et utilisé par les machines, et le langage naturel, que les humains utilisent. Le langage formel est structuré par des règles syntaxiques strictes et sans ambigüité. À l'inverse, dans le langage naturel, des mots ou des phrases peuvent avoir plusieurs sens selon l'intonation de l'énonciateur ou le contexte. **[1]**

I.3 Les principaux bénéfices de la reconnaissance vocale automatique

Si les interactions vocales sont bien menées, sont à la portée de tous les utilisateurs allant des enfants aux personnes âgées. Elles se font sans apprentissage préalable et sans pré requis sur le niveau d'éducation de l'utilisateur.

Les principaux bénéfices de la reconnaissance vocale automatique (ASR) par rapport aux traditionnelles commandes à commutateurs, claviers, écrans, souris ou autres interfaces de pointage, sont les suivants :

Les interactions vocales sont compatibles avec d'autres activités réalisées en parallèle : activités manuelles, activités sportives, conduites d'engins…

Les interactions vocales ne nécessitent pas de coordination psychomotrice entre la vue et le geste. Cela convient parfaitement aux personnes ayant une déficience motrice ou visuelle. On pense aux situations de handicap, mais aussi aux situations d'ensoleillement, de vibrations ou des environnements hostiles de froid ou d'humidité,

La commande vocale repose sur un appareillage simple, composé d'un microphone et écouteur ou haut-parleur, qui constitue la fonction de base du téléphone.

Ainsi, les commandes vocales s'utilisent parfaitement à la maison pour allumer ou éteindre des ampoules, à l'usine, à l'hôpital, en voiture pour obtenir l'itinéraire, au téléphone pour changer une date de livraison, sur son Smartphone pour dicter un email ou un mémo, etc.

La particularité de la commande vocale est aussi de s'offrir avec très peu d'apprentissage. Alors que le clavier nécessite une habitude pour trouver la bonne touche, que la souris nécessite de coordonner le geste sur l'écran, ou que le commutateur nécessite la lecture d'un manuel d'utilisation, la commande vocale est immédiatement accessible pour tous les humains qui peuvent s'exprimer dans la langue autorisée par le système.

C'est la machine qui s'adapte à l'homme et non l'homme qui s'adapte à la machine. **[2]**

I.4 Difficultés de la reconnaissance automatique de la parole

Pour l'instant, la reconnaissance automatique de la parole peut être considéré comme une boîte noire construite à partir d'un logiciel et d'un matériel (tel qu'un microphone). Il prend comme entrée la voix (speech) et renvoie la transcription des mots prononcés dans le discours. Nous détaillerons cette boîte dans les prochains blogs.

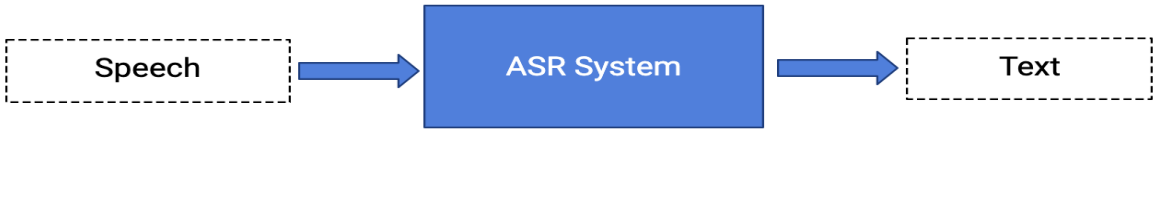


Figure I.1:Le system de reconnaissance automatique de la parole

Les humains font de la reconnaissance vocale presque sans effort. Cependant, pour les machines, la reconnaissance vocale est difficile car l'écoute est complexe et plus compliquée qu'on ne le pense. Voyons comment fonctionne le processus humain et ce qu'une machine devrait faire :

Les humains ont naturellement l'anatomie dédiée pour accepter une onde acoustique, tandis que les machines devraient convertir le signal analogique (l'onde acoustique) en représentation numérique.

- Lorsque quelqu'un nous parle (dans une salle de classe par exemple), nous devons alors séparer ses mots (son signal acoustique) de tout le bruit de fond. En général, le bruit peut être le discours du professeur, les conversations d'autres personnes, la circulation (oui, cela peut arriver même dans la salle de classe), etc. Par conséquent, la machine doit séparer la parole du bruit.

- Certaines personnes parlent lentement, tandis que d'autres parlent vite. D'autres ne s'arrêtent pas, ou ne ralentissent pas à la fin de la phrase avant d'en commencer une nouvelle. Dans ce cas, les phrases sonnent comme un long flux continu de mots (et ne peuvent donc pas distinguer les phrases). En outre, il est ambigu lorsqu'un mot se termine et qu'un autre commence. Dans ce cas, une machine devrait traiter ces séparateurs dans la parole.
- De nombreux facteurs tels que le sexe, l'âge, l'accent, le contexte, l'humeur influencent notre voix. C'est pourquoi chacun sonne différemment. Même si vous demandez à quelqu'un de dire deux fois « Quelle heure est-il ? », il ne le dira probablement pas de la même manière à chaque fois, simplement parce que les voix changent très souvent. Ici, la machine doit résister à la variabilité de la parole.
- Plus encore, supposons que nous soyons dans une conversation avec un enfant, une femme adulte, un vieil homme et plusieurs autres personnes, qui ont tous des accents différents. Nous devons alors donner un sens à l'ensemble du fil de la conversation et à ce que chacun dit. Nous devons comprendre que le mot « porte », par exemple, signifie la même chose, peu importe qui le dit. Ainsi, la machine devra reconnaître les sons même s'ils sont prononcés différemment « dog » ou « doooooor » par des personnes radicalement différentes, américaines, anglaises, espagnoles…
- Il y a beaucoup de mots qui ont un son similaire (c'est-à-dire homophones) ou identique (comme « to », « too », « two ») mais leur signification est différente. Ici, le contexte est nécessaire afin de savoir quelle signification l'orateur entend. Dans ce cas, une machine doit désambiguïser les homophones.
- Il existe de nombreux « fillers » que nous utilisons dans la parole, tels que « um », « hmm », « euh » etc. et nous savons instinctivement comment les filtrer. Ils ne nous trompent pas ou ne nous amènent pas à interpréter incorrectement les paroles de l'orateur. La machine devrait également filtrer ces remplissages.
- Parfois, nous entendons mal les phrases, donc nous les comprenons mal. Comme de telles erreurs pourraient être très gênante, la machine devrait également gérer ces malentendus, et doit être nettement meilleure que nous dans cette tâche.
- Enfin, si tout cela ne semble pas beaucoup, nous devons connaître la syntaxe et la sémantique de la langue que nous utilisons ainsi que le contexte.

Il est étonnant que nous fassions tout cela dans une simple conversation. En ce sens, nos cerveaux sont incroyables. Il n'est donc pas étonnant que les machines luttent pour faire tout cela. Pourtant, la reconnaissance vocale a fait beaucoup de chemin, et ce n'est peut-être que le début. **[3]**

I.5 Les techniques de traitement de la parole

I.5.1 Déformation temporelle dynamique

Il s'agit d'une méthode apparue dans les années 80 dans le domaine du traitement de la parole et encore utilisée dans des systèmes de reconnaissance vocale disposant de ressources matérielles limitées. Dans les systèmes de reconnaissance basés sur la DTW, chaque mot du lexique est représenté par une réalisation de référence. Le processus de reconnaissance consiste à évaluer la distance d'une observation à chacune des références. Toute la difficulté du décodage réside dans cette mesure d'un degré de similarité entre des formes acoustiques variables à la fois au niveau spectral et temporel. En effet, les réalisations acoustiques d'un mot subissent des déformations spectrales liées à divers paramètres (locuteurs, contextes, conditions d'acquisition, etc.) mais aussi des déformations temporelles globales (vitesse d'élocution) ou plus locales (accent, dynamique des organes phonatoires, etc.). Pour comparer deux segments de parole soumis à cette double déformation, il faut préalablement leur appliquer un processus d'alignement temporel. L'algorithme DTW (Dynamic Time Warping) réalise cet alignement en recherchant, parmi tous les alignements possibles, celui qui minimise une fonction de coût intégrant l'écart spectral des données alignées et un coût de distorsion temporelle. La distance retenue est celle correspondant à l'alignement de coût minimal. Rapide dans des tâches à petit vocabulaire, cette technique a un certain nombre d'inconvénients importants qui limitent son champ d'application. D'une part, la modélisation des mots par une instance est très peu robuste à l'ensemble des variabilités acoustiques. Cette faiblesse peut être partiellement limitée par l'utilisation de plusieurs références par mot, par un choix plus fin des références ou encore par l'usage de distances spectrales robustes (type Malhabolis). Néanmoins, cette technique est plus adaptée à un contexte d'utilisation mono locuteur en environnement peu bruité. D'autre part, la complexité des modèles et du décodage sont proportionnels à la taille du lexique, ce qui exclut l'utilisation de la DTW dans des systèmes grand vocabulaire. Enfin, bien que diverses extensions à la reconnaissance de la parole continue aient été expérimentées, cette méthode ne permet, dans sa version standard, que la reconnaissance de mots isolés. **[4]**

I.5.2 Les Modèles de Markov Cachés

Le Modèle de Markov Caché (Hidden Markov Model) est une méthode statistique puissante pour caractériser les échantillons de données observés d'un processus à temps discret. Elle apporte non seulement un moyen efficace de construction de modèles paramétriques, mais elle incorpore aussi le principe de programmation dynamique pour unifier la segmentation et la classification de séquence de données variant dans le temps.

Dans la modélisation d'un processus par un HMM, les échantillons peuvent être caractérisés par un processus paramétrique aléatoire dont les paramètres peuvent être estimés dans un cadre de travail bien défini. La théorie de base des HMM a été publiée dans une série de papiers par L. Baum.

Les HMMs sont devenus la méthode la plus couramment utilisée pour la modélisation des signaux de parole dans les applications suivantes : reconnaissance automatique de la parole, suivi de la fréquence fondamentale et des formants, synthèse vocale, traduction automatique, étiquetage syntaxique, compréhension du langage oral, traduction automatique... Dans une chaîne de Markov, chaque état correspond à un événement à observation déterministe (la sortie de ses sources pour un état donné n'est pas aléatoire). Une extension naturelle à la chaîne de Markov introduit un processus non déterministe qui génère des symboles de sortie pour chaque état. L'observation est donc une fonction probabiliste de l'état. Le nouveau modèle est appelé HMM, pouvant être vu comme deux processus stochastiques imbriqués dont l'un (la séquence d'états) est non observable directement. Ce processus sous-jacent est donc associé de façon probabiliste à un autre processus produisant la séquence de trames, qui elle, est observable. Ci-dessous, nous présentons les trois problèmes de base à résoudre pour l'application de cette méthode :

- Le problème d'évaluation : Quelle est la probabilité d'un modèle générant une séquence d'observation ? Ce problème est résolu par l'application de l'algorithme FORWARD.
- Le problème de décodage : Quelle est la séquence d'états la plus probable pour un modèle et une séquence d'observation donnés ? On utilise l'algorithme VITERBI pour effectuer cette tâche.
- Le problème d'apprentissage : Comment peut-on ajuster les paramètres du modèle pour maximiser la vraisemblance (probabilité jointe) de génération d'une séquence d'observation ? Les algorithmes de BAUM-WELCH et de VITERBI permettent d'effectuer l'apprentissage. Dans les applications de la parole, on utilise fréquemment les

HMM continus, où l'observation n'appartient pas à un ensemble discret mais à une distribution (le plus souvent normale). Ainsi, une topologie gauche-droite pour un HMM continu permet de modéliser les états successifs d'un phonème pour un signal de parole. Plus généralement, l'objectif à atteindre est la détermination à partir de vecteurs acoustiques de la séquence phonétique prononcée. **[5]**

I.5.3 Les réseaux de neurones

Depuis une vingtaine d'année, les réseaux de neuromimétiques constituent une technique utilisée dans les systèmes de reconnaissance automatique de la parole. Ils sont basés sur une modélisation grossière du neurone biologique (neurone formel). Tout comme le neurone biologique, le neurone formel calcule sont activation en fonction des signaux qu'il reçoit d'autres neurones, pondérés par des « poids synaptiques » et d'une fonction d'activation plus ou moins complexe.

L'ensemble de ces neurones est organisé selon des architectures plus ou moins complexes matérialisées par les connexions entre ces neurones. Selon cette architecture, ainsi que le type de la fonction d'activation, les réseaux de neurones peuvent résoudre un certain nombre de problèmes tels que des problèmes de classification, de mémorisation et de résolution de contraintes. Une particularité des réseaux de neurones est qu'ils sont dotés d'algorithmes l'apprentissage qui leur permettent d'apprendre les formes, les classes à reconnaître et à classer ou bien les problèmes à résoudre. Ces algorithmes sont soit supervisés lorsque l'on connaît déjà les classes associées aux exemples du corpus d'apprentissage, soit non-supervisés. Le but recherché est de faire en sorte que les réseaux de neurones répondent correctement à des stimuli jamais rencontrés. Etant donné le large spectre des possibilités des réseaux neuroleptiques, ils peuvent être employés à de nombreux niveaux dans un système de traitement automatique de la parole. De nombreuses études ont été menées pour les utiliser pour le traitement de signal (filtrage, annulation d'échos, séparation de sources), la modélisation acoustique mais aussi pour des tâches de plus haut niveaux telles que la modélisation linguistique. **[6]**

I.6 Conclusion

Dans ce chapitre, nous avons présenté la reconnaissance automatique de la parole, qui peut être divisée en trois étapes, à savoir : d'abord la définition de la RAP, puis la raison de son utilisation, ses difficultés, et enfin les techniques utilisées dans celui-ci.

CHAPITRE 2 :

LE SIGNAL DE LA PAROLE

II.1 Introduction

La parole est un moyen naturel de communication entre les hommes. L'auditeur humain utilise plusieurs niveaux de traitement pour la reconnaissance de la parole. Le premier niveau est la détermination des caractéristiques du signal de parole lui-même, c'est-à-dire l'analyse acoustique. Viennent ensuite les niveaux phonétiques, lexical, syntaxique, sémantique, etc. C'est dire combien chez l'homme la reconnaissance et la compréhension de la parole sont très complexes pour les machines.

Dans ce chapitre nous expliquerons des notions de base sur la production et perception de la parole, des techniques analyse acoustique.

II.2 Définition

La parole est une succession de séquences sonores et de silences, et le seul moyen qui permet de communiquer la pensée par un système de sons articulés. Les humains sont les seuls êtres vivants qui utilisent un tel type des systèmes structurés, et il est le résultat d'une variation de la pression produite par l'émission d'un son par un locuteur. **[7]**

II.3 Production de la parole

Le son que nous connaissons sous le nom de parole commence par la contraction des poumons pour expulser l'air, qui transporte le son d'une distribution de fréquence approximativement gaussienne.

Cet air est forcé à travers le tractus bronchique au-delà d'un ensemble de plis musculaires au sommet de la trachée appelés cordes vocales, et les fait vibrer. L'air pénètre ensuite à l'arrière de la cavité buccale où il suit l'un des deux chemins vers l'extérieur. Le premier chemin est au-dessus et autour de la langue, au-delà des dents et à travers la bouche. La deuxième voie passe par la cavité nasale - et c'est la seule voie possible lorsque le voile est fermé.

La figure II.1 montre un schéma de l'appareil de production de la parole (autrement connu sous le nom de tête humaine).

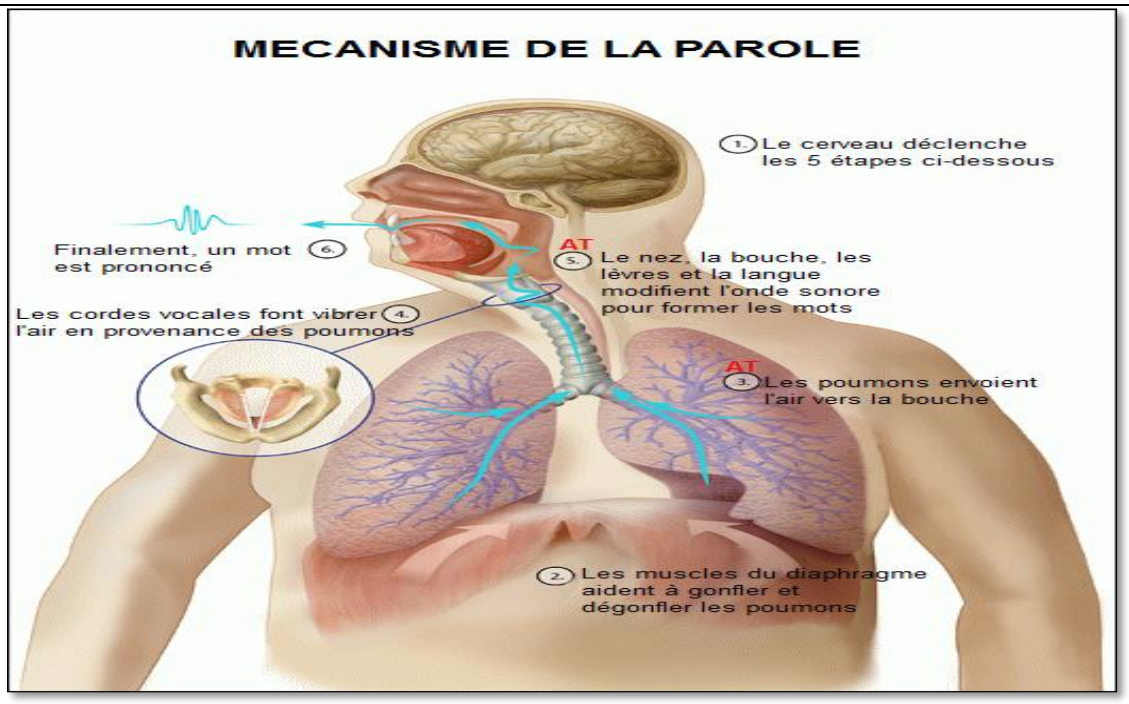


Figure II.1 : Représentation de l'appareil phonatoire et du rôle des différents organes intervenant dans la production de sons de parole voisés.

