الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

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

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Master

Filière Électronique
Spécialité Instrumentation

présenté par

Renane Amina

&

Ketir Hamida

Extraction des caractéristiques d'un signal audio en vue d'une identification vocale d'un locuteur

Proposé par : **Ykhlef Farid**

Année Universitaire 2018-2019

Remerciements

A ce titre, nous remercions vivement notre encadreur Mr YKHLEF FARID pour ses conseils et son suivi durant la réalisation de notre projet.

Aussi, nous tenons à exprimer notre reconnaissance aux membres du jury :Mr BENSLAMA ZOUBIR et Mr GUESSOUM ABDEREZZAK.

Et enfin un remerciement à tous nos enseignants, pour leurs contributions concrètes à travers l'accès à l'information et surtout pour le savoir et les efforts qu'ils ont fourni durant notre cursus d'étude.

Résumé

ملخص:

هذا المشروع جزء من الأنظمة الذكية للتعرف على صوت الناس. بدافع من تحسين دقة الاعتراف ، تم تطوير دراسة حول اختيار المعلمات الصوتية. معاملات ceptral ومشتقاتها جنبا إلى جنب مع التردد الأساسي لإشارة الكلام قد تم تحديدها. تم إجراء تجارب على قاعدة بيانات عالمية للعثور على أفضل مجموعة من المعلمات التي تعمل على تحسين جودة التعرف.

كلمات المفاتيح : التعرف على المتحدث التلقائي; MFCC ;تردد الأساس; استخراج المعلمة; تحديد المتحدث ; KNN.

Résumé :

Ce projet rentre dans le cadre des systèmes intelligents pour la reconnaissance vocale de personnes. Motivée par l'amélioration de la précision de la reconnaissance, une étude sur la sélection des paramètres vocaux a été développée. Des coefficients ceptraux ainsi que leurs dérivés combinés à la fréquence fondamentale du signal vocal ont fait l'objet de paramétrisation. Des expériences sur une base de données universelle ont été réalisées pour retrouver la meilleure combinaison des paramètres qui améliore la précession de la reconnaissance.

Mots clés : Reconnaissance automatique de locuteur ; MFCC ; pitch ; extraction des paramètres ; KNN ; identification de locuteur.

Abstract :

This project uses the intelligent system techniques for automatic voice recognition. Motivated by the improvement of the recognition accuracy, we developed a study on the features selection of voice. Ceptral coefficients and their derivatives combined with the fundamental frequency (pitch) of the speech signal have been parameterized. Experiments on a universal database were performed to find the best combination of features that improves the recognition accuracy.

Keywords: Automatic speaker recognition; MFCC; pitch; parameter extraction; speaker identification ;KNN.

Listes des acronymes et abréviations

AMDF :Autocorrelation Modified Difference Function
CMN : cepstral mean normalization
CMS : cepstral mean substraction
CMVN : cepstral mean and variance normalization
DCT : transformée en cosinus discrète
DTW : dynamic time warping
FFT : transformation de Fourier rapide
GMM : modèles de mélange Gaussien.
HMM : modèles de Markov caché
K_NN : K_nearest neighbor
LPC : coefficient de prédiction linéaire
MFCC : Mel frequency cepstral coefficients
PLP : prediction lineaire perceptuelle
RAL : reconnaissance automatique du locuteur.
SVM : support de machine vectorielle
VAD : détection d'activité vocale
VQ : vector quantization

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

La Reconnaissance Automatique du Locuteur (RAL) est l'une des branches de l'authentification permettant de sécuriser un système, qui fait référence à la reconnaissance automatique de l'identité des personnes en utilisant certaines de leurs caractéristiques intrinsèques. Il existe de nombreux autres modèles physiques et comportementaux pour l'authentification, par exemple: l'iris, les réseaux veineux de la rétine, les réseaux veineux de la paume de la main, l'empreinte digitale, etc. Pratiquement, le choix d'un modèle adéquat devrait prendre en compte au moins les considérations suivantes : la robustesse, la précision, l'accessibilité et l'acceptabilité. Par rapport à ces critères de sélection, parmi toutes les technologies d'authentification, la reconnaissance du locuteur est probablement la plus naturelle et économique pour les systèmes de communication homme-machine parce que d'une part la collecte de données parole est beaucoup plus pratique que les autres motifs, et d'autre part, la parole est le mode dominant d'échange d'information pour les êtres humains et tend à être le mode dominant pour l'échange d'information pour les systèmes de communication homme-machine.

Ce travail de master étudie d'une manière détaillée un système complet de reconnaissance du locuteur en s'appuyant sur les nouvelles approches basées sur les techniques d'intelligences artificielles (IA). Dans cette otique, nous avons utilisé le modèle K-plus proches Voisins k-ppv (k-Nearest Neighbor ou kNN, en anglais) pour développer ce système.

En plus de l'introduction générale, ce rapport de master développera trois chapitres et une conclusion générale :

Des généralités sur le système d'identification vocale d'un locuteur sont introduites dans au premier chapitre. Le deuxième chapitre présente des notions de base sur la production de la parole et son prétraitement. En chapitre trois, les résultats de l'extraction des caractéristiques d'un signal audio ainsi que la classification par KNN sont résumés dans le cadre d'une application d'identification de locuteur. A la fin, une conclusion générale clôturera le travail de ce rapport de master.

Chapitre 1 Système d'identification vocale d'un locuteur

Introduction

La reconnaissance automatique du locuteur (RAL) est définie comme étant le processus d'identification d'une personne sur la base d'informations contenues dans le signal de parole. Dans ce monde où les atteintes à la sécurité constituent une menace

majeure, la reconnaissance du locuteur est l'une des principales techniques de reconnaissance biométrique. Un grand nombre d'organisations telles que les banques, les laboratoires de défense, les industries, la surveillance médico-légale utilisent cette technologie à des fins de sécurité.

Grâce à cette technique, il est possible d'utiliser la voix des personnes pour vérifier leur identité et contrôler l'accès à des services tels que les achats par téléphone, les messageries, le démarrage automatique d'une voiture sécurisé par l'empreinte vocale.

Les systèmes RAL sont constitués de trois principales parties:

- module de prétraitement et de caractérisation qui permet d'effectuer quelques conditionnements sur le signal traité (tels que la normalisation de l'amplitude, etc.) pour le rendre plus propice à la prochaine phase.
- extraction des paramètres acoustiques de chaque locuteur pour avoir le modèle à partir des vecteurs de caractéristiques.
- le module ainsi obtenu va permettre la reconnaissance de nouveaux sujets (pendant la phase de test du système) sur les modèles existants et stockés dans une base de données (lors de la phase d'entraînement du système), afin de vérifier où d'identifier le locuteur.

Dans ce chapitre nous allons étudier les outils nécessaires à la RAL en détaillant d'une manière explicite les trois parties.

La reconnaissance automatique du locuteur

Un système de reconnaissance se compose de deux principales tâches : vérification et identification. Dans notre mémoire nous nous intéressons à la phase d'identification.

Vérification du locuteur

Un système de vérification vocale cherche à décider si l'identité revendiquée par un locuteur est compatible avec sa voix. Dans ce type d'application, il est nécessaire de choisir entre deux hypothèses, soit le locuteur est bien le locuteur autorisé, c'est à dire celui dont l'identité est revendiquée, soit nous avons affaire à un imposteur qui cherche à prétendre être un locuteur autorisé (Figure 1.1).

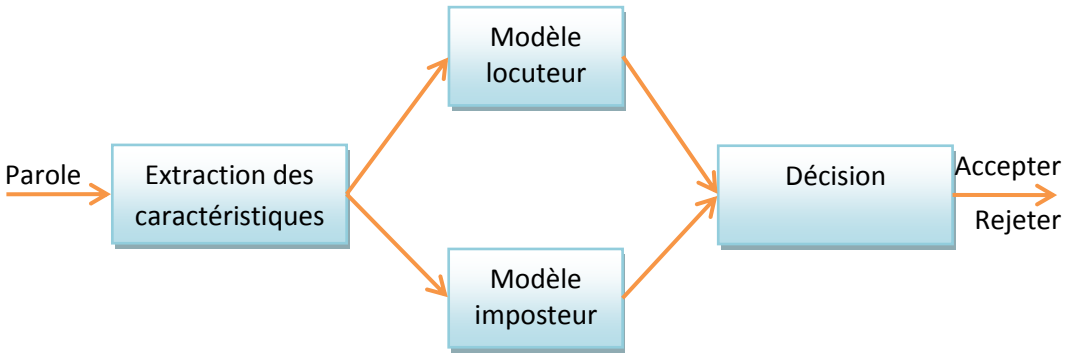

Figure 1.1 : Vérification automatique de locuteur.