Le son réellement produit dépend de nombreux critères, notamment la puissance pulmonaire et la modulation de la pression, la constriction de la glotte, la tension des cordes vocales, la forme de la bouche et la position de la langue et des dents.

a) La puissance pulmonaire affecte principalement le volume du son, mais une variation rapide distingue souvent une limite entre les syllabes.

b) Si la glotte est fermée temporairement pendant la parole, un coup de glotte se produit tel que le / t / dans une lecture à l'accent du Yorkshire de 'I went t' shops'. Un son explosif comme le /d/ dans 'dog', est un court arrêt suivi d'un relâchement explosif.

c) La tension musculaire des cordes vocales fait vibrer les cordes à des taux différents, formant les fréquences de hauteur. Les sons sans voix (par exemple /s/ dans 'six'), où les cordes vocales ne vibrent pas, ont peu ou pas de structure de hauteur.

d) Si l'air est dévié par le nez par la fermeture du velum, un son nasal tel que /m/ dans "mad" en résulte. Un timbre différent résulte également de la longueur de trajet légèrement différente des poumons au nez par rapport aux poumons à la bouche (imaginez deux tuyaux d'orgue de longueur différente).

e) Si l'air passe par la bouche, une langue bossue et l'ouverture puis la fermeture de la mâchoire inférieure provoquent un son de voyelle (par exemple /a/ dans 'card'), si la mâchoire inférieure ne se ferme pas, un glissement (par exemple /w/ dans 'won') est le résultat.

f) Des sons différents se produisent également si l'air est forcé au-delà des côtés d'une langue touchant le toit de la bouche ou les dents (par exemple, /l/ dans « luck » et le son /th/).

Les actions ci-dessus doivent être enchaînées par le locuteur afin de construire des phrases cohérentes. En pratique, les sons vont s'articuler et se fondre les uns dans les autres dans une certaine mesure, comme la dernière partie d'un son de voyelle changeant en fonction du son suivant.

Cela peut être illustré en considérant comment le son /o/ dans « or » et dans « of » diffère. **[8]**

II.4 Perception de la parole

L'oreille, comme représenté schématiquement sur la figure II.2, comprend le pavillon qui filtre le son et le focalise dans le conduit auditif externe. Le son agit alors sur le tympan où il est transmis et amplifié par les trois os, le marteau, l'enclume et l'étrier, jusqu'à la fenêtre ovale, ouvrant sur la cochlée.

La cochlée, en tant que tube enroulé, contient une paire de membranes semi-rigides d'environ 35 mm de long enfermées dans un fluide appelé endolymphe.

La membrane basilaire porte les organes de Corti, dont chacun contient un certain nombre de cellules ciliées disposées en deux rangées (environ 3500 cellules ciliées internes et 20 000 cellules ciliées externes).

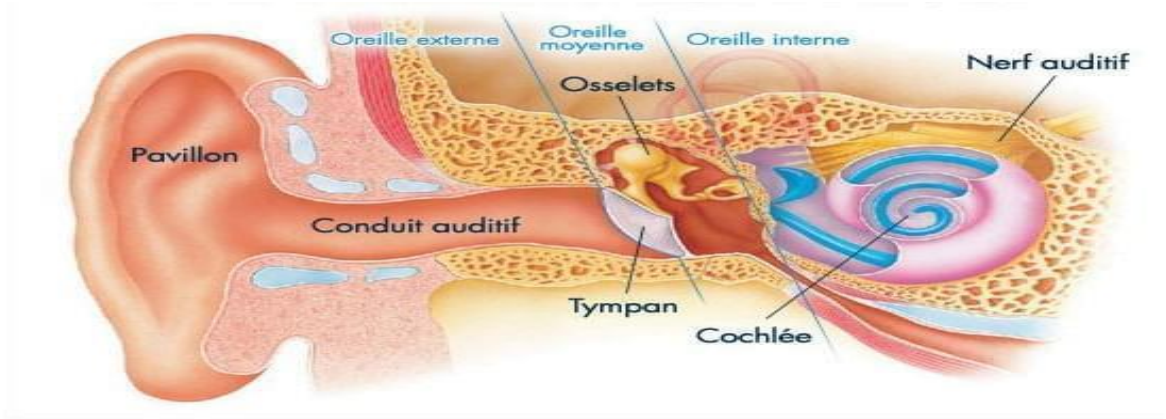


Figure II.2 : Schéma du système auditif humain

De petites impulsions électriques sont générées par les cellules ciliées lorsqu'une tension leur est appliquée (généralement lorsqu'elles sont tirées, de sorte que la réponse auditive se produit pendant la partie de raréfaction d'une oscillation dans le fluide qui les entoure).

La largeur et la rigidité de la membrane basilaire se rétrécissent sur sa longueur, et donc lorsqu'une vibration est appliquée au système, l'emplacement sur la membrane basilaire qui résonne dépend de la fréquence. Ce processus fournit une certaine sélectivité en fréquence à l'oreille, le traitement actif par le cortex auditif fournissant le reste.

En fait, le traitement du cortex auditif améliore considérablement la sélectivité en fréquence. Les nerfs transportent les impulsions électriques de l'oreille vers le cerveau où se produit un traitement inconscient approfondi. La nature du traitement n'est pas entièrement comprise ; cependant, il est hautement non linéaire, implique une corrélation entre les signaux de chaque oreille et éventuellement avec d'autres sens (tels que la sensation des vibrations et la vue). On peut supposer que l'oreille s'intègre sur de courtes périodes et traite également des schémas plus grossièrement répétitifs de différentes manières.

Des mécanismes actifs fonctionnent également dans l'oreille, notamment la tension des muscles opérant sur le marteau et l'étrier, pour protéger l'oreille des sons forts, réduisant la cadence de déclenchement de certaines cellules ciliées et produisant même des sons à l'occasion.

Une autre preuve de cela est que l'acuité audio est réduite par la maladie ou la prise de certains médicaments, et le mal d'oreille peut être causé par l'incapacité des muscles de l'oreille interne à protéger le tympan. **[8]**

II.5 Les techniques de prétraitement du signal parole

La redondance et la variabilité du signal de parole ne permettent pas son utilisation directe dans un système ASR, d'où la nécessité d'une analyse acoustique. Cette analyse également appelée extraction de caractéristiques est l'ensemble de méthodes utilisées pour extraire de l'information à partir de ce signal tout en maintenant le pouvoir discriminant du signal et en réduisant sa dimensionnalité.

Dans ce qui suit, les analyses, largement utilisées pour la création du vecteur contenant les caractéristiques discriminatifs du signal de parole sont détaillées, tout en mettant l'évidence en particulier sur l'analyse des coefficients cepstraux à fréquence Mel (MFCC), vu sa large utilisation dans le domaine de l'ASR.

Parmi ces différentes analyses, nous présentons :

- Analyse par codage prédictif linéaire ;
- Analyse par coefficients cepstraux de prédiction linéaire;
- Analyse par la prédiction linéaire perceptuelle;
- Analyse spectrale relative ;
- Analyse par coefficients cepstraux à échelle de Mel. **[9]**

II.5.1 Analyse par codage prédictif linéaire

Le codage prédictif linéaire est un modèle paramétrique du signal de parole pris du modèle humain de la production de la parole. LPC a été largement usité en particulier dans le traitement du signal vocal depuis son introduction à la fin des années 1960.

Cette technique s'appuie particulièrement sur l'hypothèse que la parole peut être modélisée par un processus linéaire, qui cherche à prédire le signal **s(n)** à un instant **n** à partir des **p** échantillons précédents. Néanmoins, la parole étant un processus non parfaitement linéaire, la somme pondérée du signal sur **p** pas de temps engendre une erreur qui doit être corrigée par l'introduction du terme **e(n)** (erreur de prédiction d'ordre **p**) illustrée par la formule **II.1**.

Le codage par prédiction linéaire s'admet alors à déterminer les coefficients ak (représentant les coefficients de prédiction linéaire d'ordre p) qui minimisent l'erreur e(n), en utilisant un ensemble de signaux constituant les données d'apprentissage. Le choix de l'ordre p est un accord entre précision spectrale, temps de calcul et mémoire de calcul.

$$s(n) = \sum_{K=1}^{P} a_k \, . \, s(n - k) + e(n). \qquad \textbf{II.1}$$

Où

s (n - k) représente les k échantillons précédents.

Les différentes étapes qui définissent la technique LPC sont expliquées ci dessous et illustrées par la Figure **II.3.**

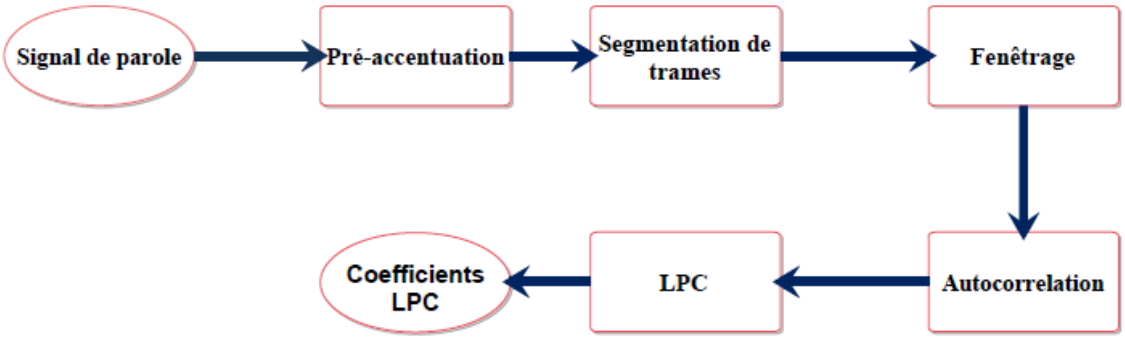


Figure II.3 : Schéma fonctionnel de la technique LPC

- **Pré-accentuation** :

  Augmentation systématique des amplitudes relatives de certaines composantes spectrales du signal de parole pour mieux couvrir le bruit de fond.

- **Segmentation** :

  Dans cette étape, le signal est scindé en trames composées de M échantillons, chacune de 20 à 40 ms avec un chevauchement standard de 10 ms entre chaque deux trame adjacente.

- **Fenêtrage** :

  Les trames résultantes sont multipliées par la fenêtre de Hamming afin d'adoucir la transition du signal sur les bords de la trame.

- **Calcul des LPCs** :

  Dans cette étape, la méthode d'auto-corrélation est appliquée sur les trames fenêtrées. **[9]**

II.5.2 Analyse par coefficients cepstraux de prédiction linéaire

La technique des coefficients cepstraux de prédiction linéaire est principalement dérivée de l'analyse prédictive linéaire, où les paramètres LPCCs (les p premiers coefficients cepstraux Cn) sont calculés en utilisant la formule **II.2** suivante :

c1 = a1,

$$c_n = \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) a_k \; c_{n-k} + a_n \qquad 1 < n \leq p. \qquad \textbf{II.2}$$

Où

c_i : le coefficient cepstre d'ordre i ;

ai : le coefficient prédicteur linéaire.

LPCC a été créée pour répondre aux limites de LPC en délivrant des coefficients moins corrélés à la place de ceux fortement corrélés fournis par LPC. La Figure **II.4** illustre le schéma fonctionnel de l'extraction des LPCCs.

Il à noter que toutes les étapes de calcul des LPCs sont maintenues dans le calcul des LPCCs. Les caractéristiques LPCC sont calculées en introduisant les coefficients cepstraux dans les paramètres LPC. **[9]**

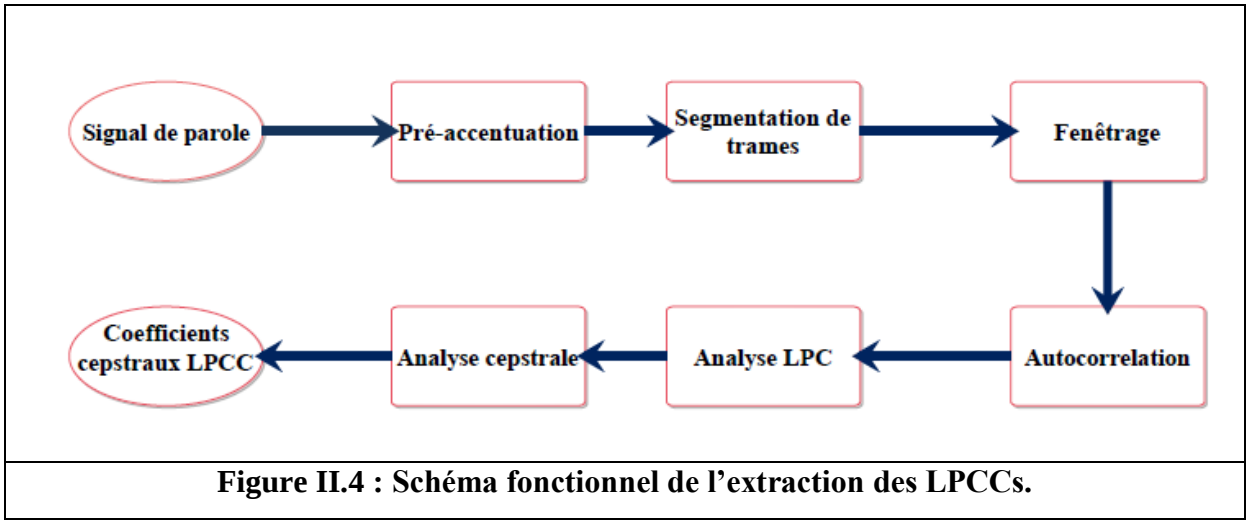


Figure II.4 : Schéma fonctionnel de l'extraction des LPCCs.

II.5.3 Analyse par la prédiction linéaire perceptuelle PLP

Une autre analyse, appelée la prédiction linéaire perceptuelle (PLP), utilise le même principe de base que la technique LPC vu qu'elle utilise le spectre à court terme du signal. En outre, PLP utilise les connaissances issues de la psycho-acoustique du système auditif humain pour optimiser l'utilisation du spectre. Cet aspect a rendu cette analyse plus proche de l'audition humaine ce qui lui a permis de fournir des paramètres plus robustes. Cette technique représente une alternative pour la technique des coefficients cepstraux (MFCC).Le processus de calcul des coefficients PLP peut être décrit par la Figure **II.5**. **[9]**

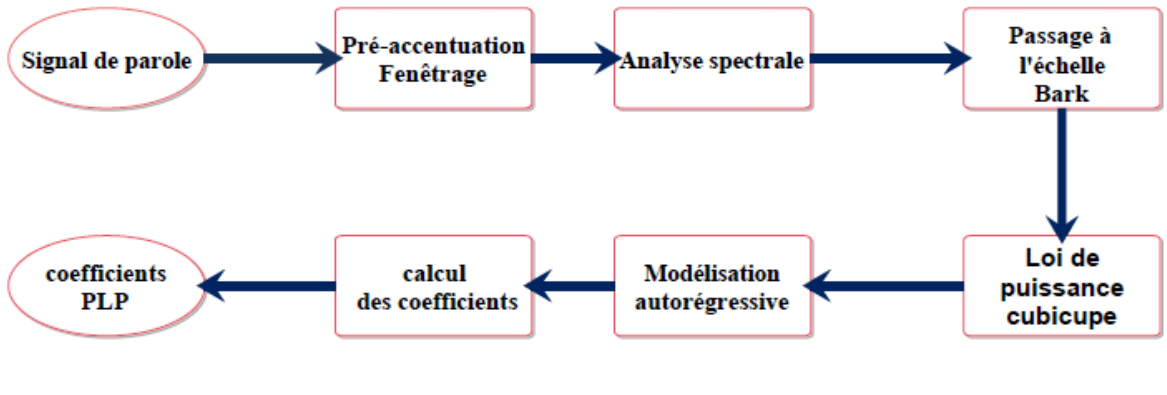


Figure II.5 : Calcul des coefficients PLP.

II.5.4 Analyse spectrale relative

L'analyse spectrale relative est dérivée de l'analyse PLP. La conception de base est de supprimer les variations trop lentes ou trop rapides par filtrage sur le spectre d'amplitude, dans le but de ne retenir que les variations liées au signal produit par l'être humain. Les articulateurs ne peuvent pas bouger trop rapidement, alors si les caractéristiques changent trop rapidement, elles ne sont peut-être pas issues de la parole. De plus, si les caractéristiques changent trop lentement, ce changement ne sera pas perçu.

L'analyse RASTA est souvent combinée avec l'analyse PLP, donnant RASTA PLP, ceci dans le but d'augmenter la robustesse des paramètres utilisés par les systèmes ASR. **[9]**

II.5.6 Analyse des Coefficients cepstraux à échelle Mel

L'analyse des Coefficients cepstraux à échelle Mel (MFCC) a été présentée par Davis et Mermelstein en 1980. Elle a été exploitée depuis cette date avec succès dans les différentes tâches d'ASR.

Cette analyse s'appuie sur un calcul de coefficients cepstraux à échelle Mel qui se rapproche de la perception fréquentielle de l'oreille humaine. L'idée principale est de moyenner le spectre dans des bandes de fréquence correspondant au filtrage effectué par la membrane basilaire.

Afin d'extraire les caractéristiques des signaux de parole, l'analyse MFCC utilise un certain nombre de bancs de filtres Mel, de 15 à 24 filtres triangulaires espacés linéairement jusqu'à 1 kHz et logarithmiquement au-dessus de 1 kHz, pour lisser et capturer les différentes caractéristiques linguistiques du spectre du signal de parole.

Les différentes phases de la technique MFCC sont expliquées ci-dessous et illustrées dans la Figure **II.6**.

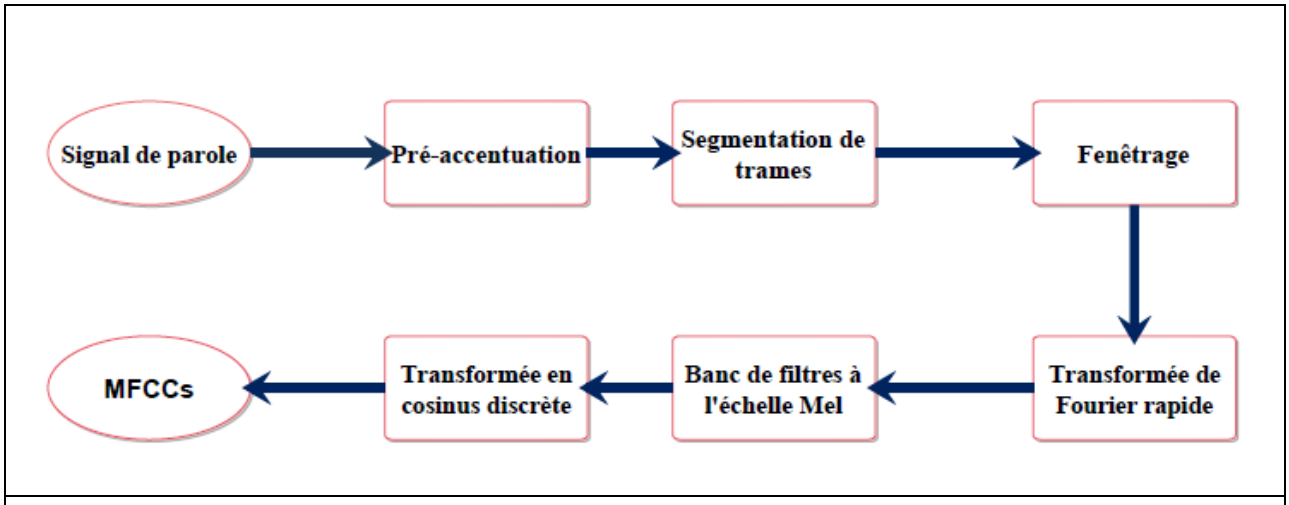


Figure II.6 : Schéma fonctionnel des étapes de calcul des MFCCs

II.5.6.a Pré-accentuation

Les recherches menées dans le domaine de l'ASR, ont montré que les segments vocaux tels que les voyelles ont plus d'énergie aux basses fréquences que dans les hautes fréquences ; ceci est causé par la nature de l'impulsion glottale.

La pré-accentuation permet d'amplifier l'énergie des hautes fréquences pour les rendre plus appropriées pour le modèle de reconnaissance. Elle est effectuée en faisant passer le signal échantillonné d'origine x[n] dans un filtre passe-haut de premier ordre dont l'équation est la suivante :

$$y[n] = x[n] - ax[n - 1]. \qquad \text{II.3}$$

Où

y[n] désigne le signal de sortie.

x[n] est la séquence d'échantillons obtenue à partir du signal temporel continu x(t).

a : facteur de pré-accentuation prenant une valeur comprise dans [0.9, 1.0].

II.5.6.b Segmentation en trames

Le signal de parole représente un processus aléatoire non-stationnaire à long terme, toutefois il est considéré stationnaire dans des fenêtres temporelles d'analyse de l'ordre de 20 à 30 ms. A cet effet, après la phase de pré-accentuation, pour avoir des caractéristiques acoustiques stables, le signal de parole doit alors être divisé en un certain nombre de trames et examiné sur chacune d'elles où la propriété de stationnarité à court terme est vérifiée avec, généralement, un chevauchement de fenêtres de 10 ms. L'intérêt de ce chevauchement est l'obtention d'une continuité temporelle des caractéristiques. De ce fait, à partir de chaque trame, un ensemble de paramètres est dérivé pour former le vecteur caractéristique.

II.5.6.c Fenêtrage

Le fenêtrage de la trame est utilisé pour minimiser les discontinuités du signal au début et à la fin de chaque trame. Dans le domaine du traitement de la parole, plusieurs types de fenêtres de pondération, également appelées fenêtres d'observation, sont définies dans la littérature et employées, où chaque fenêtre peut être décrite par trois paramètres : sa largeur appelée taille de trame, le décalage entre les fenêtres successives appelé décalage de trame ou chevauchement et la forme de la fenêtre.

Dans le domaine temporel, le fenêtrage consiste à multiplier la valeur du signal s[n] par la valeur de la fenêtre w[n] à l'instant n et fournit le signal de sortie y[n].

$$y[n] = s[n] * w[n]. \qquad \text{II.4}$$

Où

y[n] est le signal de sortie à l'instant n.

s[n] est le signal d'entrée à l'instant n.

w[n] représente la fenêtre de pondération employée.

Dans la littérature, de nombreuses fenêtres de pondération sont définies, parmi lesquelles : la fenêtre rectangulaire, la fenêtre de Hann, la fenêtre de Blackman et la fenêtre de Hamming.

Fenêtre rectangulaire : est la forme de fenêtre la plus simple. L'équation **II.5** et la Figure **II.7** décrivent cette fenêtre.

$$W_n = \begin{cases} 1 & 0 \le n \le N-1, \\ 0 & sinon \end{cases} \qquad \text{II.5}$$

Où

N est le nombre d'échantillons dans la fenêtre.

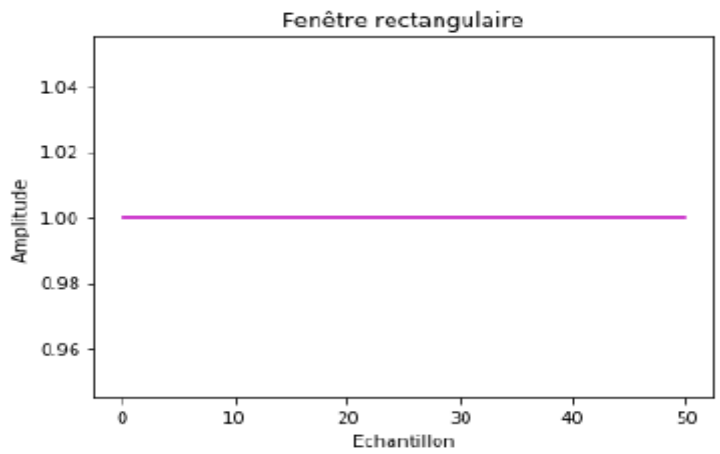


Figure II.7 : Fenêtre rectangulaire.

La fenêtre rectangulaire présente l'inconvénient de la coupure brutale à ses limites dues aux discontinuités qui posent des problèmes lors de l'analyse de Fourier.

Fenêtre de Hann : cette fenêtre est décrite par l'équation **II.6** et illustrée dans la Figure **II.8**. La fenêtre de Hann réduit complètement les données à zéro au début et à la fin de la trame.

$$W[n]=\begin{cases} 0.5-0.5cos\frac{2\pi n}{N} & 0\le n\le N-1, \\ 0 & sinon \end{cases} \qquad \textbf{II.6}$$

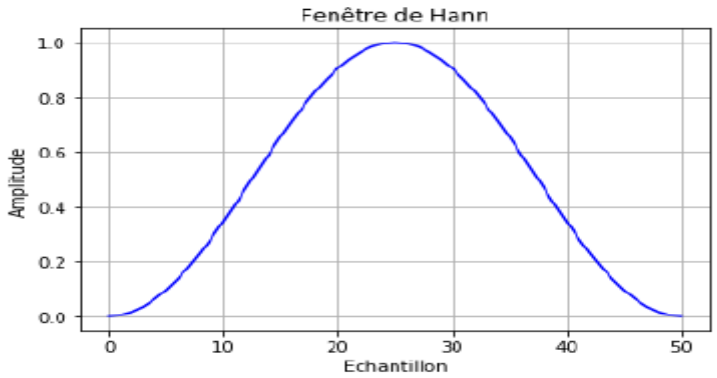


Figure II.8 : Fenêtre de Hann.

Fenêtre de Blackman : un autre type de fenêtre est la fenêtre de Blackman illustrée par l'équation **II.7** et la Figure **II.9** respectivement.

$$\mathbf{W(n)=}\begin{cases} \mathbf{0.42-0.50}cos\frac{2\pi n}{N}+\mathbf{0.08}cos\frac{2\pi n}{N} & \mathbf{0\le n\le N-1,} \\ \mathbf{0} & \boldsymbol{sinon} \end{cases} \qquad \textbf{II.7}$$

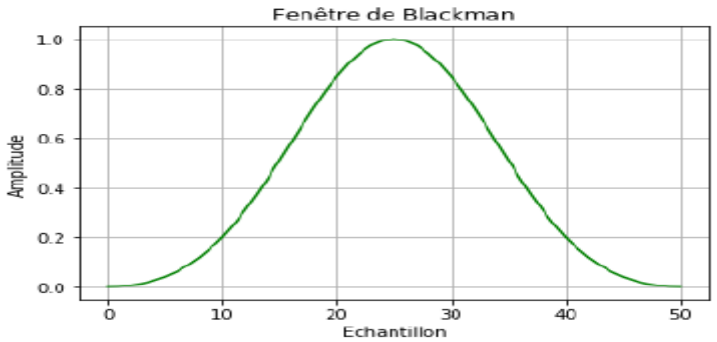


Figure II.9 : Fenêtre de Blackman.

Fenêtre de Hamming : une autre forme des fenêtres de pondération est la fenêtre de Hamming. Elle a été proposée par Richard Wesley Hamming et peut être considérée comme une forme optimisée de la fenêtre Hann. Elle a pour rôle l'atténuation de la valeur du signal à zéro lorsqu'elle s'approche des bords de la fenêtre pour éviter les discontinuités.
L'équation **II.8** et la Figure **II.10** illustrent respectivement la fenêtre de Hamming.

$$W(n)=\begin{cases} 0.54 - 0.46cos\frac{2\pi n}{N} & 0 \leq n \leq N-1, \\ 0 & sinon \end{cases} \qquad \text{II.8}$$

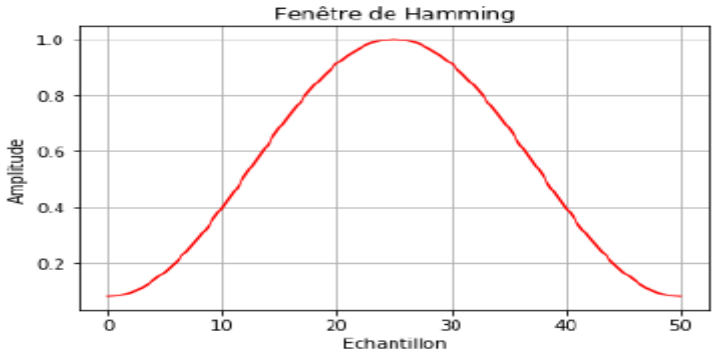


Figure II.10 : Fenêtre de Hamming.

Pour le domaine de l'ASR, différentes études ont montré que la forme de fenêtre la plus appropriée est la fenêtre de Hamming. Celle-ci est adoptée par la technique MFCC.

II.5.6.d Transformée de Fourier discrète

Après avoir effectué un fenêtrage pour atténuer la discontinuité du signal en début et en fin de trame, la phase suivante consiste à appliquer la transformée de Fourier discrète (Discrete Fourier Transform : DFT) qui se calcule via l'algorithme de la transformée de Fourier rapide (Fast Fourier Transform : FFT).

Cet algorithme largement utilisé pour évaluer la fréquence du spectre du signal permet de convertir chaque trame fenêtrée de N échantillons du domaine temporel au domaine fréquentiel. La DFT est utilisée pour extraire les informations spectrales de chaque fenêtre du signal d'entrée. A la sortie de la DFT, une valeur complexe représentant l'amplitude et la

phase de chaque composante fréquentielle de la trame est obtenue. Le calcul de la DFT est donné par l'équation **II.9**.

$$\mathbf{X_K}=\sum_{n=0}^{N-1} x_n e^{-2\pi ikn/N} \qquad \mathbf{K=0,\ldots, N-1} \qquad \mathbf{II.9}$$

Où

$\mathbf{X_k}$ est la sortie DFT.

N est le nombre d'échantillons dans la trame.

II.5.6.e Banc de filtres à l'échelle Mel et Log

L'échelle Mel a été inspirée pour la première fois par Stevens et Volkman en 1937. Elle modélise la membrane basilaire et redistribue les fréquences en fonction de la fréquence perçue.

Le banc de filtres à l'échelle Mel est constitué d'une série de filtres passe-bande de forme triangulaire qui se chevauchent dont l'objectif est la réduction de la taille des caractéristiques impliquées. Ce type de structure de filtre est largement utilisé pour la modélisation spectrale auditive dans le domaine du traitement de la parole en particulier dans le cadre du calcul des coefficients cepstraux de fréquence Mel. Chaque réponse de filtre commence à une valeur d'amplitude nulle à l'extrémité inférieure et augmente linéairement jusqu'à la fréquence centrale, puis décroît linéairement à zéro à son extrémité supérieure.

Les filtres sont disposés de telle sorte que le premier filtre commence à une fréquence nulle et se termine à la fréquence centrale du filtre suivant.

Ensuite, le deuxième filtre commence à la fréquence centrale du filtre précédent et se termine à la fréquence centrale du filtre suivant, etc. La Figure **II.11** illustre la forme générale du banc de filtres à l'échelle Mel.

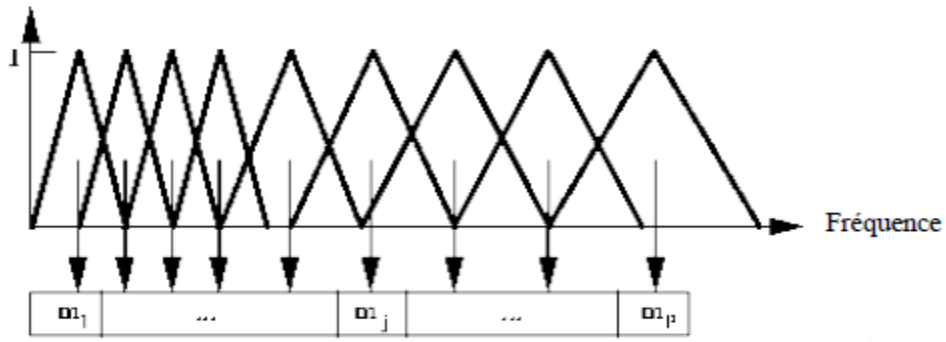


Figure II.11 : Banc de filtres à l'échelle Mel.

Les bancs de filtres à l'échelle Mel convertissent la puissance du spectre obtenu à partir de la transformation FFT sur l'échelle Mel en utilisant l'équation **II.10**.

$$\mathbf{Mel(f) = 2595 * \log_{10}(1 + f/700)} \qquad \mathbf{II.10}$$

Où

f représente la fréquence dans l'échelle linéaire et Mel(f) est celle perçue.

La relation entre la fréquence en Hertz et celle de l'échelle Mel est linéaire en dessous de 1000 Hz et logarithmique au dessus de 1000 Hz comme illustré dans la Figure **II.12**.

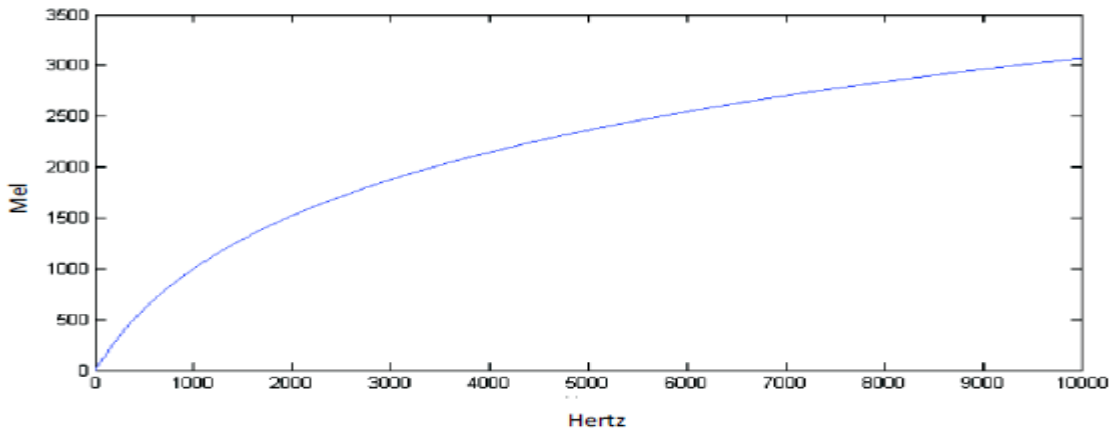


Figure II.12 : Relation entre la fréquence en Hertz et en échelle Mel.

Enfin, l'utilisation de l'opérateur Log rend les estimations des coefficients moins sensibles aux variations d'entrée, telles que les variations dues au rapprochement ou à l'éloignement de la bouche du haut-parleur du microphone.

Les coefficients log du banc de filtres à l'échelle Mel (Filter bank : FB) peuvent être calculés à partir des sorties des filtres par l'équation **II.11**:

$$\mathbf{S(m)= 20\ log_{10}}\left(\sum_{k=0}^{N-1}|X(K)|H(K)\right), \quad 0< m < M \qquad \mathbf{II.11}$$

Où,

M est le nombre de filtres à l'échelle Mel de 20 à 40.

X(k) est la FFT de la trame.

H(k) est la fonction de transfert du filtre Mel.

Les coefficients obtenus à la sortie des filtres peuvent être utilisés directement pour la reconnaissance de la parole, néanmoins, d'autres coefficients plus discriminatifs, plus robustes au bruit et bien particulièrement dé corrélés entre eux sont privilégiés obtenus en utilisant la transformation en cosinus discrète.

II.5.6.f Transformée en cosinus discrète

La transformation en cosinus discrète (Discrete Cosine Transform : DCT) est une transformation mathématique linéaire ayant la capacité de générer des coefficients décorrélés et de concentrer la majeure partie de l'énergie du signal dans un nombre réduit de coefficients. La DCT convertit le signal du domaine fréquentiel au domaine temporel, avec la possibilité de le reconvertir dans le domaine fréquentiel en utilisant la DCT inverse (Inverse Discrete Cosine Transform : IDCT). Les coefficients c_n sont calculés par l'équation **II.12** :

$$\mathbf{C(n)=}\sum_{m=0}^{N-1} S(m)\ \mathbf{cos}\left(\pi n\left(m-\frac{1}{2}\right)/M\right), \qquad \mathbf{0\leq n\leq M} \qquad \mathbf{II.12}$$

Où

C(n) : les coefficients MFCC.

$\mathbf{S_m}$: spectre logarithmique.

N : le nombre d'échantillons dans chaque trame.

M : le nombre des bancs de filtres.

Cette transformation dé-corrèle les sorties du banc de filtres à échelle Mel, et les premiers coefficients sont concaténés pour former le vecteur de caractéristiques MFCCs comme le montre la Figure **II.13**. **[9]**

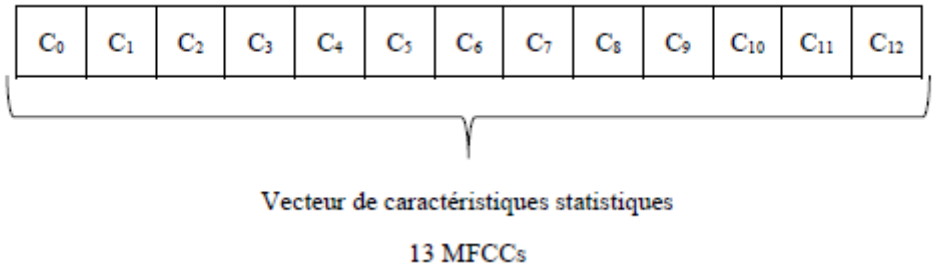


Figure II.13 : Les caractéristiques statiques MFCCs.

II.6 Conclusion

Dans cette partie, nous avons présenté les mécanismes de production et de perception de la parole chez les humains ainsi que les caractéristiques du signal vocal responsables des difficultés de la tâche de reconnaissance.

Par la suite, nous avons décrit les différentes méthodes de prétraitement du signal parole.

CHAPITRE 3:

LA TECHNIQUE DTW

III.1 Introduction

Un locuteur, même entraîné, ne peut prononcer plusieurs fois une même séquence vocale avec exactement le même rythme et la même durée. Les échelles temporelles de deux occurrences d'un même mot ne coïncident donc pas, et les formes acoustiques issues de l'étage de paramétrisation ne peuvent être simplement comparés point à point. On peut distinguer a priori deux sources de modification de l'échelle temporelle ; le changement de la vitesse d'élocution qui est représentable par une transformation linéaire de l'axe de temps, et, les variations dans le rythme de prononciation qui se traduisent par une transformation non linéaire. En fait, tout changement de vitesse d'élocution s'accompagne d'une transformation non linéaire de l'échelle temporelle, car les parties stables du signale sont plus affectés par les changements que les transitions.