Identification du locuteur

L'identification du locuteur consiste à déterminer l'identité d'un individu parmi une population de personnes connues. À partir d'un échantillon d'une voix enregistrée, on cherche à déterminer quel locuteur de la base de données a parlé. Pour ce faire, les données de la base sont comparées à une référence caractéristique de chaque utilisateur connu du système. Le résultat de chaque comparaison est un score, fonction de la similarité observée par le système entre les données du locuteur et la référence considérée. Le score le plus élevé correspond à la référence la plus proche des données de test et l'identité du locuteur correspondant à cette référence est renvoyée par le système [1].

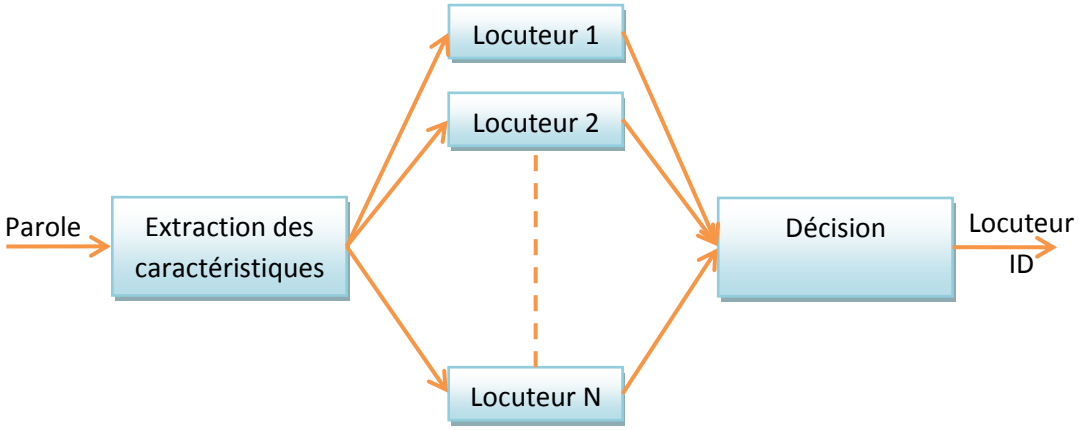

Figure 1.2 : Identification automatique de locuteur.

Phases de la reconnaissance vocale

Tous les systèmes de reconnaissance du locuteur doivent suivre les deux phases suivantes :

La phase d'entraînement : dans laquelle chaque locuteur susceptible d'être testé à l'avenir doit fournir des échantillons de parole afin que le système puisse former un modèle de référence pour chacun d'eux.

La phase de test : où le signal de parole à identifier est associé aux modèles de référence collectés après l'apprentissage. Ainsi, la décision finale est prise [1] [2].

Problèmes rencontrées dans la reconnaissance du locuteur

Les systèmes RAL soufferts des difficultés liées au domaine applicatif, comme l'utilisation des systèmes dans des conditions difficiles, les tentatives d'imposture, etc.

Afin de mieux mettre en œuvre les techniques qui améliorent les performances des systèmes RAL, dans ce qui suit, on introduira les variabilités nuisibles pouvant influer négativement sur les performances d'un système RAL.

Variabilité intra-locuteur

Le signal de parole varie pour le même individu car la voix d'une personne peut changer entre le début et la fin de la journée. Cette variabilité intra-locuteur est induite par l'évolution naturelle ou volontaire de la voix d'une personne (fatigue, état émotionnel, maladies des organes de la voie).

Variabilité interlocuteur

Ce type de variabilités est dit "de haut-niveau" et reflète différents scénarios concernant l'interaction vocale avec une autre personne ou un système technologique. Cette variabilité peut aussi être issue de différences par rapport à la langue ou au dialecte parlés.

Variabilité due au matériel

Cette variabilité est due aux équipements d'acquisition du signal vocal : microphone, combiné téléphonique, ligne de transmission (ex : lignes téléphoniques), convertisseurs. Ces informations apparaissent le plus souvent sous la forme de déformations/dégradations du signal de parole.

Robustesse en environnements et tentatives d'imposture

Les systèmes RAL doivent être robuste face au bruit ambiant et les environnements des canaux digitaux (téléphone, réseaux mobile, internet...) [3].

Outils de l’identification du locuteur

Un système RAL est basé sur la connaissance de "N" clients d'un système, où chacun étant représenté par un modèle. Lors de l'arrivée d'un signal de parole, le système doit déterminer l'identité de la personne qui parle, parmi les N connus. Un système de vérification est utilisé pour connaître le type d’identité revendiquée. Le système apporte si le locuteur "I" parle ou non dans l'enregistrement en cours. La majorité des systèmes de reconnaissance du locuteur, qu’il s’agisse de tâches d’identification ou de vérification, utilisent des modèles de mélange gaussien (GMM) dans la modélisation du locuteur, soit exclusivement, soit en combinaison avec d’autres techniques telles que HMM (modèle de Markov caché) ou SVM (support de machine vectorielle). Un système de reconnaissance (identification ou vérification) comprend plusieurs composants: un module d’extraction de paramètres, un bloc d’appariement (matching), un module de normalisation de score d’appariement et un module de décision. La figure 1.1 illustre l'architecture d'un système RAL, y compris son identification et sa vérification [1].

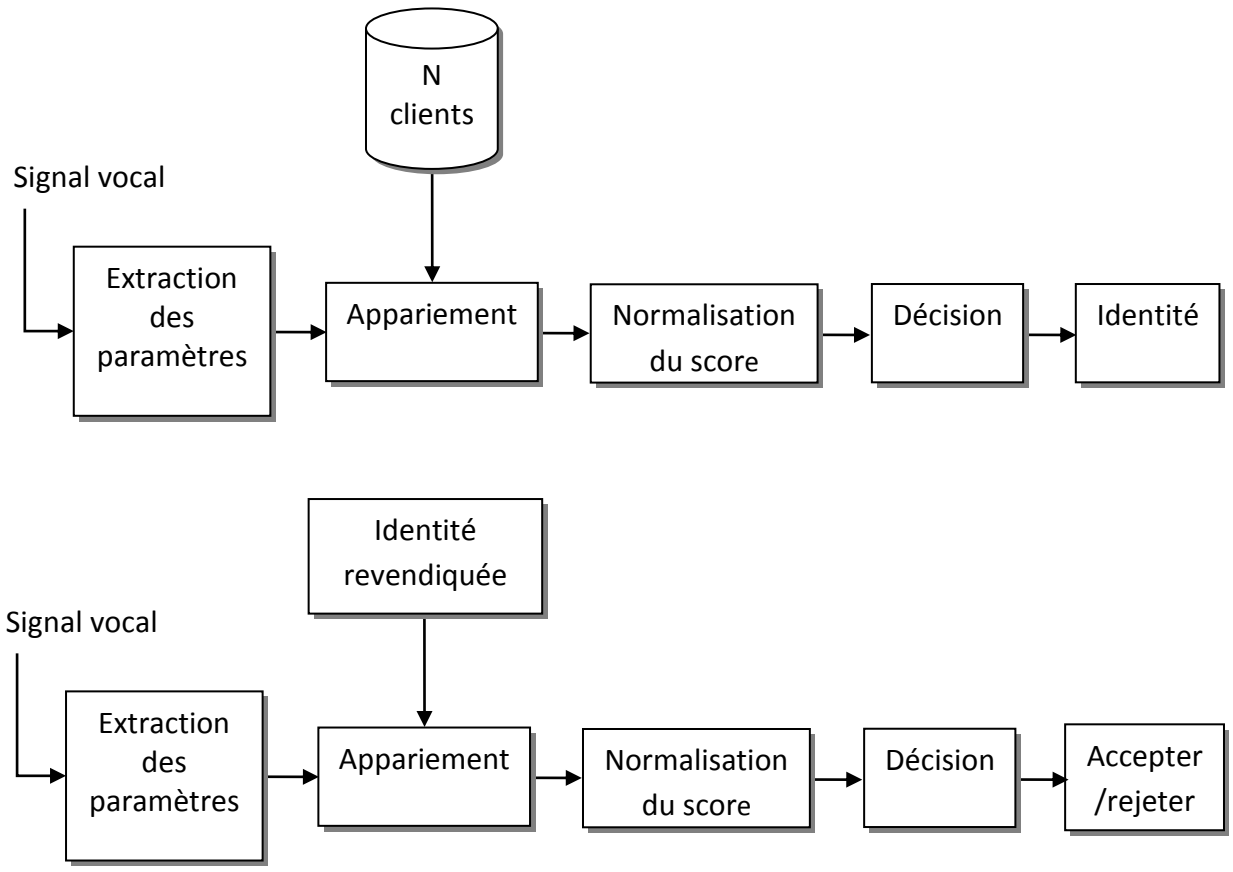

Figure 1.3 : Architecture d'un système RAL (Identification en Haut, Vérification en Bas).

Nous présentons dans les sous-sections suivantes les différentes approches et techniques utilisées dans l'extraction de paramètres, la modélisation et la normalisation des scores.

Extraction des paramètres

La phase d'extraction des paramètres (caractéristiques) a pour but de générer des éléments vectoriels riches en information [2].