Le problème de l'alignement temporel entre un mot inconnu et une référence peut être résolu de manière très efficace par un algorithme de comparaison dynamique qui va mettre en correspondance optimale les échelles temporelles des deux mots.

Dans ce chapitre on basée sur la technique déformation temporelle dynamique.

III.2 Généralités sur les systèmes de reconnaissance de la parole

Un système de reconnaissance de la parole est capable de transcrire une voix humaine en informations numériques, compréhensibles et reconnaissables par l'ordinateur.

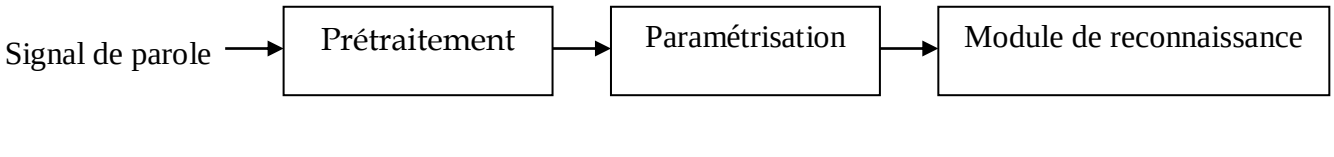


Figure III.1 : Schéma bloc d'un système de reconnaissance de la parole

Le signal de la parole est d'abord numérisé puis modélisé, généralement, sous une forme fréquentielle. Le module suivant, dans la chaîne de traitement acoustique (Figure III .1), est celui qui extrait des paramètres pertinents pour la reconnaissance de la parole. Ces paramètres sont, ensuite, envoyés à un module de reconnaissance qui identifie les sons présents dans le signal.

Parmi les méthodes utilisées dans le module de reconnaissance de la parole on peut citer la méthode d'alignement temporelle dynamique ou DTW. **[10]**

III.3 Structure d'un système de reconnaissance basé sur la méthode DTW

La structure d'un système de reconnaissance de la parole basé sur cette méthode est représentée par la figure suivante :

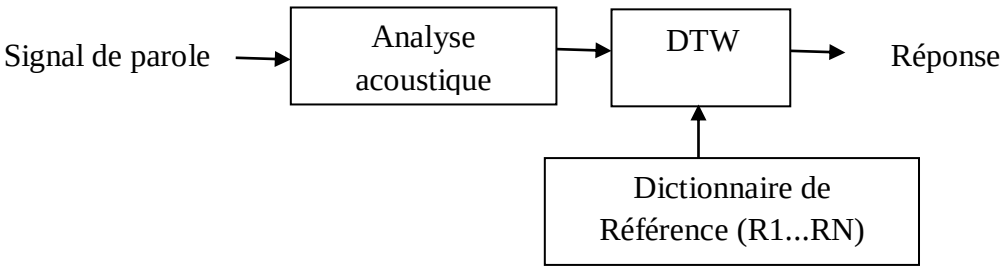


Figure III.2 : Système de reconnaissance de la parole basé sur DTW

Dans ce système on peut distinguer deux phases :

- Une phase d'apprentissage pour constituer le dictionnaire de référence (R1...RN).
- Une phase de reconnaissance durant laquelle le système va rechercher la référence la plus proche du mot à identifier par le calcul de la distance entre le mot inconnu et les mots de références. **[10]**

III.4 Définition

La déformation temporelle dynamique est un algorithme permettant de mesurer la similarité entre deux suites qui peuvent varier au cours du temps. Par exemple des similarités entre des pas dans des vidéos peuvent être détectées même si dans l'une ou l'autre des vidéos le sujet a marché plus rapidement ou plus lentement, ou encore si au cours de l'une ou l'autre le sujet a accéléré ou ralenti.

L'algorithme DTW a été exploité en vidéo, audio, graphique par ordinateur, bioinformatique,... et peut être appliqué dans toute situation où les données peuvent être

transformées en une représentation linéaire. Une application célèbre est l'application en reconnaissance de locuteur, où il est nécessaire de tenir compte de vitesses de locution très variables.

De façon générale, DTW est une méthode qui recherche un appariement optimal entre deux séries temporelles, sous certaines restrictions. Les séries temporelles sont déformées par transformation non-linéaire de la variable temporelle, pour déterminer une mesure de leur similarité, indépendamment de certaines transformations non-linéaires du temps. **[11]**

III.5 Les types de DTW [12]

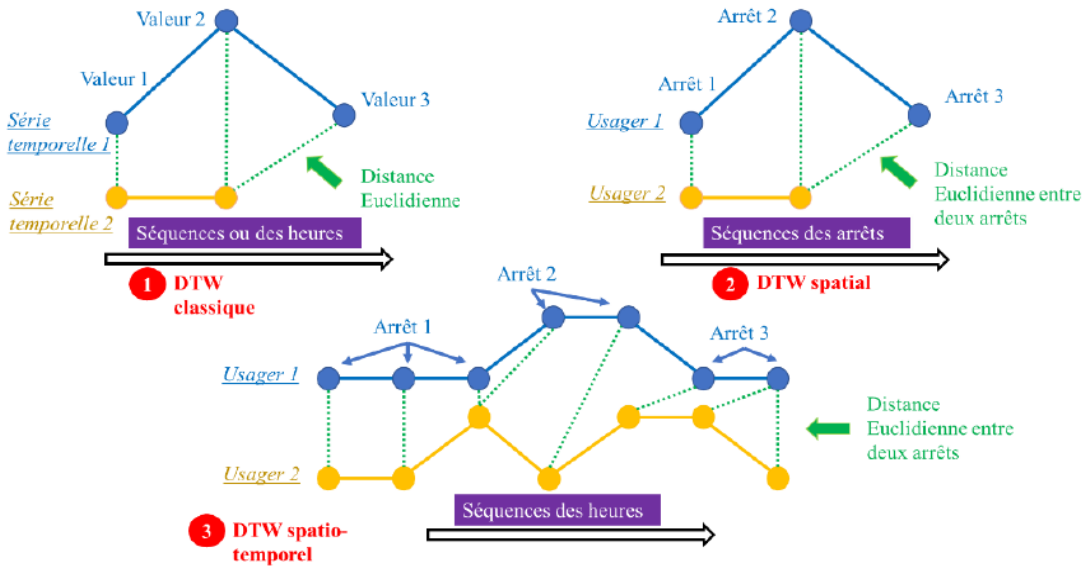


Figure III.3 : Comparaison des trois méthodes DTW

Conception	DTW Classique	DTW Spatial	DTW Spatio-temporelle
Objet à traiter	Séries temporelles	Trajectoires d'utilisateur dans le profil d'une journée (Séquence d'arrêts)	Localisation-heure d'utilisateur dans le profil d'une journée (Séquence d'arrêts à une heure donnée (moment))
Point	Point du temps (moment)	Arrêt	Arrêt à chaque moment donné
Séquence (Série temporelle)	Série temporelle	Séquence d'arrêt (inégale par rapport au temps)	Séquence d'arrêt (inégale par rapport au temps)
Distance entre le point de la grille	Peut être défini comme distance euclidienne, distance de Manhattan, etc.	Distance entre deux arrêts donnés (uniquement distance euclidienne)	Distance entre deux arrêts donnés (uniquement distance euclidienne)
Distance euclidienne	Au sens du temps (X : temps ; Y : valeur en x)	Au sens de la géographie (X: longitude; Y: latitude)	Au sens de la géographie (X: longitude; Y: latitude)

Tableau III.1 : Conception de trois types

III.6 Problèmes de DTW

L'alignement temporel, plus connu sous l'acronyme de DTW, est une méthode fondée sur un principe de comparaison d'un signal à analyser avec un ensemble de signaux stockés dans une base de référence. Le signal à analyser est comparé avec chacune des références et est classé en fonction de sa proximité avec une des références stockées. Le DTW est en fait une application au domaine de la reconnaissance de la parole de la méthode plus générale de la programmation dynamique. Elle peut ainsi être vue comme un problème de cheminement dans un graphe.

Ce type de méthode pose 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, variables, de reconnaissance du système d'alignement temporel. Chacun des signaux de référence est en effet stocké dans son état brut, sans compression d'aucune sorte.

Ce stockage permet de disposer d'un vocabulaire dont la taille correspond au nombre de mots du vocabulaire multiplié par le nombre de locuteurs et le nombre des éventuelles répétitions

des mots. Cette base de référence permet d'effectuer une mise en correspondance entre le signal stocké, d'une part, et sa retranscription symbolique d'autre part.

La taille de la base de référence est importante et implique une charge de travail non négligeable puisque la classification de chaque forme à analyser impose de la comparer à chaque forme de la base de référence. Donc, si la constitution de la base de référence est assez rapide et si le processus d'apprentissage est inexistant dans la méthode de l'alignement temporel, la phase d'utilisation nécessite une puissance de calcul non négligeable pour chaque référence atomique de signal à analyser. Le schéma de principe de la méthode est présenté dans la figure III.4.

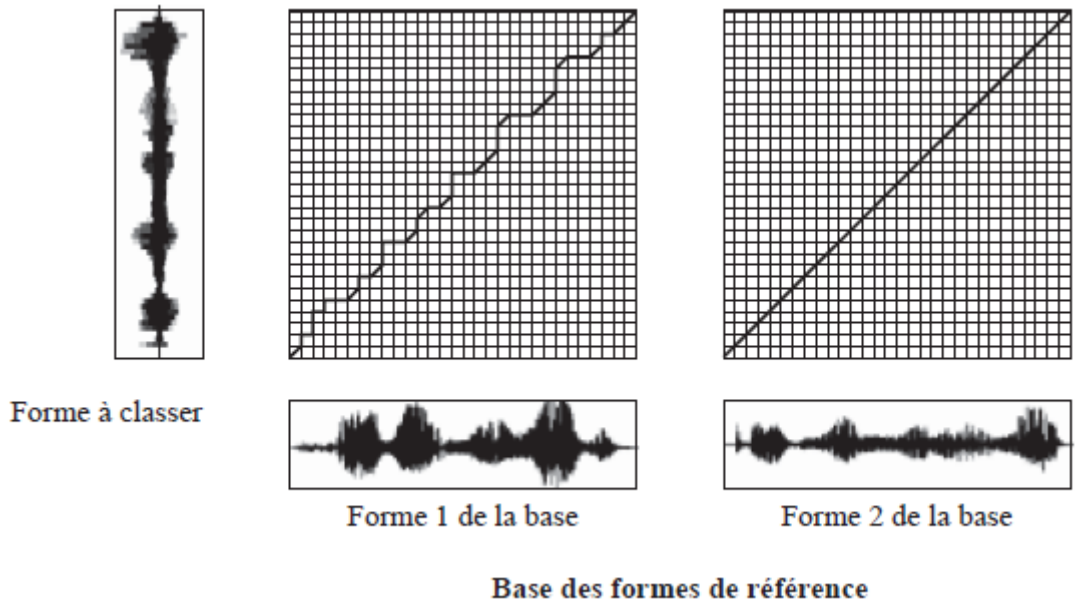


Figure III.4 : Visualisation du cheminement de l'alignement temporel pour des formes de la base de référence

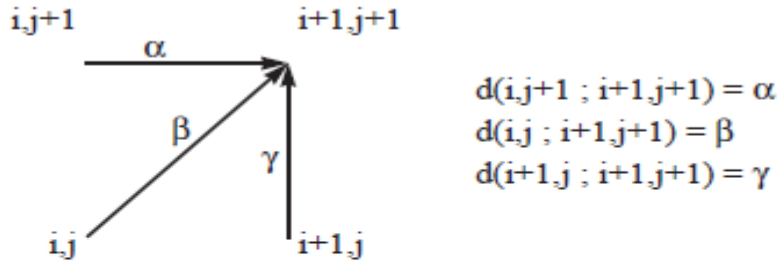
Comme la montre le schéma de la figure III.4, la forme choisie sera celle pour laquelle le chemin de mise en correspondance est le plus court, cette taille minimale marquant le peu de différences entre la forme à analyser et la forme de référence.

L'autre partie importante de l'alignement temporel est la définition de la fonction de recalage qui permet de calculer, selon certaines contraintes, la distance entre la forme à comparer et la forme de référence. La forme à analyser est mise en correspondance dans le plan temporel par l'algorithme d'alignement qui essaie de trouver le plus court chemin dans le graphe ainsi constitué. Cette fonction de mise en correspondance définit une valeur pour

chaque arc du graphe, ces valeurs favorisant l'axe médian qui correspond à une parfaite mise en relation de la forme à analyser et d'une forme de référence comme le montre la figure III.4.

- La fonction de recalage suit typiquement le schéma présenté dans la figure III.5.

La fonction *d(i,j)* est la fonction de calcul de la distance entre deux points successifs du graphe. Les valeurs α, β et γ permettent de définir une partie du comportement de la fonction *d* qui peut être soit symétrique (α = γ) soit asymétrique (α ≠ γ). Ce calcul de distance entre deux nœuds successifs du graphe n'est cependant pas suffisant pour calculer la longueur totale du chemin parcouru dans le graphe. Une fonction supplémentaire, *G*, calcule une longueur totale qui permettra, après le calcul de cette longueur des chemins sur toutes les formes de la base de référence, de savoir à quel mot du vocabulaire préenregistré correspond la forme à classer. D'un point de vue mathématique, *M* et *N* étant les longueurs respectives de la forme à classer et de la forme de référence, on cherche sur l'ensemble du corpus le *G (M, N)* minimal. Le calcul de cette fonction *G* répond au même principe que le principe général énoncé par Bellman pour la programmation dynamique : toute sous-partie du chemin optimal est lui-même un chemin optimal. Des exemples de fonctions *d* et *G* de calcul de distance, qui peuvent être bien plus complexes que la fonction de recalage présentée en figure III.5.



$$G(i+1,j+1) = \min \begin{bmatrix} G(i,j+1) + d(i,j+1 ; i+1,j+1) \\ G(i,j) + d(i,j ; i+1,j+1) \\ G(i+1,j) + d(i+1,j ; i+1,j+1) \end{bmatrix}$$

Figure III.5: Schéma typique d'une fonction de recalage en alignement temporel

Cette méthode de reconnaissance des formes est, initialement, bien adaptée à la reconnaissance de mots isolés mais des extensions ont été développées pour permettre de l'appliquer à la parole continue.

D'autres méthodes complémentaires ont par ailleurs été développées pour tenter de réduire la taille de la base des formes de référence par sélection optimale des formes à conserver. Ces méthodes reposent surtout sur une exploration statistique de la base des formes de référence et permettent d'obtenir une caractérisation des différents ensembles la constituant, ces ensembles correspondant aux différents symboles référencés dans la base. Une des techniques qu'il est possible d'employer pour ce faire est, par exemple, la méthode des plus proches voisins.

Certaines méthodes permettent de réduire ce temps de calcul à l'utilisation par apprentissage a priori de coefficients qui permettent de compacter la connaissance présente dans la base de référence qui devient ainsi un corpus d'apprentissage. **[13]**

III.7 Principe de fonctionnement

La normalisation temporelle peut être réalisée de manière optimale au cours de la phase de comparaison en ajustant les échelles temporelles des deux mots à comparer des transformations non linéaires. Cette technique désignée par le terme de comparaison dynamique ou alignement temporel dynamique est utilisée pour la RAP depuis 1968 .

Si A et B sont deux images acoustique de longueur I et J, on notera d (i, j) la distance entre les événements ai et bj. L'ajustement non linéaire de A et B est représenté par un chemin {C(k) = (n(k), m(k)), k = 1, K} dans [1, I] x [1, J] (figure III.6).

Pour correspondre à une réalité physique, les fonctions n(k) et m(k) doivent respecter certaines conditions :

- Conditions de frontière : C(1)= (1, 1) et C(K)= (I, J) ;
- Continuité : Supposons que C(K)= (i, j) et C (K+1) = (i', j') donc $i' - i \leq 1$ et $j' - j \leq 1$.

Cette contrainte est garante que C(K+1) doit être une cellule adjacente de C(K) (cela inclut la cellule adjacente diagonale) ;

- Monotonie : Supposons que C(K) = (i, j) et C (K+1)= (i', j') donc i' – i ≥ 0 et j' – j ≥ 0. Cette contrainte force le chemin à être monotone.

Ainsi, nous supposons que les seuls chemins valides arrivant au point (i , j) viennent des point (i-1,j), (i-1,j-1) ou (i, j-1). Cette contrainte est illustrée par la figure III.7 a (les figures III.7 b et III.7 c nous montrent d'autres solutions employées). En supposant que les frontières des mots sont correctement définies, on prendra C(1) = (1,1) et C(K) = (I, J).

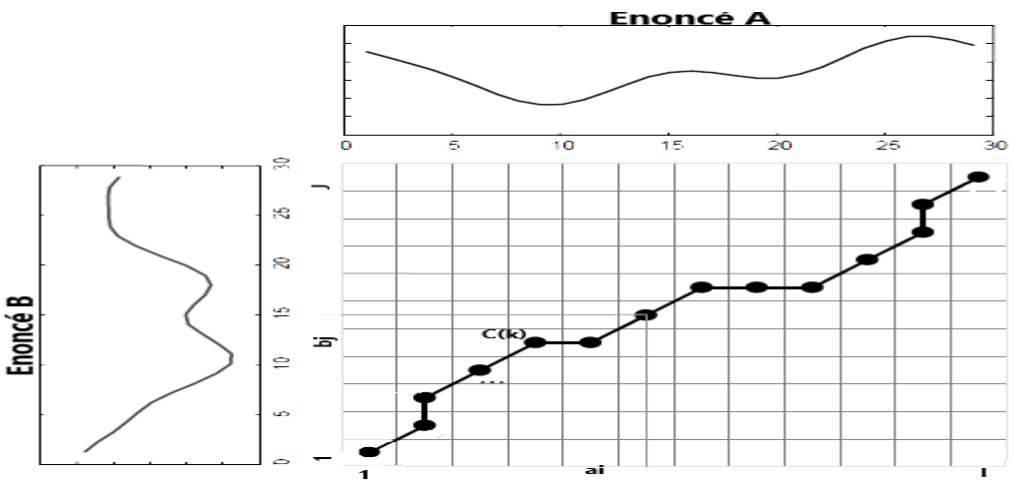


Figure III.6 : Comparaison dynamique - Mots isolés

La méthode consiste à choisir parmi tous les chemins physiquement possible, celui qui passe par les distances d (i, j) les plus petites, de sorte que la somme des distances le long du chemin soit minimale, la dissemblance entre A et B étant définie comme ainsi :

$$D(A,B) = min\left[\frac{\sum_{k=1}^{k} d(c(k)w(k))}{N(w)}\right] \qquad \text{III.1}$$

Où w(k) est un coefficient de pondération appliqué sur le k^{ième} segment du chemin C et N (w) est un coefficient de normalisation qui dépend de la fonction w. Pour évaluer D (A, B)

définie par l'équation (III.1), nous devons définir fonction de pondération w, et le coefficient de normalisation N (w).

Plusieurs types de fonction de pondération on été proposés ; en voici quatre exemples :

$W(k) = n(k) – n(k-1) + m(k) – m(k-1)$ (III.2)

$W(k) = n(k) – n(k-1)$ (III.3)

$W(k) = m(k) – m(k-1)$ (III.4)

$W(k) = \max \{n(k) – n(k-1), m(k) – m(k-1)\}$ (III.5)

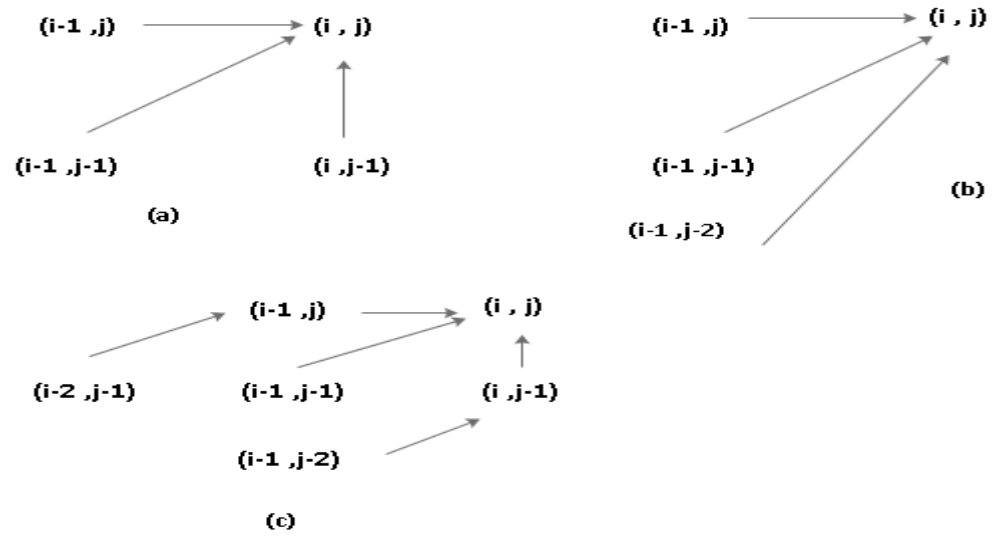


Figure III.7 : Contraintes locales

Le coefficient N (w) est généralement choisi tel que D(A, B) soit indépendante des longueurs de A et B, et, dans la mesure du possible pour la forme symétrique de w .

$$N(w)=\sum w(k)= I + J \qquad \text{(III.6)}$$

Le problème d'optimisation décrit par l'équation (III.1) peut être efficacement résolu par un algorithme de programmation dynamique. Ce dernier est grandement simplifié par le fait que N (w) est indépendant de C, car le problème se réduit alors à la minimisation du numérateur de (III.1). Ainsi si les seules transitions autorisées sont celles indiquées sur la figure III.7 a, et si l'on retient une fonction de pondération symétrique avec le

coefficient de normalisation défini par (III.6), on obtient la relation récurrente locale suivante :

$$g(i,j) = \min\begin{cases} g(i-1)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \\ g(j-1)+d(i,j) \end{cases} \qquad \text{III.7}$$

Où g (i, j) est la distance cumulée le long du chemin optimal allant du point (1,1) au point (i,j). g (i, j) est évaluée sur tout le domaine [1,I] x [1,J] qui est parcouru "colonne par colonne" ou "ligne par ligne" en partant du point (1,1). **[14]**

On obtient finalement :

$$D(A,B)=\frac{g(I,J)}{I+J} \qquad \text{III.8}$$

III.8 Identification de mots à l'aide de l'algorithme DTW

L'identification des mots peut être effectuée par comparaison directe des formes numériques des signaux ou par comparaison de spectrogrammes de signaux.

Le processus de comparaison dans les deux cas doit compenser à la fois la longueur différente des séquences et la nature non linéaire du son.

L'algorithme DTW parvient à résoudre ces problèmes en trouvant le chemin de distorsion correspondant aux distances optimales entre deux séries de longueurs différentes.

Il y a quelques particularités lorsque l'algorithme est appliqué aux deux cas :

1. Comparaison directe des formes numériques ou des signaux. Dans ce cas, pour chaque séquence numérique, une nouvelle séquence est créée, séquence dont les dimensions sont beaucoup plus petites. L'algorithme traite ces séquences. La séquence numérique peut avoir quelques milliers de valeurs numériques, tandis qu'une sous-séquence peut en avoir une centaine. La diminution du nombre de valeurs numériques peut être effectuée en supprimant celles entre les points extrêmes. Ce processus de réduction de la longueur de la séquence numérique ne doit pas modifier sa forme. Apparemment, le processus conduit à une diminution de la précision de reconnaissance. Cependant, compte tenu d'une augmentation de la vitesse, la précision est, en fait, augmentée en augmentant le nombre de mots dans le dictionnaire.

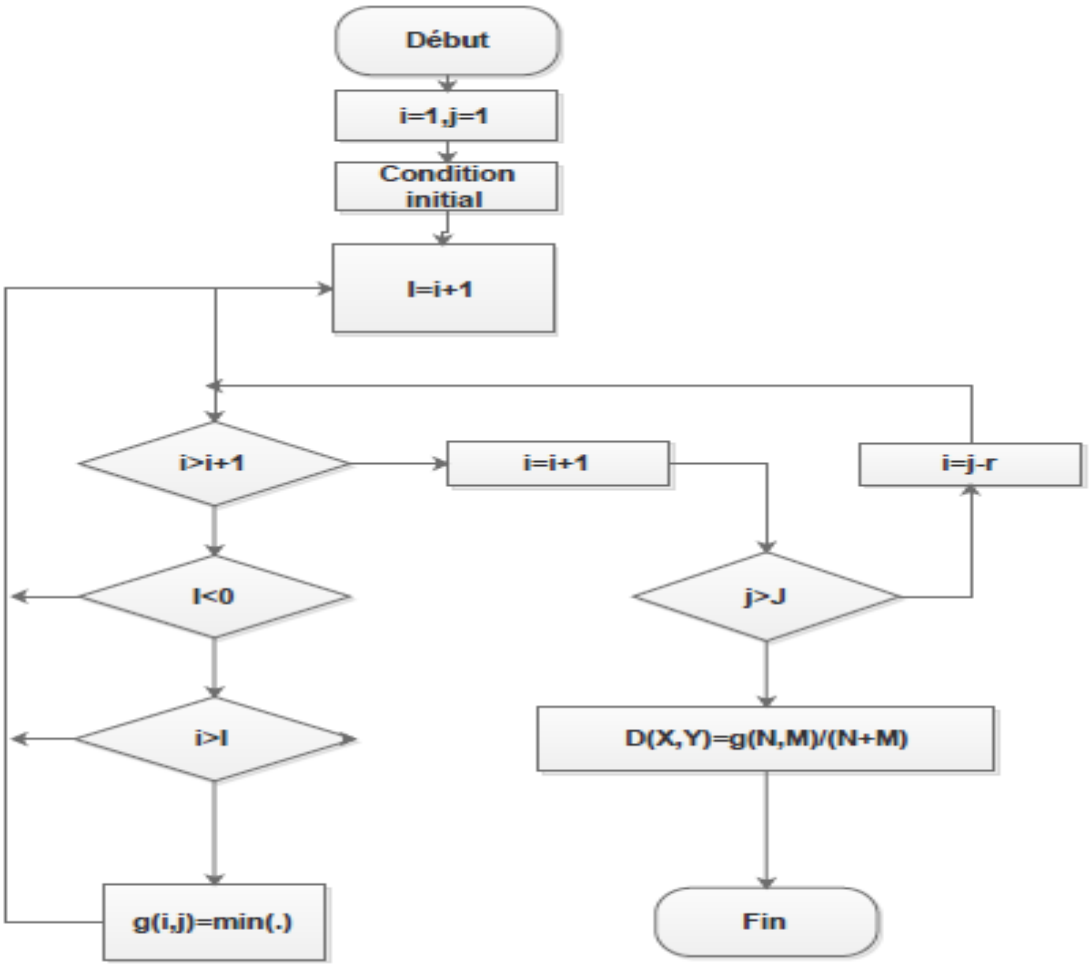
2. Représentations des spectrogrammes de signaux et application de l'algorithme DTW pour la comparaison de deux spectrogrammes.

La méthode consiste à diviser le signal numérique en un certain nombre de "fenêtres" (intervalles) qui se chevaucheront. Pour chaque fenêtre, des intervalles de nombres réels, (fréquences sonores) la transformée de Fourier rapide… seront calculés et seront stockés dans une matrice : le spectrogramme sonore.

Les paramètres seront les mêmes pour toutes les opérations de calcul de la : la longueur de la fenêtre, la longueur de la transformée de Fourier, la longueur de recouvrement pour deux fenêtres successives. La transformée de Fourier est symétrique par rapport au centre et les nombres complexes de la seconde moitié sont le nombre complexe conjugué des nombres symétriques de la première moitié. De ce fait, seules les valeurs de la première moitié peuvent être retenues, de sorte que le spectrogramme sera une matrice de nombres complexes, son nombre de lignes égalant la moitié de la longueur de la transformée de Fourier et son nombre de colonnes dépendant de la longueur du son.

La DTW sera appliquée sur une matrice de nombres réels résultant de la conjugaison des valeurs du spectrogramme, matrice dite matrice des énergies. **[15]**

III.9 Organigramme de la DTW [10]



III.10 Conclusion

Cet algorithme propose une nouvelle technique de formation simple pour préparer des modèles de référence pour les systèmes de reconnaissance vocale basés sur DTW. Il y a eu des progrès significatifs dans cette technique de formation. Ce DTW est destiné aux petites applications de vocabulaire. En utilisant DTW, des mots de différentes longueurs peuvent être reconnus. MFCC est la meilleure fonctionnalité pour la reconnaissance vocale et en utilisant cette fonctionnalité, nous pouvons obtenir une grande précision.

CHAPITRE 4:

RÉSULTATS ET INTERPRÉTATIONS

IV.1 Introduction

Après avoir présenté des notions de bases nécessaires et avoir situé notre recherche dans un cadre théorique, nous verrons dans ce chapitre la partie de mise en œuvre de notre système de reconnaissance automatique des chiffres arabes par l'utilisation du technique de déformation temporelle dynamique utilisée pour les signaux de parole, nous allons présenter aussi les étapes utilisées pour appliquer cette technique ainsi que les résultats obtenus .

IV.2 Outil de travail

- **MATLAB**

MATLAB est un langage de script permet d'appliquer les algorithmes DTW et MFCC qu'ils aider de traitement de parole.

- **Adobe Audition**

(Anciennement Cool Edit Pro) est un logiciel de traitement de données audio numériques, Il fut racheté en mai 2003 par Adobe.

IV.3 Spécification des besoins

Afin de nous assurer que nos méthodes sont bien développées et de mieux organiser notre travail, nous avons identifié des objectifs. Celles-ci peuvent être résumées dans les points suivants :

- notre principal objectif est de réaliser un système de reconnaissance automatique des chiffres prononcés en dix premiers chiffres arabes en mode mono-locuteur.
- on exposera les modules de bases qui le composent depuis l'acquisition du signale de parole jusqu'aux la décision prise. Nous étalons les algorithmes implémentés pour la méthode DTW, ainsi que les différents résultats obtenus, la tâche de reconnaissance est réalisée en passant par deux grandes étapes :
  - une analyse du signale de parole lu avec la méthode Cepstrale pour la génération des coefficients MFCC;
  - application de la méthode DTW pour la comparaison du signal avec la base sonore.

IV.4 But de notre Travail

En reconnaissance de la parole, l'étape d'extraction des caractéristiques, appelée communément l'étape d'analyse, peut-être réalisée de plusieurs manières. En effet, les vecteurs acoustiques sont généralement extraits à l'aide de méthodes temporelles comme le codage linéaire prédictif (LPC) ou de méthodes Cepstrales comme le codage MFCC, ainsi que le codage PLP (Perceptual Linear Predictive coding) qui est un exemple de l'application des connaissances du système auditif humain en reconnaissance de la parole.

L'extraction de caractéristiques est un élément clé pour la mise au point d'un système de reconnaissance. De nombreux travaux ont montré l'importance de cette étape.

Notre Objectif ici est d'élaborer un système de reconnaissance automatique des dix premiers chiffres arabes en mode mono-locuteur.

IV.5 La méthode de comparaison (DTW)

Notre système de reconnaissance de la parole est basé sur l'algorithme de DTW, il essaie d'évaluer la distance entre une observation et une liste de références (dictionnaire), ainsi la référence pour laquelle cette distance est minimale permet de dire de quel mot il s'agit.

L'évaluation de la distance entre deux signaux ne s'effectue pas avec les signaux eux-mêmes. Cela ferait beaucoup trop de calculs. Il s'agit donc dans un premier temps de trouver une meilleure représentation des signaux. C'est ici qu'intervient l'étape d'analyse; les coefficients MFCC.

Nous avons programmé la méthode de DTW en utilisant, pour la comparaison, des coefficients MFCC. La partie apprentissage concerne l'enregistrement des corpus de sons afin de concevoir le dictionnaire avec lequel seront comparés les signaux de test.

Des problèmes de reconnaissance peuvent apparaître selon les conditions dans lesquelles le signal à tester est enregistré. Si le mot est prononcé plus ou moins proche du microphone les taux de reconnaissance peuvent varier grandement. Cependant si l'utilisateur prononce le mot toujours à la même distance et avec la même intensité, les taux de reconnaissance sont très satisfaisants.

Il résulte néanmoins que la représentation à l'aide des coefficients MFCC fournit de meilleurs résultats, et supporte mieux les limitations exposées liées au problème de la capture du signal.

L'algorithme principal de la DTW comporte trois étapes suivantes :

- l'acquisition du fichier du son à tester ;
- l'extraction des coefficients MFCC ;
- la comparaison avec le dictionnaire des références.

IV.6 Les résultats sous MATLAB

- **L'enregistrement d'audio (test et référence) pour nbr = 1**

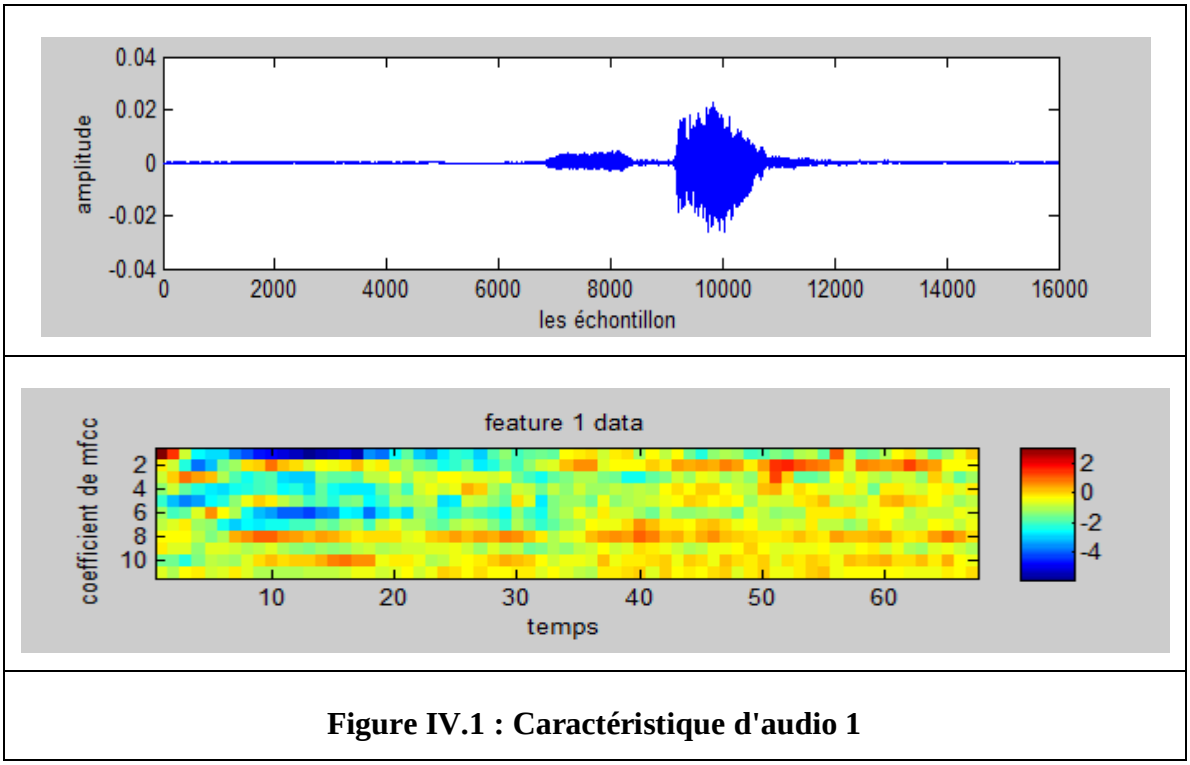


Figure IV.1 : Caractéristique d'audio 1

➢ La figure précédente représente le signal audio de test ;

Courbe 1 : Le signal de parole en domaine temporelle.

Courbe2 : Les coefficients MFCC par rapport au temps.

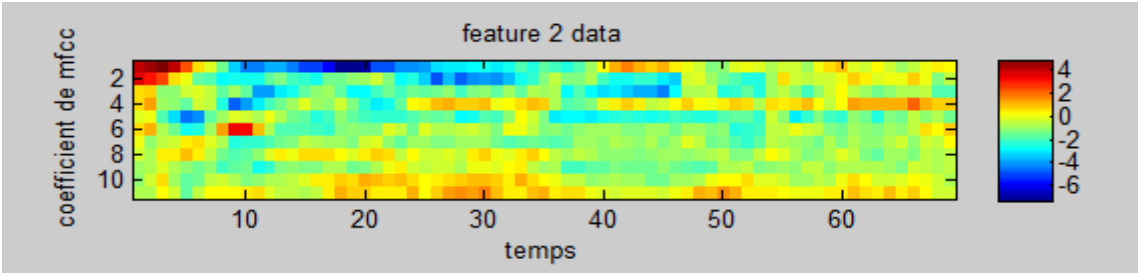


Figure IV.2 : Caractéristique d'audio 2

➢ La figure précédente représente le signal audio de référence du numéro 1.