Paramétrisation

Le signal de parole est intrinsèquement complexe et redondant et présente une grande variabilité, ce qui rend difficile son utilisation directe par les systèmes RAL. Cette complexité provient du fait de la combinaison de plusieurs facteurs ; la grande variabilité inter et intra-locuteurs, les effets de la coarticulation en parole continue, les conditions d'enregistrement, etc. Par conséquent, il devient nécessaire d'effectuer une étape de paramétrisation qui vise à extraire une représentation plus compacte de cette information acoustique qui réduit la redondance et améliore les propriétés spécifiques du locuteur.

De plus, ces paramètres doivent être robustes aux différents interférences et variations inter-sessions, et difficiles à reproduire par un imposteur. Les paramètres spectraux à court terme, les paramètres liés à la source de voix, les paramètres prosodiques et les paramètres de haut niveau peuvent être aussi utilisés [4].

Paramètres spectraux à court terme

Le signal de la parole varie de manière continue dans le temps, en fonction des mouvements articulatoires. Par conséquent, sa paramétrisation doit être effectuée sur de courtes fenêtres d'analyse (généralement entre 10 à 30 ms) où le signal est considéré comme quasi-stationnaire. L'analyse utilise des fenêtres glissantes qui se chevauchent, avec un décalage régulier de 5 à 10 ms. Le signal de parole est d'abord filtré par un filtre de préaccentuation numérique (un filtre passe-haut) afin d'intensifier les hautes fréquences, dont l'énergie est toujours inférieure à celle des basses fréquences. Pour améliorer l'analyse et limiter les effets de bord, les trames du signal sont pondérées par une fenêtre temporelle aplatie aux extrémités, ce qui réduit les discontinuités dans le signal.

Dans le domaine du traitement du signal, de nombreuses fenêtres ont été proposées et étudiées (par exemple: Hanin, Blakmann, Kaiser et Hamming). Dans la représentation source / filtre du signal de parole, ce signal résulte d'une convolution (dans le domaine temporel) de la source et du conduit vocal (filtre): $s(n) = e(n) * h(n)$.

Le passage dans le domaine log-spectral permet de remplacer cette convolution par une somme: $\log|S(f)| = \log|E(f)| + \log|H(f)|$. Le cepstre réel d'un signal numérique est obtenu en appliquant une transformée de Fourier inverse au logarithme de son spectre. Dans ce nouveau domaine, la séparation source-conduit peut être effectuée facilement via une simple fenêtre temporelle (appelée liftrage) [4].

Au cours des années, de nombreux paramètres à court terme ont été développés pour les applications de reconnaissance vocale, puis utilisés pour la reconnaissance des locuteurs. Par la suite, nous en citons quelques-uns. Les paramètres du vecteur acoustique peuvent être:

Descripteurs de spectre

- Le spectre à court terme (calcul de la FFT)est utilisé pour caractériser les signaux de parole.
- Les coefficients LPC (Linear Prediction Coefficients : coefficients de prédiction linéaire) utilisés pour l'extraction de formants.
- L'énergie par bande spectrale (Mel, Bark, Harmonique).
- Les coefficients PLP (Perceptual Linear Prediction : prédiction linéaire perceptuelle) sont les coefficients LPC obtenus sur une échelle de Bark perceptive.

Descripteurs du Cepstre

Le Cepstre est le résultat de la transformée de Fourier appliquée à spectre en décibel d'un signal. Son nom dérive du renversement des quatre premières lettres du mot « spectre ».Il est parfois appelé spectre de fréquences.

Les descripteurs dérivé du Cepstre sont :

- Les coefficients cepstraux (MFCC) généralement sur 13 points
- Les coefficients LPCC (Linear Prediction Cepstral Coefficients) sont des LPC obtenus sur le Cepstre.
- L'analyse en banc de filtres (MFCC : Mel Frequency Cepstral Coefficients) : est une technique d'analyse Cepstrale non-paramétrique qui utilise l'échelle de Mel.

Les MFCCs seront traités plus en détails dans la prochaine sous section.

MFCC

Les MFCC (Mel Frequency Cepstral Coefficients) sont utilisés pour les tâches de reconnaissance de la parole, du langage ou du locuteur. Malgré que le grand nombre d'ouvrages proposent d'autres méthodes, les MFCC restent les plus utilisées dans le domaine du traitement de la parole. Le processus d'extraction des MFCC est décrit par figure 1.4.

Le signal de parole est analysé localement en utilisant le fenêtrage temporel (souvent Hanning ou Hamming) afin de réduire les effets de bord causés par la troncature (fenêtre rectangulaire). La longueur de la fenêtre glissante utilisée (entre 20 et 30 millisecondes) est choisie pour respecter la stationnarité. Le décalage des fenêtres temporelles utilisées pour extraire deux segments consécutifs du signal est choisi de sorte que ces fenêtres chevauchent en partie le segment du signal tramé et ensuite appliqué à une transformation de Fourier rapide (Fast Fourier Transform - FFT).

Le module du spectre obtenu est filtré par un banc de filtres qui permet de réduire la taille du vecteur spectral en calculant la moyenne du spectre sur la bande de fréquence correspondant à chacun des filtres.

Les fréquences centrales de chaque filtre sont fixées par l'échelle MEL. Le logarithme de ces valeurs est calculé et multiplié par 20 pour obtenir l'enveloppe spectrale en décibels.

Les coefficients acoustiques ainsi obtenus à ce stade peuvent être directement utilisés pour les prochaines étapes de traitement. Dans ce cas, on parlera dans ce cas de banc de filtres logarithmique ou log filter-bank (FB) en anglais.

La dernière étape de la paramétrisation consiste à appliquer une transformée en cosinus discrète (DCT) à partir de laquelle résultent les coefficients Cepstraux (MFCC). La transformée en cosinus discrète est utilisée ici pour sa capacité à décorréler les données.

Des informations dynamiques sont ajoutées à ces coefficients en les concaténant avec leurs dérivées première et seconde temps inférieures à celles des basses fréquences.

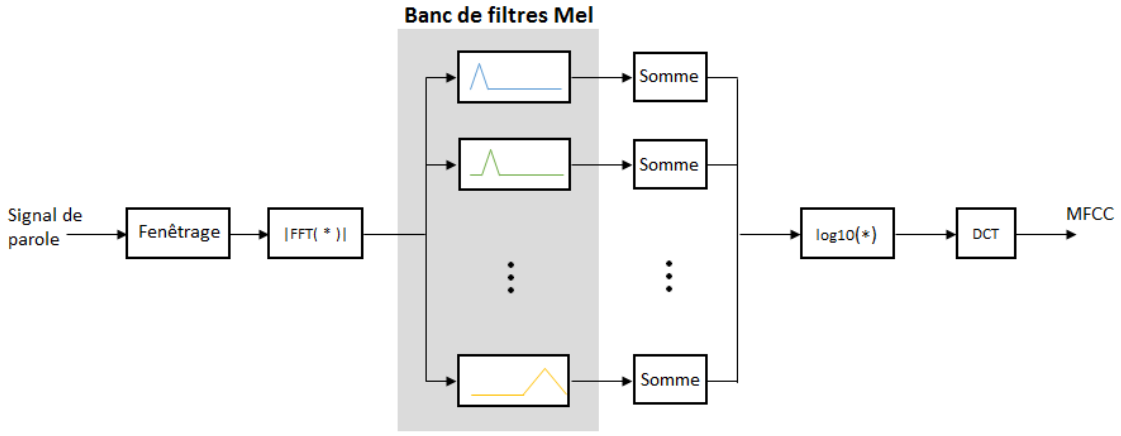

Figure 1.4 : Calcul des coefficients MFCC.

Le principe de calcul des coefficients MFCC repose sur des recherches psychoacoustiques sur la tonie et la perception de différentes bandes de fréquences par l'oreille humaine. La FFT passe dans un banc de filtres à l'échelle de Mel. Cette échelle non linéaire tient principalement compte du fait que la perception des intervalles change en fonction de la zone du spectre à laquelle appartiennent les hauteurs qui les composent. L'intérêt principal de ces coefficients est d'extraire des informations pertinentes en nombre limité, à la fois sur la base de la production (théorie de Cepstale) et sur la perception de la parole (échelle de Mels).

Le calcul se déroule comme suit:

- La FFT est calculée sur un fragment (trame).
- Cette dernière est filtrée par un banc de filtres triangulaires répartis le long de l'échelle de Mel.
- Le logarithme du module de l'énergie de sortie du banc de filtres est calculé.
- Une transformation en cosinus discrète inverse (équivalente à la transformée FFT inverse pour un signal réel) est appliquée. Seuls les premiers coefficients sont retenus.

Coefficients différentiels (dynamiques)

Les changements temporels dans le cepstre jouent un rôle important dans la perception humaine et c'est à travers les dérivées des coefficients des MFFC statiques que ces changements peuvent être mesurés.

Généralement les coefficients MFCC sont désignés comme des paramètres statiques, puisqu'ils contiennent seulement l'information sur une trame donnée. Afin d'améliorer la représentation de la trame, il est souvent proposé d'introduire de nouveaux paramètres dans le vecteur des paramètres. L'auteur dans [5] a proposé l'utilisation des paramètres dynamiques qui présentent l'information de transition spectrale dans

le signal vocal. En particulier, il a proposé des coefficients différentiels du premier ordre appelés aussi coefficients delta, issus des coefficients cepstraux ou de l'énergie. Soit $C_k(t)$ le coefficient cepstral d'indice k de la trame t, alors le coefficient différentiel $\Delta C_k(t)$ correspondant est calculé sur $2n\Delta$ trames d'analyse par l'estimation de la pente de la régression linéaire du coefficient C_k à l'instant t [5]:

$$\Delta C_k(t) = \frac{\sum_{i=-n_\Delta}^{n_\Delta} i C_k(t+i)}{2\sum_{i=-n_\Delta}^{n_\Delta} i}$$

Les coefficients différentiels du second ordre ΔΔ (delta-delta ou d'accélération) sont calculés de la même manière à partir des coefficients du premier ordre.

Remarque sur l'échelle de MEL

C'est une échelle de fréquences basée sur la vision humaine. Elle se mesure en Mels. On considère que l'oreille humaine perçoit linéairement le son jusqu'à 1000 Hz, mais après, elle perçoit moins d'une octave par doublement de fréquence. L'échelle de Mels modélise assez fidèlement la perception de l'oreille : linéairement jusqu'à 1000 Hz, puis logarithmiquement au-dessus. La conversion du Hertz en Mels se fait avec une des formules suivantes :

$$M = \frac{1000 \ln\left(1+\frac{f}{700}\right)}{\ln\left(1+\frac{1000}{700}\right)} = 1127 \ln\left(1+\frac{f}{700}\right) = 2595 \log_{10}\left(1+\frac{f}{700}\right)$$

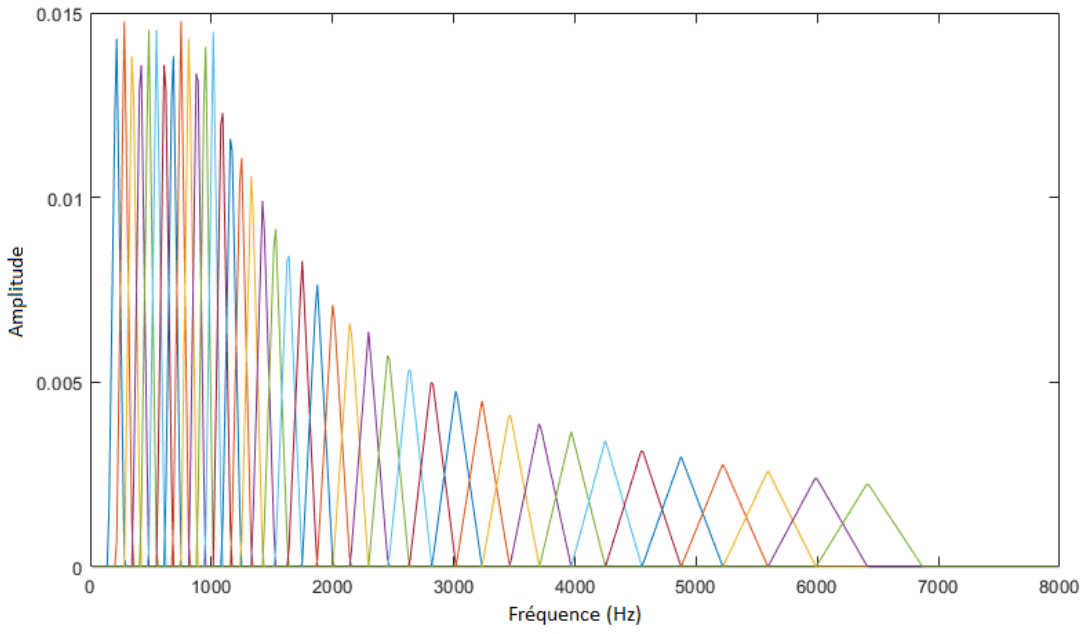

Figure 1.5 : Bancs de filtres MEL.

Paramètres prosodiques

Contrairement aux paramètres spectraux à court terme traditionnels, les paramètres prosodiques peuvent impliquer des segments de parole plus longs (syllabe, mot, expression).

Des exemples de paramètres prosodiques sont: fréquence fondamentale, intonation, accentuation, énergie, rythme et débit de la parole. Ces derniers caractérisent le style d'élocution du locuteur. La combinaison de paramètres prosodiques et de paramètres spectraux à court terme permet d'améliorer les performances d'identification et de vérification des systèmes de reconnaissance du locuteur [manuscrit]. Cette combinaison est l'objectif de notre travail de master.

Le terme "paramètres prosodiques" combine énergie, durée et fréquence fondamentale (ou pitch). Ces paramètres sont toutefois fragiles dans la pratique et ne permettent pas, à eux seuls, de discriminer les locuteurs. En conséquence, ils sont souvent associés aux paramètres de l'analyse spectrale [1].

Énergie

L'énergie d'un signal est un indice qui peut contribuer à la détection de zones de voisement d'un segment de parole. L'énergie totale est calculée dans le domaine temporel sur une trame du signal $s(n)$:

$$E_0 = \frac{1}{N}\sum_{n=0}^{N-1} |s(n)|^2$$

L'énergie ainsi obtenue est sensible au niveau d'enregistrement; on choisit généralement de le normaliser et d'exprimer sa valeur en décibels par rapport à un niveau de référence [1].

Fréquence fondamentale (ou pitch)

Le pitch est un paramètre très important pour l'étude acoustique et la synthèse de la parole. L'oreille humaine est très sensible à ses variations du pitch qui constituent la prosodie. L'évolution de la fréquence en fonction du temps au niveau du phonème constitue la micro mélodie, par contre son évolution au niveau des groupes syntaxiques de la phrase est la macro mélodie, l'intonation du message est directement liée au pitch. Plusieurs techniques pour l'extraction de la fréquence fondamentale peuvent être utilisées parmi celles-ci, nous pouvons citer la méthode de AMDF (en anglais "Autocorrelation Modified Difference Function") qui est donnée par :

$$AMDF(k) = \frac{1}{N}\sum_{n=0}^{N-1}|a(n) - a(n-k)|$$

Cette fonction a un minimum de multiples de la période fondamentale.

En général, les méthodes d'estimation de la fréquence fondamentale comprennent trois étapes:

- le prétraitement du signal de parole pour l'adaptation du signal (filtrage, fenêtrage, préaccentuation, etc.),
- le traitement pour l'extraction de la fréquence fondamentale,
- Post-traitement pour corriger les erreurs de calcul, notamment pour les transitions voisées/non-voisées [1].

Segmentation parole/silence

Les trames résultantes de la phase de paramétrisation ne sont pas toutes utiles pour le processus de reconnaissance du locuteur. La paramétrisation est suivie par une détection d'activité vocale (VAD : Voice Activity Detection), où l'étiquette non activité vocale désigne en réalité toute trame jugée non utile au sens de la reconnaissance. La VAD peut également intervenir avant la paramétrisation. Elle a tendance à éliminer les trames à faible et moyenne énergie, qui représentent le silence, le bruit ou l'écho. Ces trames rendent la reconnaissance des locuteurs plus difficile. On ne conserve donc que les trames de haute énergie qui correspondent principalement aux zones stables des voyelles. En pratique, il est difficile d'avoir une VAD indépendante des différentes conditions d'enregistrement. Par conséquent, on réalise une VAD par condition. La segmentation la plus simple et la plus utilisée reste celle basée sur l'énergie du signal. La répartition de l'énergie des trames est généralement modélisée par un modèle GMM à 2 (ou 3) composantes dont la loi Gaussienne à faible énergie (respectivement élevée) modélise les trames à faible énergie (respectivement élevée). L'apprentissage des lois gaussiennes permet de calculer des seuils de décision (deux seuils de décision pour un modèle à 3 composantes) en fonction des paramètres du modèle, ce qui permet d'affecter les trames à l'une des classes (parole/silence) [4].

Les approches de détection d'activité vocale les plus populaires en RAL se basent sur la distribution d'énergie des trames et utilisent un seuil de décision pour détecter les zones de parole.

Normalisation des paramètres

En réalité, il n'est pas possible de concevoir des paramètres acoustiques qui restent inchangés quelles que soient les conditions d'enregistrement. Cependant, ces

changements peuvent être atténués de diverses manières en utilisant des techniques de normalisation.

Soustraction du cepstre moyen : C'est une méthode efficace pour traiter la variabilité du canal est la soustraction du cepstre moyen (CMN, Cepstral Mean Normalization, également appelé CMS, Cepstral Mean Substraction). Cette technique permet de compenser les bruits non intentionnels en retardant autant que possible les composants continus ou lents. La transformation appliquée peut être écrite sous la forme :

$$c = c_t - \mu_{cmn} \text{ avec } \mu_{cmn} = \frac{1}{T}\sum_{t=1}^{T} c_t$$

Où c_t représente un vecteur de paramètres cepstraux à normaliser, μ_{cmn} est la moyenne à calculée en utilisant les trames de parole correspondant à un locuteur donné.