- **Utilisation de la technique DTW (distance euclidienne) pour nbr=1**

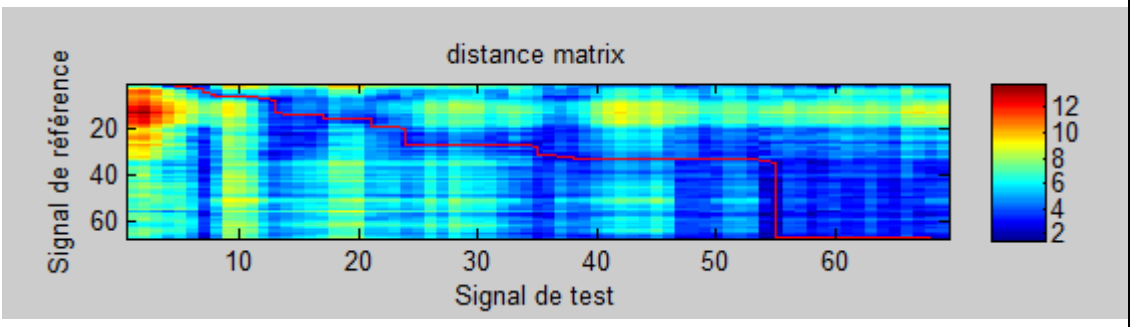


Figure IV.3 : Matrice de distance

Signal de référence	Signal de test	La distance
واحد	واحد	**3.092109508412095**
اثنان	واحد	3.306408582828009
ثلاثة	واحد	4.032643120924964
أربعة	واحد	4.888490363272837
خمسة	واحد	3.717593195200751
ستة	واحد	4.280430030730291
سبعة	واحد	3.818770647800282
ثمانية	واحد	3.691347646254057
تسعة	واحد	4.870097082049739
عشرة	واحد	3.460736705517696

Tableau IV.1 : Comparaison entre les distances

- Le tableau précédent montre que la valeur distance minimale entre le signal de test et les signaux références correspondant le nombre 1.
  - **L'enregistrement d'audio (test et référence) pour nbr=2**

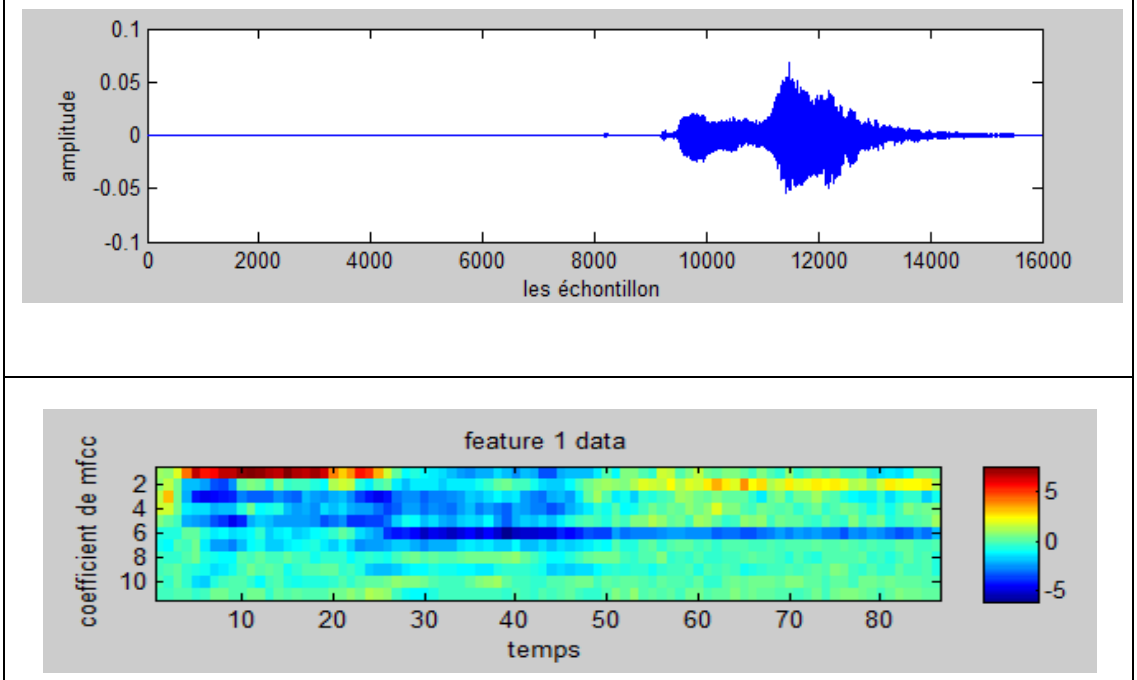


Figure IV.4: Caractéristique d'audio 1

- La figure précédente représente le signal audio de test ;

Courbe 1 : Le signal de parole en domaine temporelle.

Courbe2 : Les coefficients MFCC par rapport au temps.

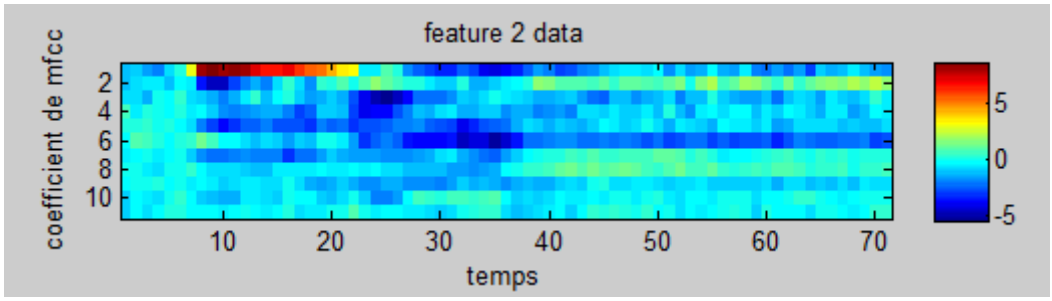


Figure IV.5: Caractéristique d'audio 2

- La figure précédente représente le signal audio de référence du numéro 2.

- **Utilisation de la technique DTW (distance euclidienne) pour nbr=2**

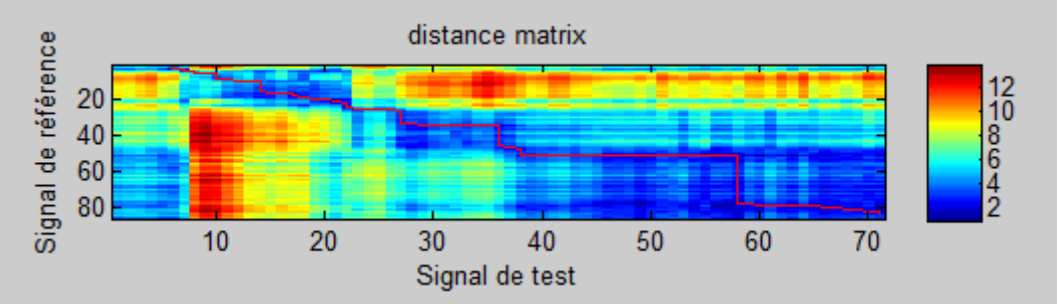


Figure IV.6 : Matrice de distance

Signal de référence	Signal de test	La distance
واحد	اثنان	4.567255085479711
اثنان	اثنان	**2.880081601349839**
ثلاثة	اثنان	4.775416079465079
أربعة	اثنان	7.326753724415371
خمسة	اثنان	5.412192976152908
ستة	اثنان	4.475464746078555
سبعة	اثنان	4.330032689016918
ثمانية	اثنان	4.300620532713574
تسعة	اثنان	5.442435575146797
عشرة	اثنان	5.025609662432221

Tableau IV.2 : Comparaison entre les distances

- Le tableau montre que la valeur distance minimale entre le signal de test et les signaux références est le nombre de référence 2.

- **L'enregistrement d'audio (test et référence) pour nbr = 3**

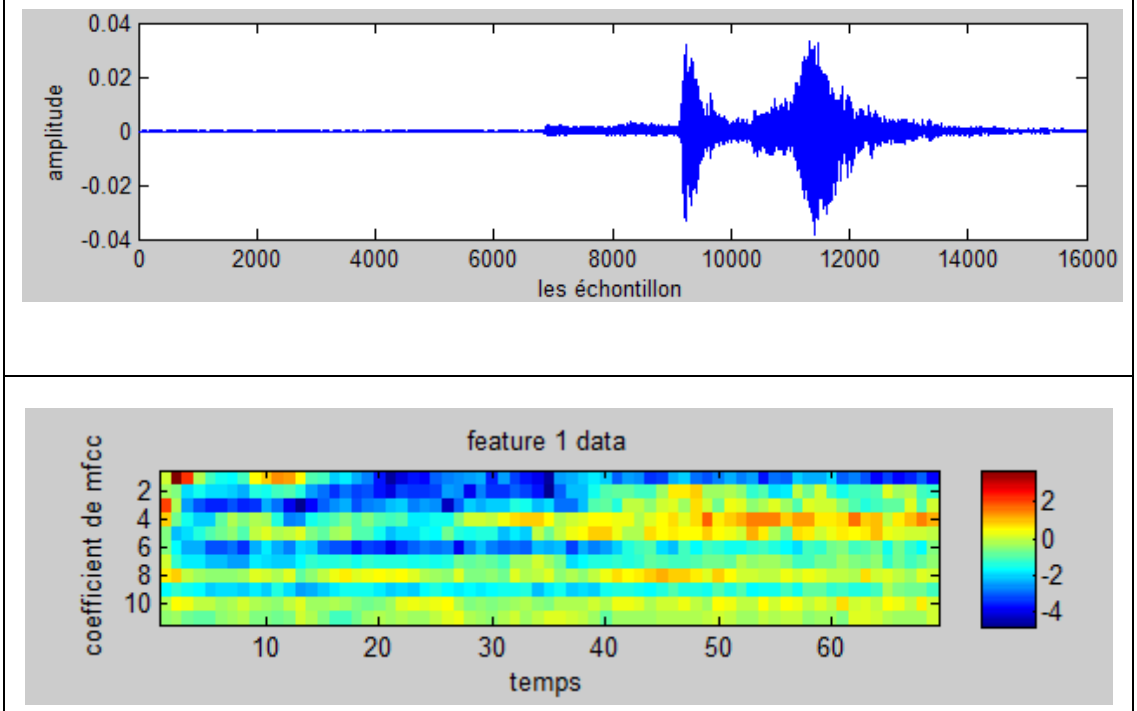


Figure IV.7: Caractéristique d'audio 1

➢ La figure précédente représente le signal audio de test ;

Courbe 1 : Le signal de parole en domaine temporelle.

Courbe2 : Les coefficients MFCC par rapport au temps.

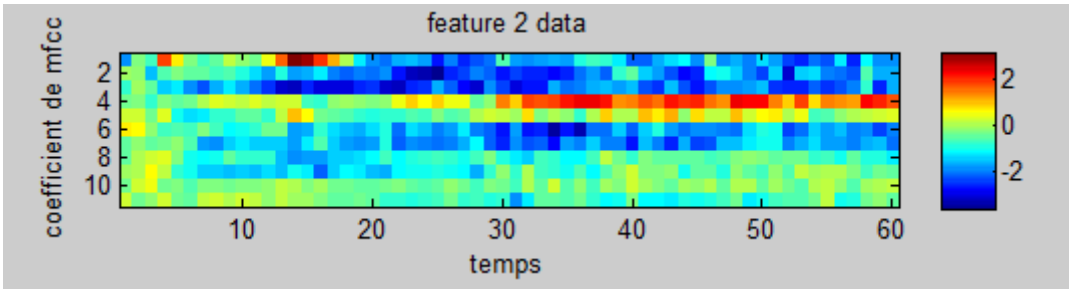


Figure IV.8 : Caractéristique d'audio 2

➢ La figure précédente représente le signal audio de référence du numéro 3.

- **Utilisation de la technique DTW (distance euclidienne) pour nbr=3**

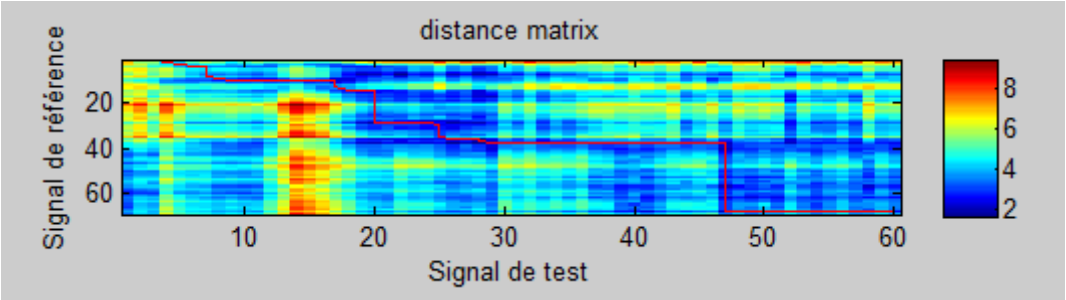


Figure IV.9 : Matrice de distance

Signal de référence	Signal de test	La distance
واحد	ثلاثة	3.882245430862250
اثنان	ثلاثة	3.721630522432190
ثلاثة	ثلاثة	**2.973933198521649**
أربعة	ثلاثة	5.214781153215549
خمسة	ثلاثة	4.005000906516835
ستة	ثلاثة	4.358412312204879
سبعة	ثلاثة	3.058722763760808
ثمانية	ثلاثة	4.349092971511587
تسعة	ثلاثة	4.949840826194915
عشرة	ثلاثة	3.672141683919171

Tableau IV.3 : Comparaison entre les distances

- Le tableau précédent montre que la valeur distance minimale entre le signal de test et les signaux références correspondant le nombre 3.

- **L'enregistrement d'audio (test et référence) pour nbr = 4**

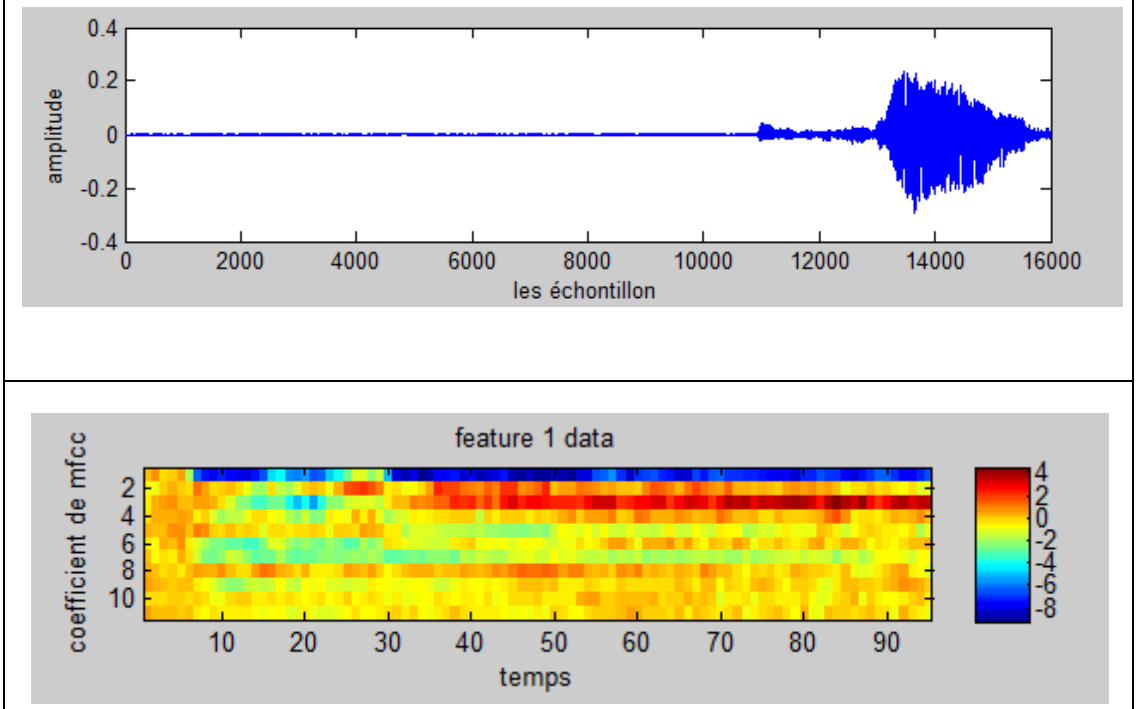


Figure IV.10: Caractéristique d'audio 1

➢ La figure précédente représente le signal audio de test ;

Courbe 1 : Le signal de parole en domaine temporelle.

Courbe2 : Les coefficients MFCC par rapport au temps.

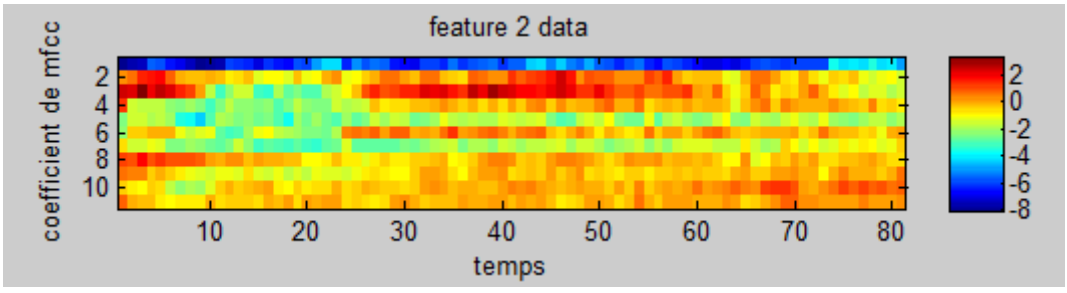


Figure IV.11: Caractéristique d'audio 2

➢ La figure précédente représente le signal audio de référence du numéro 4.

- **Utilisation de la technique DTW (distance euclidienne) pour nbr=4**

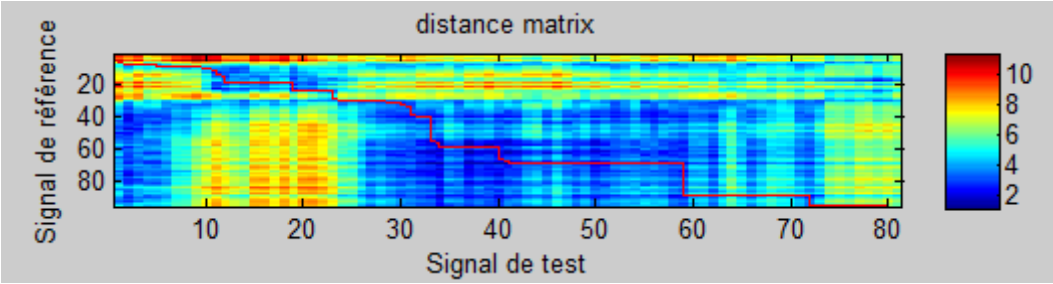


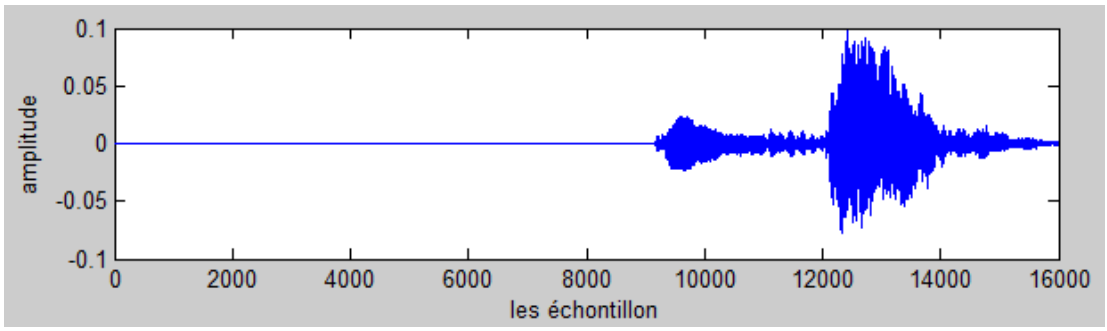
Figure IV.12 : Matrice de distance

Signal de référence	Signal de test	Minimum distance
واحد	أربعة	6.664096333890781
اثنان	أربعة	6.595515961583611
ثلاثة	أربعة	6.762667560451759
أربعة	أربعة	**3.526151229606711**
خمسة	أربعة	4.567200251761431
ستة	أربعة	7.657428700762020
سبعة	أربعة	3.690884354903544
ثمانية	أربعة	7.899929916751431
تسعة	أربعة	4.201140989964634
عشرة	أربعة	4.358176273652052

Tableau IV.4 : Comparaison entre les distances

- Le tableau précédent montre que la valeur distance minimale entre le signal de test et les signaux références correspondant le nombre 4.