Normalisation CMVN: C'est une extension de la CMN qui consiste à la normalisation de la moyenne et de la variance cepstrale (CMVN : Cepstral Mean and Variance Normalization). Elle est très simple et très utilisée, et elle consiste à retirer la moyenne de la distribution de chacun des paramètres cepstraux (la composante continue), et à ramener la variance à une variance unitaire en les divisant par l'écart type global des paramètres acoustique. Cette technique permet d'avoir le même ordre de grandeur sur les trames provenant de différents segments de parole [4].

Normalisation par feature warping : La méthode de normalisation feature warping est aussi une approche populaire dans le domaine de la RAL. Le principe consiste à une transformation non-linéaire de la distribution des coefficients cepstraux pour la correspondre à une distribution Gaussienne (comme illustré par la figure1.6). Cette transformation est appliquée sur une fenêtre glissante à durée courte (3 secondes).

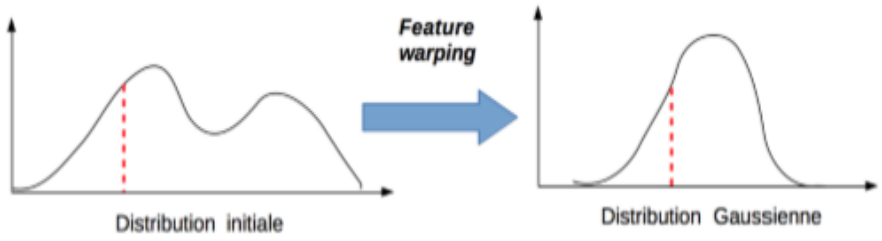

Figure 1.6 : Principe de la normalisation feature warping.

Modélisation

Le module de modélisation et d'apprentissage a pour but de classer les objets d'intérêt, c'est-à-dire les séquences de vecteurs acoustiques extraits du signal de parole, en une ou plusieurs catégories ou classes. Chaque classe est ici attachée à un locuteur [2]. La modélisation des paramètres et les techniques d'extraction sont les composants principaux d'un système de reconnaissance de locuteur. Dans cette partie, nous explorons quelques techniques de modélisation utilisées dans la RAL. Nous distinguons l'approche vectorielle, connexionniste, prédictive et statistique [3].

Approche vectorielle

Dans l'approche vectorielle, un modèle de locuteur est un ensemble de vecteurs de paramètres représentatifs de l'espace acoustique résultant de la phase de paramétrisation des signaux d'apprentissage. Dans cette approche, nous trouvons deux techniques principales: la programmation dynamique et la quantification vectorielle.

Programmation dynamique

La programmation dynamique (Dynamic Time Warping : DTW) consiste à aligner temporellement une séquence de vecteurs de paramètres de test sur une séquence de vecteurs d'apprentissage. Bien que facile à mettre en œuvre, très rapide et offrant des performances relativement bonnes, la programmation dynamique est cependant très sensible à la qualité de l'alignement et en particulier au choix du point de départ des deux formes à comparer.

Quantification vectorielle

La quantification vectorielle (Vector Quantization : VQ) repose sur un partitionnement de l'espace acoustique en sous-espaces. Chaque sous-espace est associé à leur vecteur centroïde, i. e., à un vecteur de paramètres représentant l'ensemble des vecteurs composant le sous-espace. Dans ces conditions, un modèle de locuteur est composé d'un ensemble de vecteurs centroïdes, appelé dictionnaire de quantification (codebook). Au cours de la phase de reconnaissance, une distance est calculée entre un vecteur test et chaque vecteur centroïde du dictionnaire. La distance minimale est conservée. La quantification vectorielle s'applique en mode dépendant du texte ou indépendant du texte. La vitesse et les performances de cette technique dépendent fortement de la taille du dictionnaire : plus la taille du dictionnaire est grande, meilleures sont les performances. Néanmoins, le processus devient plus lent.

Approche prédictive

Les modèles prédictifs sont basés sur le principe selon lequel une trame d'un signal de parole peut être générée à partir des trames précédentes du signal. Un locuteur donné est représenté par une fonction de prédiction estimée sur ses données d'entraînement (apprentissage). Deux stratégies peuvent alors être adoptées pour la reconnaissance: soit calculer une erreur de prédiction en tant que mesure de similarité, entre les trames et les trames réellement observés ; ou comparer la fonction de prédiction du locuteur concerné avec une nouvelle fonction de prédiction estimée cette fois sur les nouvelles données, en fonction d'une mesure de distance donnée. Dans la littérature, deux grandes familles de fonctions prédictives existent : les modèles vectoriels autorégressifs (AR-Vector Models) et les réseaux neuronaux prédictifs [3].

Approche statistique

Les techniques statistiques considèrent le locuteur comme une source probabiliste et le modélisent selon une densité de probabilité connue. La phase d'apprentissage consiste à estimer les paramètres de la fonction de la densité de probabilité. La décision est prise en calculant la vraisemblance des données par rapport au modèle appris préalablement. Les modèles de Markov cachés HMM ont été utilisés dans des applications dépendantes du texte dépendant de reconnaissance automatique du locuteur, tandis que les modèles de mélange de lois Gaussiennes GMM et les machines à vecteurs de support SVM sont largement utilisées comme applications indépendantes du texte ainsi dans des applications de vérification du locuteur [3].

L'inconvénient commun aux méthodes présentées ci-dessus est qu'elles ne tiennent pas compte de l'ordre dans lequel les vecteurs de paramètres sont présentés. L'approche statistique résout ce problème en utilisant des techniques permettant de construire des modèles prenant en compte l'aspect temporel du signal de parole. Les vecteurs acoustiques résultant de la paramétalisation sont donc représentés par des statistiques à long terme.

Classification

C'est le processus de reconnaissance en intention (par leurs propriétés) des classes décrites en extension (par les valeurs de leurs descripteurs). Si les valeurs à prédire sont des classes en petit nombre, on parle alors de classification.

Modèles de Markov cachés (HMM)

Les modèles de Markov (HMM : Hidden Markov Models) ont été largement utilisés dans le domaine de la reconnaissance automatique de la parole. Plus récemment, leur utilisation s'est étendue à la reconnaissance automatique du locuteur. La modélisation dans ce cas se fait par une succession d'états avec des probabilités de passage d'un état à un autre. La reconnaissance est effectuée en calculant la probabilité qu'une séquence de vecteurs de test soit dérivée de la chaîne de Markov. L'utilisation de HMMs dépendants du texte fournit d'excellents résultats [4].

Modèle de mélange de Gaussiennes (GMM)

Ces modèles permettent d'estimer les paramètres d'une distribution de variables aléatoires en les modélisant par une densité mélange. Habituellement, ils sont utilisés en classification (supervisée ou non) et on considère que chaque composant du mélange caractérise une classe. Ces modèles présentent deux avantages principaux:

C'est une méthode probabiliste pour obtenir une classification des observations. Une probabilité d'appartenance à chaque classe est calculée et une classification est généralement obtenue en affectant chacune des observations à la classe la plus probable. Ces probabilités permettent également d'interpréter certaines classifications suspectes. Ils offrent une grande souplesse de modélisation et permettent ainsi de modéliser un grand nombre de phénomènes. Cette nouvelle modélisation présente l'avantage d'utiliser la densité des paramètres acoustiques pour représenter les locuteurs (au lieu d'utiliser des vecteurs prototypes dans le cas de la VQ) et permet de prendre en compte la variance des données ainsi que la contribution des paramètres acoustiques à chaque zone de l'espace acoustique sous forme de probabilité. Cette modélisation a obtenu de bonnes performances et est encore utilisée par la plupart des systèmes RAL à ce jour. La figure 1.7 montre la structure de l'espace acoustique utilisant cette modélisation.

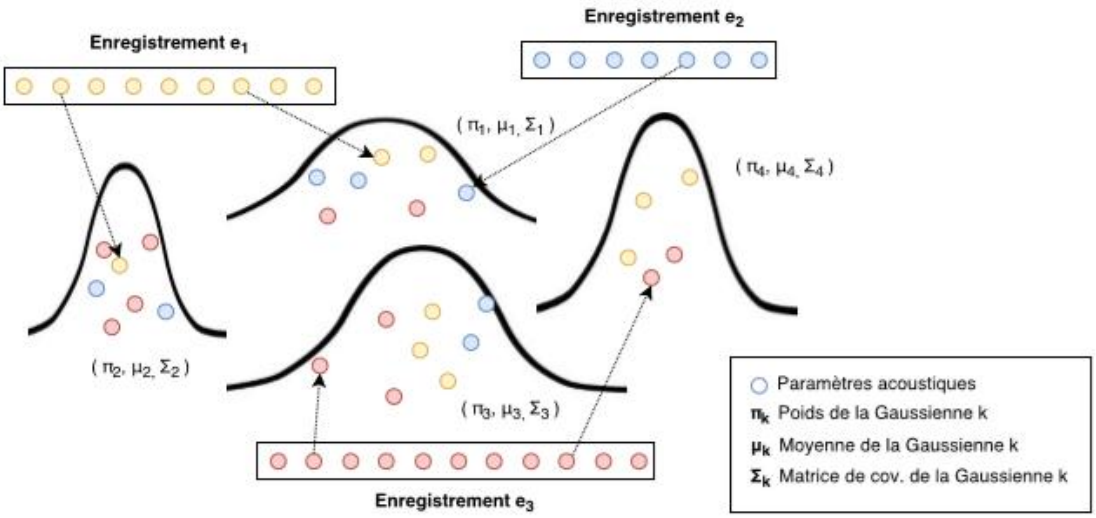

Figure 1.7: Un mélange de Gaussiennes (GMM) construit en utilisant des paramètres acoustiques issus de plusieurs enregistrements.

Machine à vecteurs de support

Les machines à vecteurs de support sont basées sur le concept des plans de décision qui définissent les limites de décision. Un plan de décision est un plan qui sépare un ensemble d'objets ayant des appartenances de classe différentes. Un exemple schématique est présenté dans l'illustration ci-dessous. Dans cet exemple, les objets appartiennent à la classe VERT ou ROUGE. La ligne de séparation définit une limite sur le côté droit de laquelle tous les objets sont VERTS et sur le côté gauche de laquelle tous les objets sont ROUGES. Tout nouvel objet (cercle blanc) qui tombe vers la droite est étiqueté, c'est-à-dire classé VERT (ou classé ROUGE s'il tombe à gauche de la ligne de séparation).

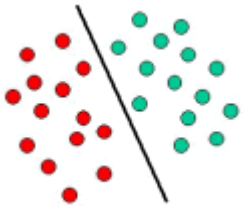

Figure 1.8 : Classificateur SVM linéaire.

C'est un exemple classique de classificateur linéaire, c'est-à-dire un classificateur qui sépare un ensemble d'objets en leurs groupes respectifs (VERT et ROUGE dans ce cas) avec une ligne. Cependant, la plupart des tâches de classification ne sont pas aussi simples et des structures plus complexes sont souvent nécessaires pour effectuer une séparation optimale, c'est-à-dire pour classer correctement de nouveaux objets (cas de test) sur la base des exemples disponibles (cas de d'entrainement). Cette situation est illustrée ci-dessous. Par rapport au schéma

précédent, il est clair qu'une séparation complète des objets VERT et ROUGE nécessiterait une courbe (qui est plus complexe qu'une ligne). Les tâches de classification basées sur le tracé de lignes de séparation pour distinguer les objets de différentes appartenances de classe sont connues sous le nom de classificateurs hyperplan. Les SVMs sont particulièrement adaptées à ce type de tâches.

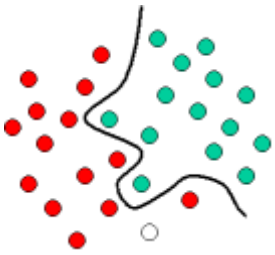

Figure 1.9 : Classificateur SVM hyperplan.

L'illustration ci-dessous montre l'idée de base derrière les SVMs. Ici, nous observons les objets originaux (à gauche du schéma) mappés, c'est-à-dire réarrangés, à l'aide d'un ensemble de fonctions mathématiques, connues sous le nom de noyaux. Le processus de réarrangement des objets est connu sous le nom de cartographie (transformation). Il faut noter que dans ce nouveau paramètre, les objets mappés (côté droit du schéma) sont séparables linéairement et, par conséquent, au lieu de construire la courbe complexe (schéma de gauche), tout ce que nous avons à faire est de trouver une ligne optimale qui peut séparer les objets VERT et ROUGE.

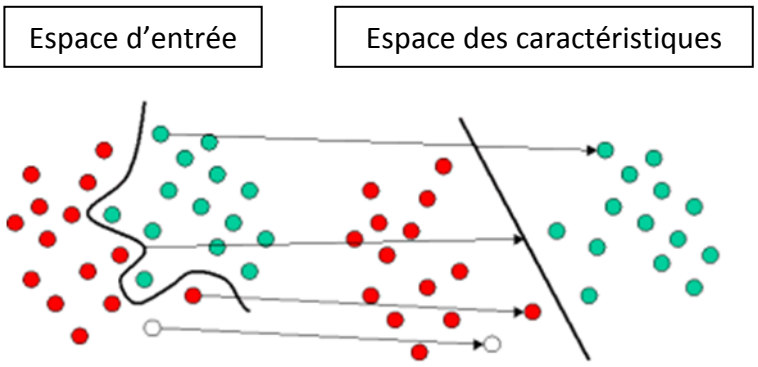

Figure 1.10 : Processus de réarrangement des objets.

K-plus proches Voisins

Le classifieur des k plus proches voisins ou k-ppv (k-Nearest Neighbor ou k-NN, en anglais) est l'un des algorithmes de classification les plus simples. Son principe est de classifier par vote majoritaire de ses k "voisins" (par mesure de distance), c'est-à-dire qu'il est prédit de classe C si la classe la plus représentée parmi ses k voisins est la classe C.

Un cas particulier est le cas où $k = 1$, l'exemple est alors affecté à la classe de son plus proche voisin.

L'opérateur de distance le plus souvent utilisé est la distance Euclidienne, cependant, en fonction du problème, on peut encore utiliser les distances de Hamming, de Mahalanobis, etc.

Remarques

- Le choix du k est très important pour la classification.
- On s'abstient de choisir des valeurs paires de k, pour éviter les cas d'égalité.

Exemple : sur la figure 1.11, on peut voir l'effet du choix de k sur le résultat de la classification. En effet, si $k = 1, 2, 3$ l'exemple à prédire (noté "?") serait classifié comme étant de la classe "X", mais si $k = 5$, il serait classifié comme étant de la classe "O".

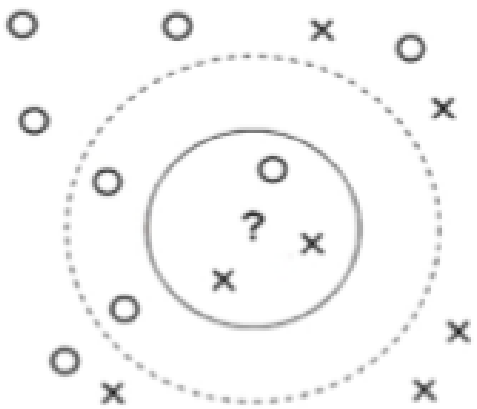

Figure 1.11 : Schéma de classification par la méthode de k-NN [mokhtarjefal].

Notion de distance

Distance d'un objet à une classe

La distance d'un point $x \in E$ (donnée ou objet) à une partie (partition ou classe) $A \subset E$, partie de l'espace métrique E, est donnée par :

$$d\,(x, y) = \; inf\,\{d\,(x, y) | y \in A\}$$

Distance entre deux classes

La distance entre deux ensembles (classes) d'objets E_1 et E_2, avec E_1 et E_2 deux parties non vides d'un espace métrique E muni d'une distance d, est donnée par :

$$(E1, E2) = inf \{d(x, y) | (x, y) \in E_1 \times E_2\}$$

Distances Usuelles

- **Distance de Manhattan**

$$1 - d(A, B) = d_1(A, B) = \sum_{i=1}^{n} |b_i - a_i|$$

- **Distance Euclidienne**

Dans le plan 2D :

$$d(A, B) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

De manière générale, pour des descripteurs de dimension n :

$$2 - distance(A, B) = d_2(A, B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

- **Distance de Mahalanobis**

Elle se base sur la corrélation entre les variables (pour mesurer la ressemblance entre les descripteurs des échantillons) par lesquelles différents modèles peuvent être identifiés et analysés. C'est, aussi, une manière utile pour déterminer la similarité entre une série de données (où chaque donnée est un vecteur de valeurs) inconnue et d'autres connues.

La distance de Mahalanobis tient compte de la variation (variance) et de la covariation entre les données ; son calcul de ce fait à travers la matrice de covariance qui permet de quantifier les écarts conjoints des donnés ou variables par rapport à leurs espérances respectives

Décision

Après avoir comparé le signal de test à tous les modèles de Locuteurs connus du système, on obtient un ensemble de mesures de similarité qui va servir d'entrée au module de décision. Ce dernier a pour tâche de rechercher la mesure de similarité maximale ou bien minimale en termes de distance et d'indiquer l'identité du Locuteur.