- **L'enregistrement d'audio (test et référence) pour nbr = 5**



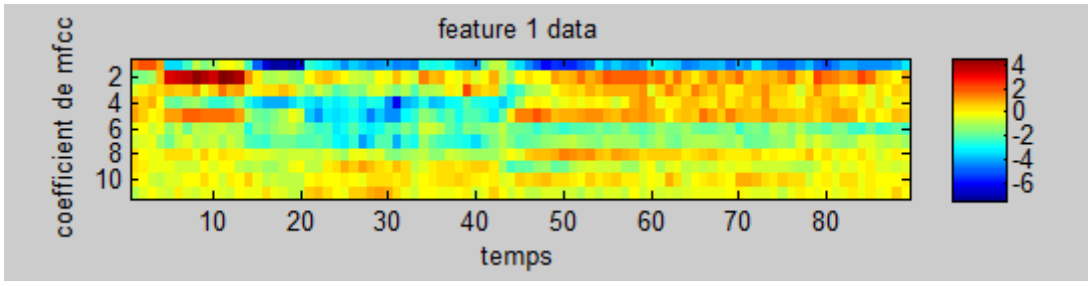


Figure IV.13: Caractéristique d'audio 1

➢ La figure précédente représente le signal audio de test ;

Courbe 1 : Le signal de parole en domaine temporelle.

Courbe2 : Les coefficients MFCC par rapport au temps.

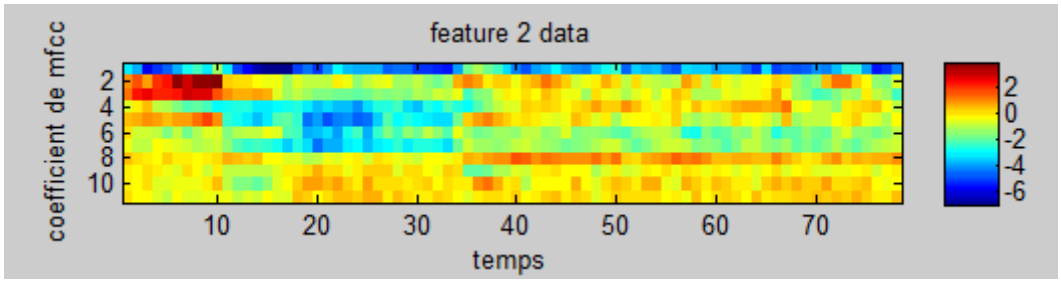


Figure IV.14: Caractéristique d'audio 2

➢ La figure précédente représente le signal audio de référence du numéro 5.

- **Utilisation de la technique DTW (distance euclidienne) pour nbr=5**

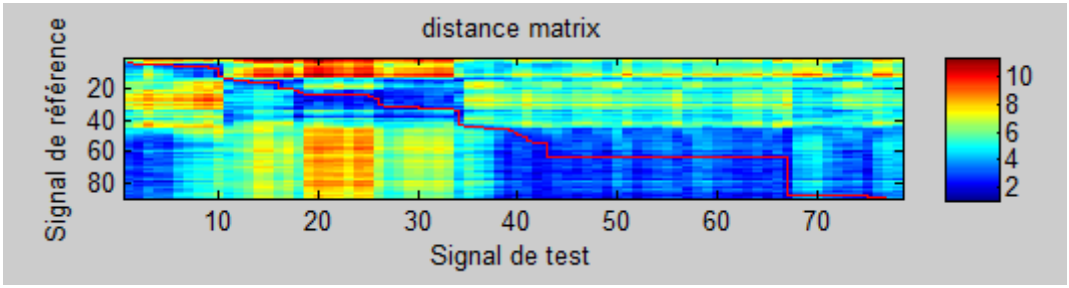


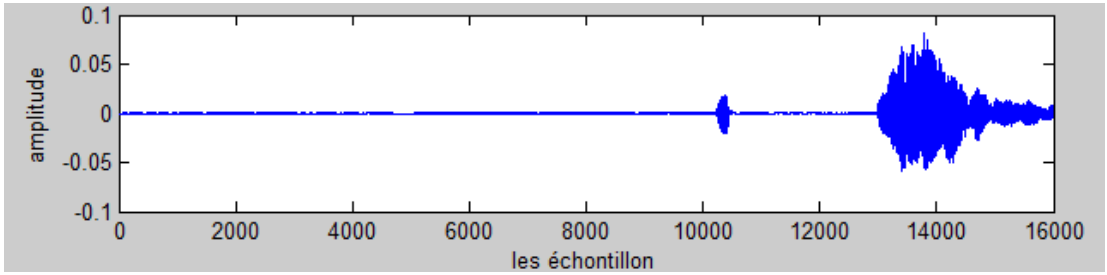
Figure IV.15 : Matrice de distance

Signal de référence	Signal de test	La distance
واحد	خمسة	5.168484611892199
اثنان	خمسة	4.566567479037812
ثلاثة	خمسة	4.975264878097213
أربعة	خمسة	4.203496512789303
خمسة	خمسة	**2.684867023439056**
ستة	خمسة	5.537082421773914
سبعة	خمسة	3.640450576403239
ثمانية	خمسة	5.654125769382355
تسعة	خمسة	4.364055874013143
عشرة	خمسة	3.123487885372305

Tableau IV.5 : Comparaison entre les distances

- Le tableau précédent montre que la valeur distance minimale entre le signal de test et les signaux références correspondant le nombre 5.

- **L'enregistrement d'audio (test et référence) pour nbr = 6**



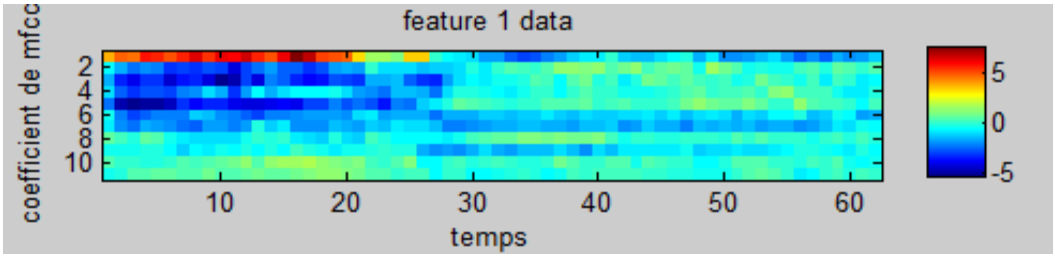


Figure IV.16: Caractéristique d'audio 1

➢ La figure précédente représente le signal audio de test ;

Courbe 1 : Le signal de parole en domaine temporelle.

Courbe2 : Les coefficients MFCC par rapport au temps.

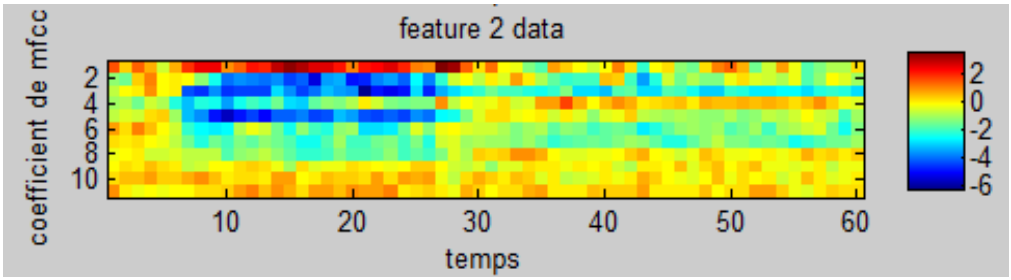


Figure IV.17: Caractéristique d'audio 2

➢ La figure précédente représente le signal audio de référence du numéro 6.

- **Utilisation de la technique DTW (distance euclidienne) pour nbr=6**

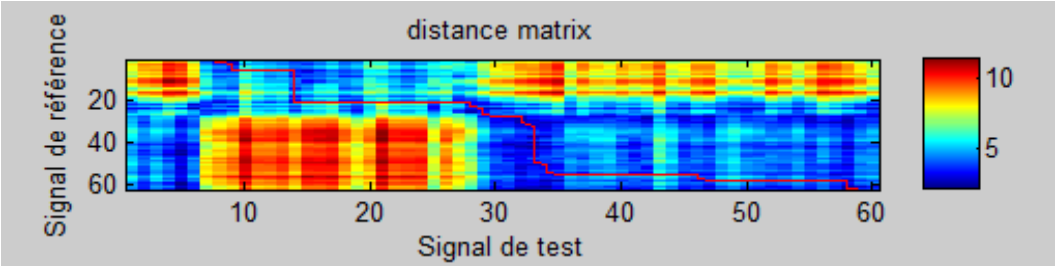


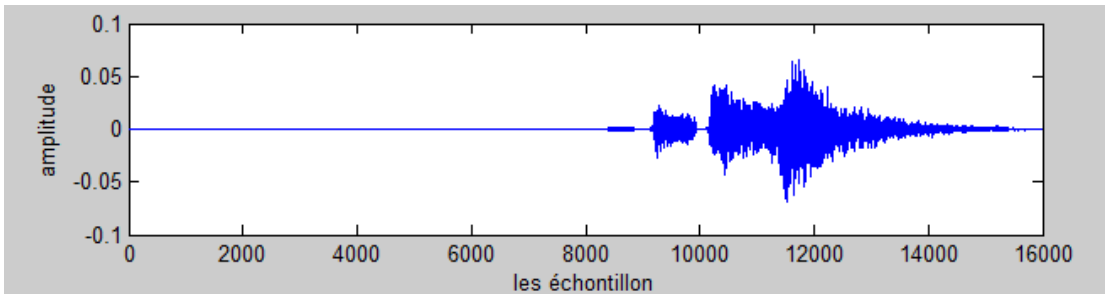
Figure IV.18 : Matrice de distance

Signal de référence	Signal de test	La distance
واحد	ستة	4.254472644841317
اثنان	ستة	3.639125505400430
ثلاثة	ستة	4.163135808590998
أربعة	ستة	6.760718032433901
خمسة	ستة	5.282799142659931
ستة	ستة	**3.600320018002890**
سبعة	ستة	3.849212503474624
ثمانية	ستة	4.124117347371747
تسعة	ستة	4.249145792791579
عشرة	ستة	4.543644689208943

Tableau IV.6 : Comparaison entre les distances

- Le tableau précédent montre que la valeur distance minimale entre le signal de test et les signaux références correspondant le nombre 6.

- **L'enregistrement d'audio (test et référence) pour nbr = 7**



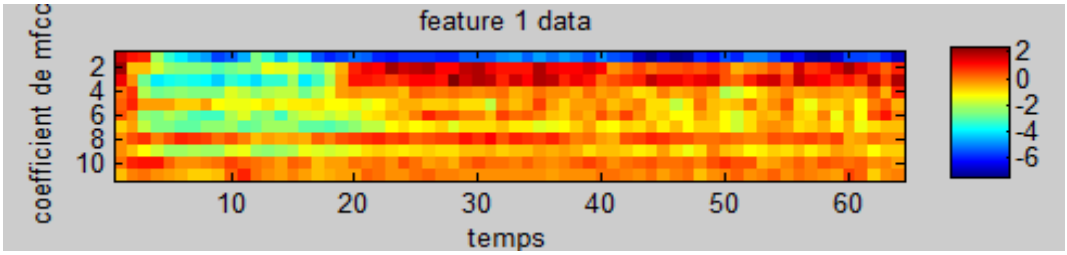


Figure IV.19: Caractéristique d'audio 1

➢ La figure précédente représente le signal audio de test ;

Courbe 1 : Le signal de parole en domaine temporelle.

Courbe2 : Les coefficients MFCC par rapport au temps.

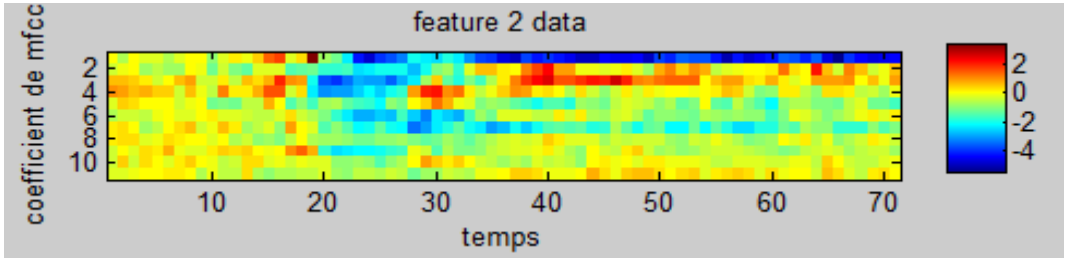


Figure IV.20: Caractéristique d'audio 2

➢ La figure précédente représente le signal audio de référence du numéro 7.

- **Utilisation de la technique DTW (distance euclidienne) pour nbr=7**

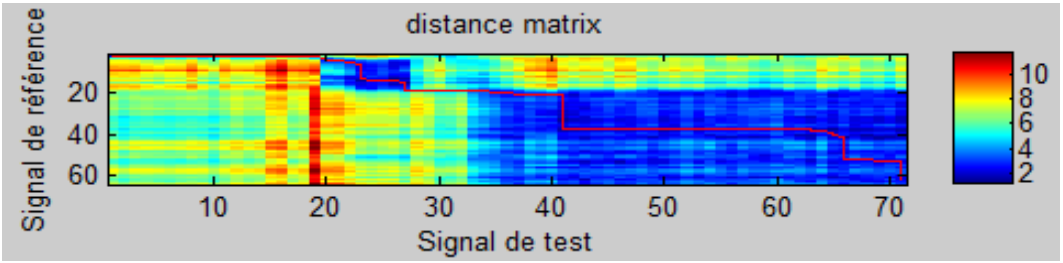


Figure IV.21 : Matrice de distance

Signal de référence	Signal de test	La distance
واحد	سبعة	5.259619908738626
اثنان	سبعة	4.844854887733671
ثلاثة	سبعة	5.270131346679301
أربعة	سبعة	3.416533925440303
خمسة	سبعة	3.895498925941796
ستة	سبعة	6.200800580491643
سبعة	سبعة	**2.583236709460974**
ثمانية	سبعة	6.518359268235096
تسعة	سبعة	3.441816346754972
عشرة	سبعة	3.603549805935532

Tableau IV.7 : Comparaison entre les distances

- Le tableau précédent montre que la valeur distance minimale entre le signal de test et les signaux références correspondant le nombre 7.

- **L'enregistrement d'audio (test et référence) pour nbr = 8**

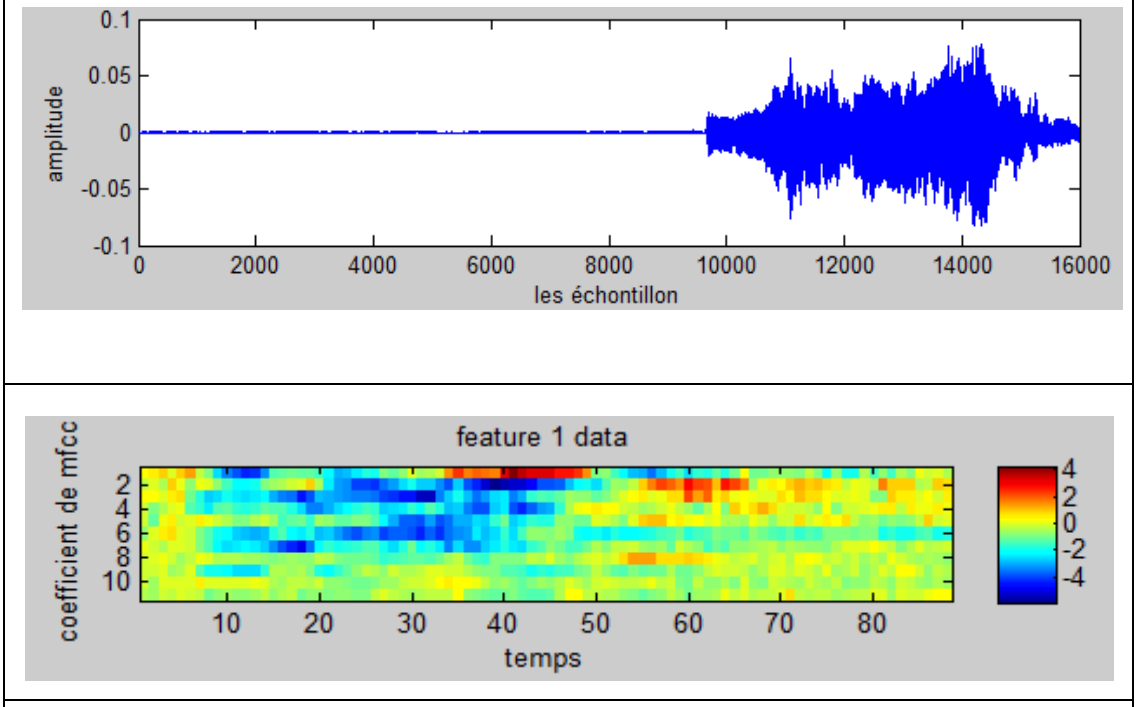


Figure IV.22: Caractéristique d'audio 1

➢ La figure précédente représente le signal audio de test ;

Courbe 1 : Le signal de parole en domaine temporelle.

Courbe2 : Les coefficients MFCC par rapport au temps.