Pour mesurer les performances d'un système d'IAL, on utilise généralement le taux d'identification correcte I_c ou incorrecte I_i qu'on obtient par les formulations suivantes :

$$I_c = \frac{nombre\ de\ teste\ correctements\ identifiés}{nombre\ total\ de\ tentatives}$$

et

$$I_i = \frac{nombre\ de\ test\ mal\ identifiés}{nombre\ total\ de\ tentatives}$$

avec

$$I_c + I_i = 100\%$$

Conclusion

Un système de reconnaissance automatique du locuteur, quelle que soit la tâche considérée, se résume en trois étapes principales : l'analyse acoustique du signal parole, la modélisation du locuteur et la décision soit une décision et vérification. Également, tout système de RAL dépend de la technique d'extraction de paramètres utilisé, modélisation, décision et ainsi la phase de prétraitement.

Chapitre 2 Production et prétraitement de la parole

Introduction

La parole est l'une des formes les plus importantes de la communication humaine. Elle joue un rôle important dans le domaine du traitement du signal. Il existe deux caractéristiques fondamentales de la parole : les caractéristiques prosodiques et les caractéristiques articulatoires :

– Une caractéristique prosodique est construite à partir de trois paramètres acoustiques appropriés au signal numérique de la parole produite : sa fréquence fondamentale F0, son énergie et son spectre. Chaque trait acoustique est lui-même intimement lié à sa grandeur perceptuelle (respectivement) : pitch, intensité et timbre.

– Une caractéristique articulatoire se traduit physiquement, dans la parole, par une variation de la pression de l'air causée et émise par le système articulatoire.

Production de la parole

En traitement de la parole, et plus particulièrement en reconnaissance vocale, il est essentiel d'avoir aussi une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille si on désire extraire l'information de façon pertinente. Tout ce qui peut être mesuré acoustiquement ou observé par la phonétique articulatoire n'est pas nécessairement perçu et par conséquent il y a une partie du signal qu'il est inutile d'analyser. Ceci nous permet alors une réduction des données, et un gain en vitesse de traitement.

La parole naît de l'excitation de la cavité résonante. L'appareil respiratoire fournit l'énergie nécessaire à la production de sons en poussant l'air à travers l'appareil phonatoire vers la source du résonateur. Selon Joseph Campbell, la source du résonateur est en réalité décomposable en deux émissions distinctes d'origines différentes:

- Les cordes vocales, qui possèdent la particularité de produire, en plus de leur fréquence fondamentale, un spectre riche en harmoniques ; elles produisent les sons voisés.
- Le bruit d'écoulement de l'air en provenance des poumons, dont le spectre est similaire à un bruit blanc ; il crée les sons non-voisés.

L'évolution temporelle de la fréquence fondamentale (F0) ou pitch est une information spécifique à chaque locuteur, qui varie en fonction des phonèmes qu'il prononce au cours d'une phrase.

Description de l'appareil phonatoire

L'appareil phonatoire est composé principalement de trois éléments qui contribuent ensemble à la production de la parole. Ces éléments dont le contrôle et la coordination sont assurés par le système nerveux central, sont :

- les poumons :ils fournissent l'énergie (l'air) nécessaire à la production du son.
- le larynx : son rôle est la production des sons. C'est un ensemble de cartilages articulés comprenant les deux "cordes vocales". Ces dernières sont des organes vibratoires constituées de tissu musculaire et de tissu conjonctif résistant.
- le conduit vocal : c'est le conduit entre le larynx et les lèvres, il est composé de plusieurs cavités reliées entre elles. On retrouve la cavité pharyngale (le pharynx), la cavité nasale (les fosses nasales), la cavité buccale (la bouche) et la cavité labiale (les lèvres).

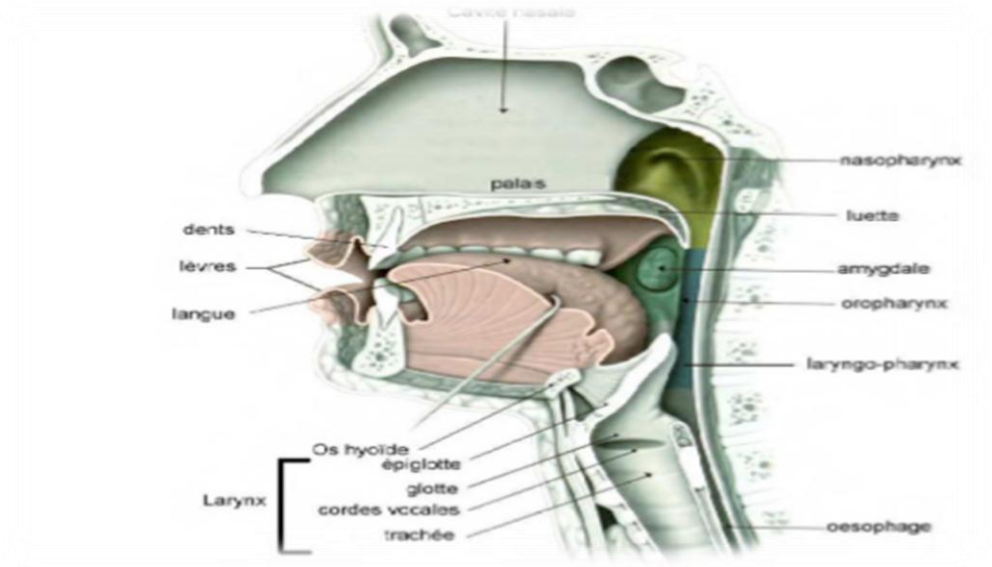

Figure 2.1 : Vue schématique de l'appareil vocal.

Bases fréquentielles de l'appareil phonatoire

La fréquence fondamentale (ou pitch) de la voix est propre à chaque individu. Elle est fonction de différents paramètres physiologiques tels que le volume et la masse de la glotte, la section de la trachée, sa longueur etc.

- Pour les hommes, cette fréquence fondamentale F0 se situe aux environs des 100 Hz,
- Pour les femmes, cette fréquence F0 se situe plutôt aux environs des 200Hz.
- Pour les enfants, cette fréquence F0 se situe plutôt aux environs des 300 à 400 Hz

La fréquence fondamentale F0 n'est pas dominante par rapport à ses harmoniques. L'énergie est répartie sur différentes fréquences supérieures de manière légèrement différente, d'un individu à l'autre pour des sons ou phonèmes équivalents.

Importance du larynx

Le larynx ou la pression de l'air est modulée avant d'être appliquée au conduit vocal. Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée (Figure 2.2). Les cordes vocales sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx et, en s'écartant progressivement, déterminer une ouverture triangulaire appelée glotte. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non-voisés (ou sourds7). Les sons voisés (ou sonores) résultent au contraire d'une vibration périodique des cordes vocales. Le larynx est d'abord complètement fermé, ce qui accroît la pression en amont des cordes vocales, et les force à s'ouvrir, ce qui fait tomber la pression, et permet aux cordes vocales de se refermer; des impulsions périodiques de pression sont ainsi appliquées au conduit vocal, composé des cavités pharyngienne et buccale pour la plupart des sons. Lorsque la luette est en position basse, la cavité nasale vient s'y ajouter en dérivation. Notons pour terminer le rôle prépondérant de la langue dans le processus phonatoire. Sa hauteur détermine la hauteur du pharynx : plus la langue est basse, plus le pharynx est court. Elle détermine aussi le lieu d'articulation, région de rétrécissement maximal du canal buccal, ainsi que l'aperture, écartement des organes au point d'articulation [6].

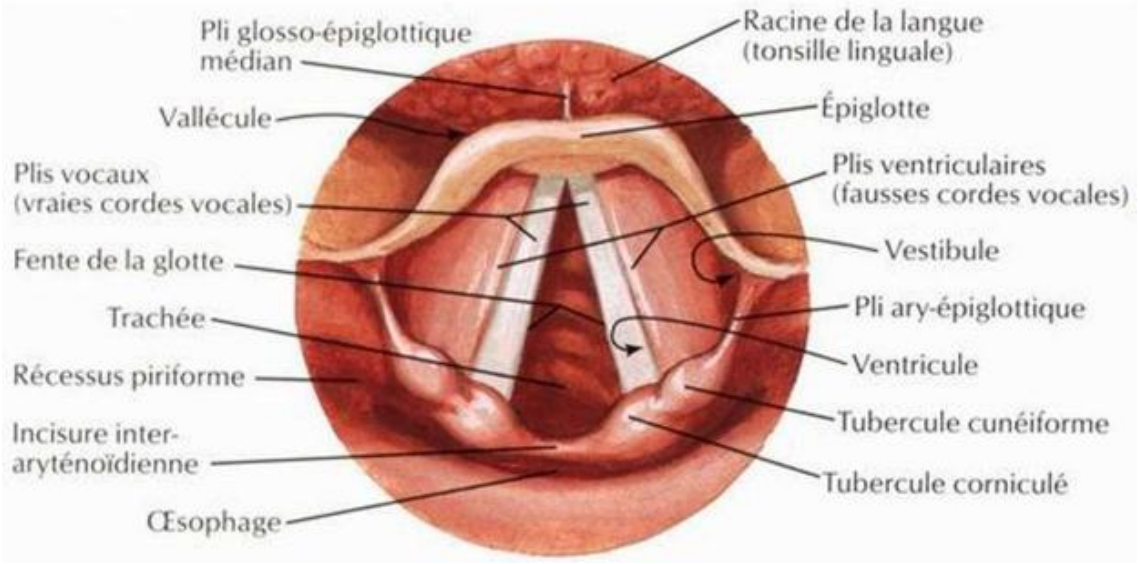

Figure 2.2 : Section du larynx, vu de haut.