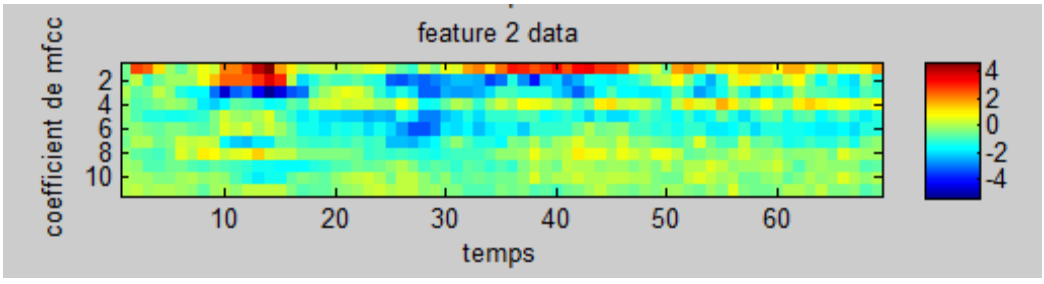


Figure IV.23: Caractéristique d'audio 2

➢ La figure précédente représente le signal audio de référence du numéro 8.

- **Utilisation de la technique DTW (distance euclidienne) pour nbr=8**

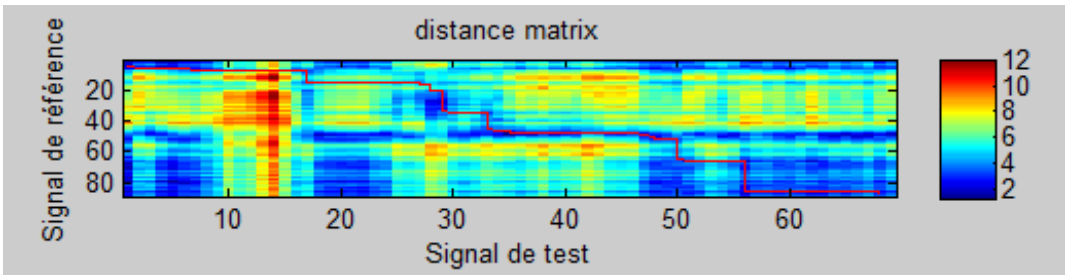


Figure IV.24 : Matrice de distance

Signal de référence	Signal de test	Minimum distance
واحد	ثمانية	4.082355683103309
اثنان	ثمانية	4.044630427857132
ثلاثة	ثمانية	3.810584550893914
أربعة	ثمانية	5.813980650500193
خمسة	ثمانية	4.580851640215461
ستة	ثمانية	4.028250954525059
سبعة	ثمانية	3.680203912649823
ثمانية	ثمانية	**3.453451236203882**
تسعة	ثمانية	5.843830831860378
عشرة	ثمانية	4.146577031359570

Tableau IV.8 : Comparaison entre les distances

- Le tableau précédent montre que la valeur distance minimale entre le signal de test et les signaux références correspondant le nombre 8.

- **L'enregistrement d'audio (test et référence) pour nbr = 9**

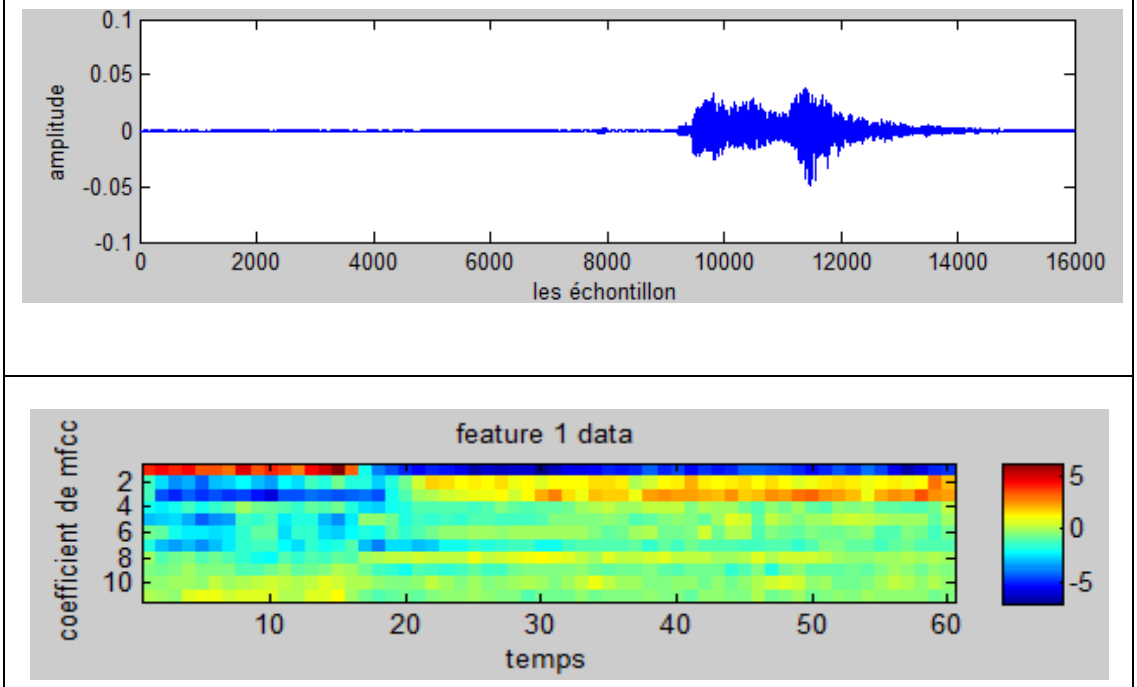


Figure IV.25: Caractéristique d'audio 1

➢ La figure précédente représente le signal audio de test ;

Courbe 1 : Le signal de parole en domaine temporelle.

Courbe2 : Les coefficients MFCC par rapport au temps.

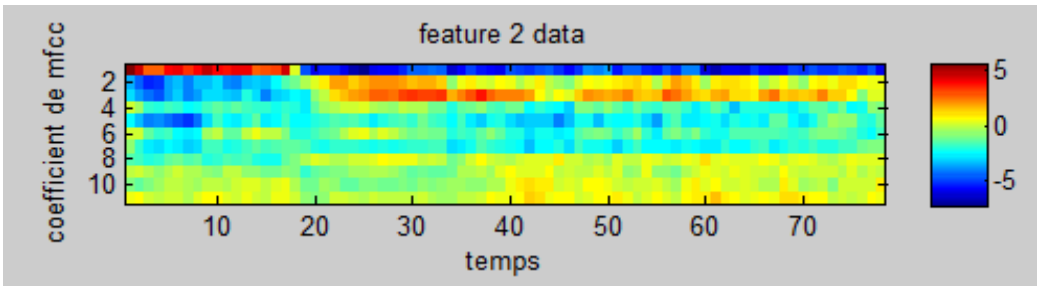


Figure IV.26: Caractéristique d'audio 2

➢ La figure précédente représente le signal audio de référence du numéro 9.

- **Utilisation de la technique DTW (distance euclidienne) pour nbr=9**

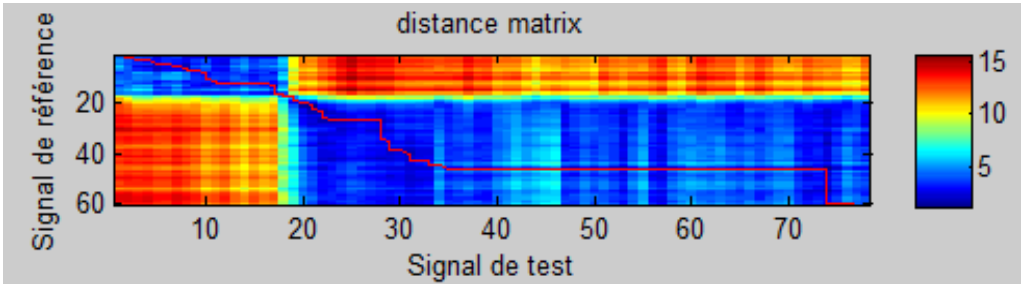


Figure IV.27 : Matrice de distance

Signal de référence	Signal de test	La distance
واحد	تسعة	5.738282904842016
اثنان	تسعة	4.823545495169489
ثلاثة	تسعة	5.679967452526049
أربعة	تسعة	4.754322443461990
خمسة	تسعة	4.903880666889912
ستة	تسعة	5.295507296872810
سبعة	تسعة	3.562628212473319
ثمانية	تسعة	6.172930074808471
تسعة	تسعة	**2.782505311094708**
عشرة	تسعة	4.307692971978156

Tableau IV.9 : Comparaison entre les distances

- Le tableau précédent montre que la valeur distance minimale entre le signal de test et les signaux références correspondant le nombre 9.

- **L'enregistrement d'audio (test et référence) pour nbr = 10**

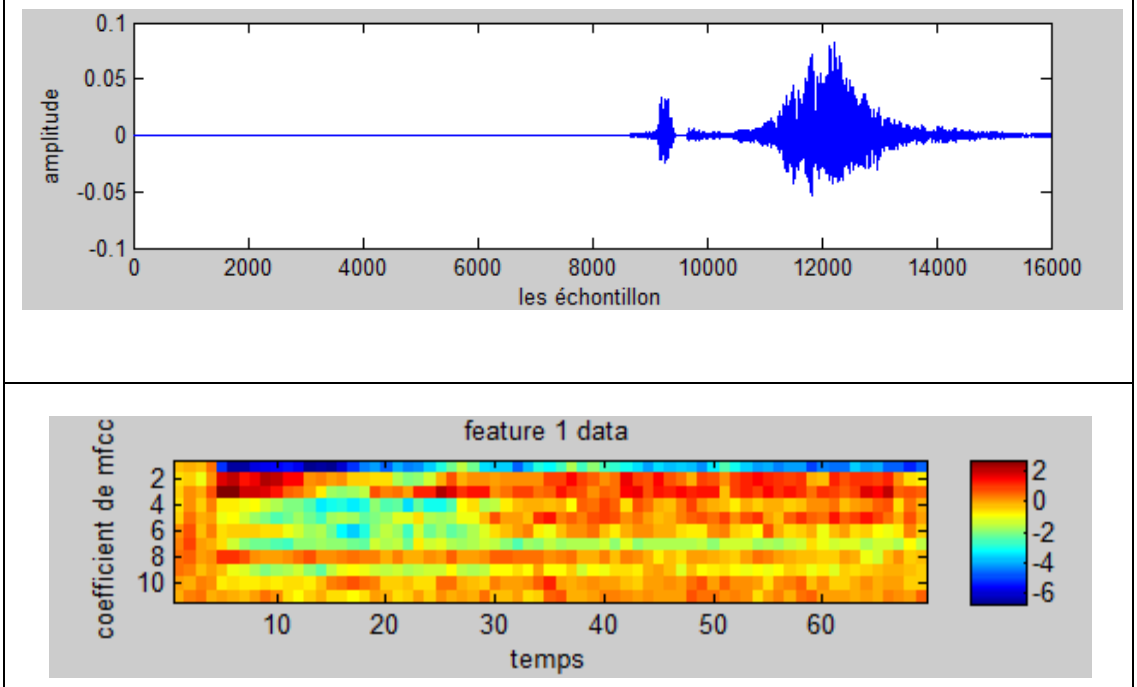


Figure IV.28: Caractéristique d'audio 1

➢ La figure précédente représente le signal audio de test ;

Courbe 1 : Le signal de parole en domaine temporelle.

Courbe2 : Les coefficients MFCC par rapport au temps.

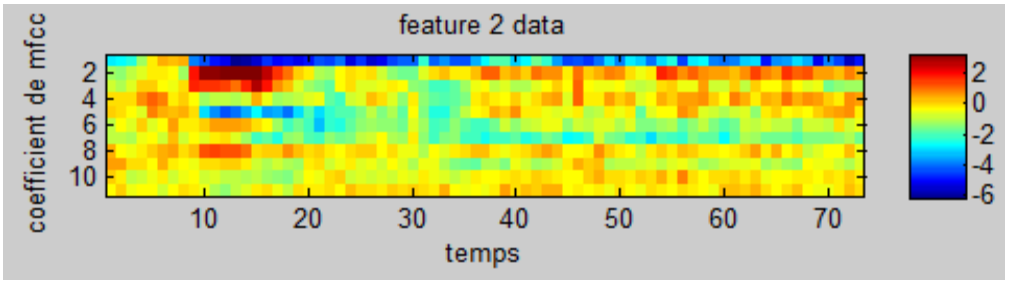


Figure IV.29: Caractéristique d'audio 2

➢ La figure précédente représente le signal audio de référence du numéro 10.

- **Utilisation de la technique DTW (distance euclidienne) pour nbr=10**

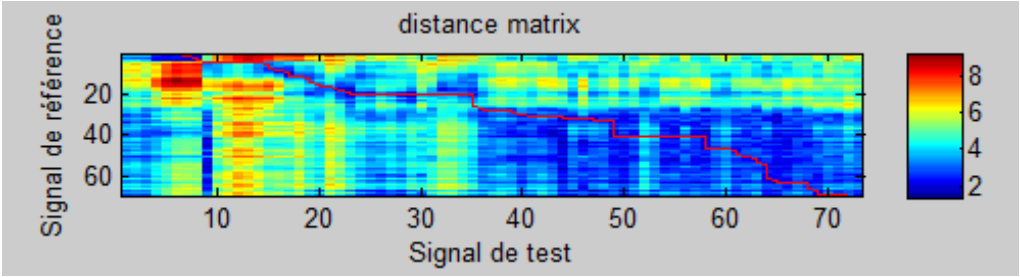


Figure IV.30 : Matrice de distance

Signal de référence	Signal de test	La distance
واحد	عشرة	4.885520091039913
اثنان	عشرة	4.688839252769768
ثلاثة	عشرة	4.739606209626393
أربعة	عشرة	3.288685716809667
خمسة	عشرة	2.916591664236267
ستة	عشرة	5.480813947840328
سبعة	عشرة	2.985867006692041
ثمانية	عشرة	5.417635737358191
تسعة	عشرة	3.875433222722234
عشرة	عشرة	**2.567912879975022**

Tableau IV.10: Comparaison entre les distances

- Le tableau précédent montre que la valeur distance minimale entre le signal de test et les signaux références correspondant le nombre 10.

Remarques

- Notre travail a permis d'extraire les caractéristiques de chaque signal parole enregistré à travers MATLAB et de produire une matrice de distance pour comparer chaque nombre avec le dictionnaire de référence (1 à 10) via le calcule de DTW.
- Plus le chemin dans la matrice est diagonal, plus le mot parlé est proche de l'une des références enregistrées.
- Chaque fois le tableau montre que le mot qui a été dit est celui qui a le moins cout (distance minimale).

IV.7 Conclusion

Dans ce chapitre, nous avons expliqué la partie conception et implémentation, dans laquelle nous présentons un système de reconnaissance des dix premiers chiffres de l'Arabe standard, dont la partie reconnaissance est basée sur la méthode DTW, et dans sa partie analyse cepstrale pour en extraire les coefficients MFCC, sur la base des tests que nous avons menés, ça permit effectivement la reconnaissance de mots isolés en mode mono locuteur des dix premiers chiffres de l'Arabe Standard.

Conclusion générale

La reconnaissance automatique de la parole 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 système (programme) de reconnaissance de la parole. Les résultats des études examinées sur la reconnaissance automatique de la parole 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 automatique de la parole, il restait une quantité surprenante de problèmes latéraux à comprendre et à résoudre. La partie traitement de signal de la reconnaissance vocale, Matériel d'enregistrement et environnement de travail (l'écho et le bruit), par exemple, ne doit pas être sous-estimée, il faut donc chercher de nouvelles approches dans ces domaines 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, où elle est dans sa version standard, quand on fait la reconnaissance de mots isolés. Néanmoins d'autres méthodes existent, comme les modèles de Markov cachés (HMM) et les réseaux de neurones par exemple bien plus puissant que l'algorithme DTW mais bien plus complexe.

Comme travaux futures on propose les points suivants :

- Ajouter des algorithmes de prétraitement tels que la détection et l'activation par voix (VAD) pour réduire les calcules et l'amélioration de reconnaissance.
- Augmenter la base de données de référence (corpus) pour améliorer l'efficacité du system de reconnaisse.
- Faire une étude comparative entre cette méthode et les autres tels que les HMM et les ANN.

Bibliographie

1 https://fr.wikipedia.org/wiki/Reconnaissance_automatique_de_la_parole#cite_ref-1 30/01/2023.

2 https://www.altervoice.com/application-vocale-developpement-sur-mesure/pourquoi-la-commande-vocale 31/01/2023.

3 https://vivoka.com/fr/pourquoi-reconnaissance-automatique-parole-complexe 31/01/2023.

4 Georges Linares pour l'AFCP (L'Association Francophone de la Communication Parlée), le 17 novembre 2003.

5 Nicolas Scheffer pour l'AFCP (L'Association Francophone de la Communication Parlée), le 18 novembre 2003.

6 Pascal Nocera pour l'AFCP (L'Association Francophone de la Communication Parlée), le 17 novembre 2003.

7 DEBILOU Chaima et BOUDAOUD Samiha - Amélioration d'un synthétiseur de la parole par concaténation-MASTER ACADEMIQUE- Université Echahid Hamma Lakhdar El-Oued-Juin 2019.

8 IAN MCLOUGHLIN -Applied Speech and Audio Processing -École d'ingénierie informatique-Université technologique de Nanyang-Singapour.

9 Naima ZERARI-INTÉGRATION D'UN MODULE DE RECONNAISSANCE DE LA PAROLE AU NIVEAU D'UN SYSTÈME AUDIOVISUEL - APPLICATION TÉLÉVISEUR-Doctorat en Sciences en Génie Industriel-AVRIL 2021.

10 Zahira BENKHELLAT et Ali BELMEHDI-Utilisation des Algorithmes Génétiques pour la Reconnaissance de la Parole-Université de Bejaia Algérie-March 22-26, 2009.

11 Mlle.BELGHITRI KARIMA-Système sécurisé à base vocale- Master en Réseaux et Système distribué -Université Abou Bakr Belkaid, Tlemcen- Année universitaire 2014-2015.

12 Classification spatio-temporelle des localisations d'activité des utilisateurs de cartes à puce en transport en commun- LI HE1, MARTIN TRÉPANIER , BRUNO AGARD - Département de mathématiques et génie industriel- Canada.

13 Laurent BUNIET- Traitement automatique de la parole en milieu bruité : étude de modèles connexionnistes statiques et dynamiques- Doctorat de l'Université Henri Poincaré - Nancy 1spécialité informatique- lundi 10 février 1997.

14 HAMADOUCHE Maamar-Techniques d'analyse en vue de la reconnaissance automatique de la parole –MEMOIRE DE MAGISTER -université SAAD DAHLAB de BLIDA-Mai 2008.

15 **Titus Felix FURTUNĂ- Dynamic Programming Algorithms in Speech Recognition- Academy of Economic Studies, Bucharest- 2008.**