Phonèmes

Le phonème est la plus petite unité présente dans la parole et susceptible par sa présence de changer la signification d'un mot. Le nombre de phonèmes est toujours très limité, en générale il est inférieur à 50.

La notion de phonème ne tient compte que des caractéristiques acoustiques qui permettent une distinction entre des mots, elle ne tient pas compte des phénomènes physiques de la production du son.

Classification des sons de la parole

Voisé

Le flux d'air est découpé en un train d'impulsions quasi périodique qui résonne dans les différentes cavités. Physiquement, le train d'impulsion quasi périodique subit une modulation en fréquence en passant par les différentes cavités.

Non-voisé

Les cordes vocales sont relâchées, l'air passe librement au niveau du larynx sans les faire vibrer.

Figure 2.3 : Exemples de son voisé et non-voisé .

Production de la parole

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

- **Générateur :** l'air expulsé des poumons passe à travers l'appareil vocal en tant qu'instrument à vent et crée la pression nécessaire pour générer un signal acoustique.
- **Vibrateur :** l'air expulsé des poumons traverse la trachée pour arriver dans le larynx où se trouvent les cordes vocales. Les cordes vocales sont une paire de muscles dont la longueur moyenne se situe entre 20 et 25 millimètres. Cette longueur varie cependant d'un individu à l'autre. L'air traversant le larynx met en vibration les cordes vocales. La fréquence de vibration des cordes vocales est modulée en fonction de leur degré de contraction. Le locuteur peut ainsi moduler la hauteur des sons qu'il émet.
- **Résonateur** : Les vibrations des cordes vocales sont modifiées par le passage de l'air dans les différentes cavités qui composent le pharynx et dans les fosses nasales, la bouche et le larynx avec lesquels il communique. Ces résonateurs influent sur le son en atténuant certaines fréquences et en amplifiant d'autres. La forme et le volume de ces cavités, spécifiques au locuteur, modifient fortement le son produit.
- **Modulateurs** : Enfin les organes modulateurs que sont la langue, les lèvres et la mâchoire sculptent le son pour produire les phonèmes qui composent la parole. La position de ces différents organes est le mécanisme final qui permet la production de parole articulée [6] [7].

Mécanisme de production de la parole

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

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

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

3. Articulation et l'exécution motrice de la parole: elle consiste en l'exécution de la séquence articulatoire correspondant au message. Dans cette étape, le locuteur exécute une série de signaux neuromusculaires qui servent de commandes et contrôlent les cordes vocales, les lèvres, la mâchoire, la langue et le vélum (voile du palais), produisant ainsi la séquence sonore souhaitée à la sortie [7].

Propriétés acoustiques du conduit vocal

Continuité

La production d'un son est fortement influencée par les sons qui le précédent et le suivie à cause de l'anticipation du geste articulatoire. L'identification correcte d'un segment de parole isolé de son contexte est parfois impossible. Certainement, il est plus facile de reconnaître des mots isolés bien séparés par des périodes de silence que de reconnaître la séquence de mots constituant une phrase. En fait, dans ce dernier cas, non seulement la frontière entre les mots n'est plus connue, mais, en outre, les mots deviennent fortement articulés [7].

Variabilité de la parole

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

variabilité spécifique au locuteur en fonction de son état physique mais également psychologique. C'est la variabilité intra-locuteur. De plus, les conditions environnementales influencent l'onde acoustique du signal de parole. Les bruits ambiants additifs ou de convolutions causées par l'enregistrement sonore modifient aussi le signal de parole [7].

Redondance

La redondance naturelle du signal de parole permet de réduire très fortement le débit binaire dans une très large mesure, au prix d'un traitement plus ou moins complexes et du risque d'une certaine dégradation de la qualité de la représentation [7].

Audition

L'appareil phonatoire, émetteur d'informations, ne serait d'aucune utilité si l'information générée ne pouvait être captée et analysée par un récepteur. Parmi tous les récepteurs existants, l'homme a acquis la capacité de découvrir le sens caché sous les sons produits par son interlocuteur. L'oreille, organe récepteur de l'information sonore, et les capacités de perception qui caractérisent cet organe lorsqu'il est en parfait état et n'a subi aucune atteinte venue amoindrir ses capacités intrinsèques. L'oreille est divisée en trois parties distinctes, cette division se faisant en fonction de la distance par rapport à l'environnement aérien, porteur des sons. Une première partie, l'oreille externe, correspond à la partie visible de l'organe, pavillon et lobe, à laquelle est rattaché le conduit auditif externe qui permet de propager le son jusqu'au tympan. Le tympan marque la frontière entre l'oreille externe et l'oreille moyenne. Les organes de l'oreille moyenne permettent de transformer les sons en vibrations grâce au contact qu'ils ont avec le tympan. Ces vibrations, une fois générées, sont transmises à la cochlée qui constitue l'organe majeur de l'oreille interne. La cochlée permet de transformer les vibrations en influx nerveux par le biais de cellules ciliées qui captent les vibrations produites dans le fluide de la membrane basilaire par l'étrier, le dernier os de l'oreille moyenne. Cet influx nerveux est alors transmis au cerveau en charge du traitement.

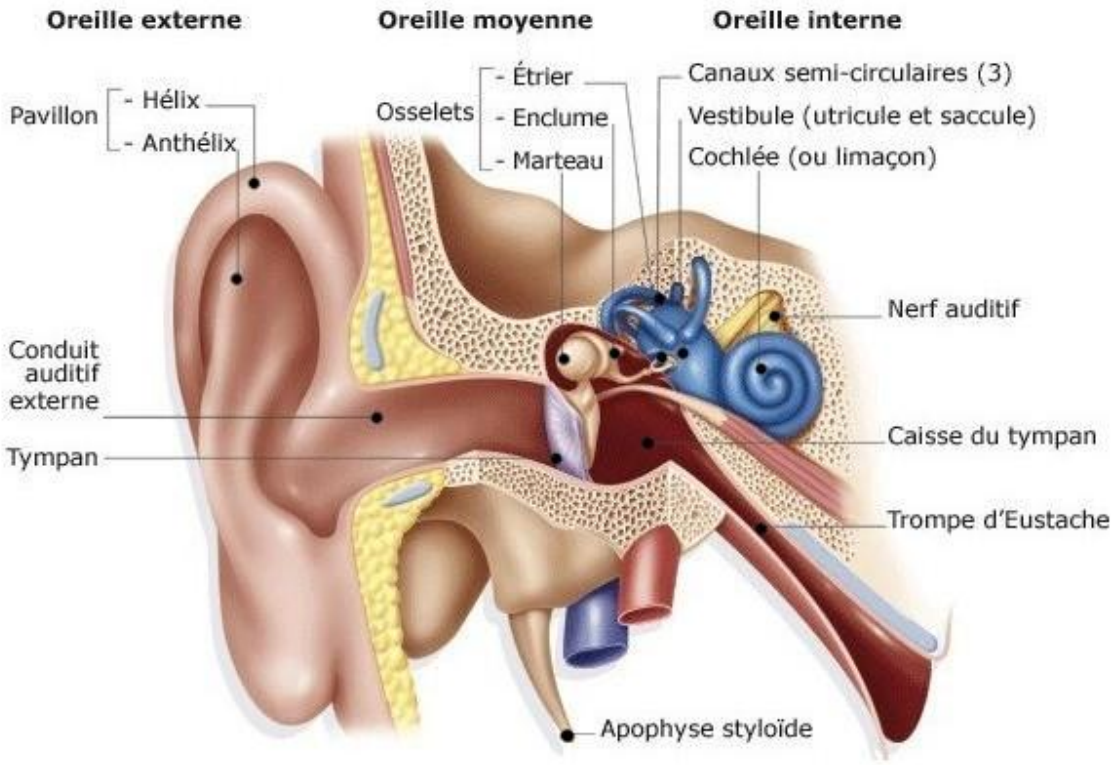

Figure 2.4 : L'oreille humaine.

Identité du locuteur

L'identité vocale d'un locuteur est largement définie par l'anatomie de son appareil de parole. Le ton fondamental de sa voix est défini par la taille et la masse de ses cordes vocales (ainsi que par la tension qu'il leur applique). La longueur de son tract vocal définit une échelle de grandeur acoustique que l'auditeur associera à la taille du locuteur. En combinant ces deux informations, il est possible d'estimer la taille, l'âge et le genre du locuteur à partir de sa voix. Cependant, la fréquence fondamentale est rarement fixe et varie pour participer à la prosodie. De même, la longueur du tract vocal varie légèrement pour transmettre certaines émotions.

Prétraitement du signal de parole

Échantillonnage du signal

L'échantillonnage consiste à transformer une fonction $x(t)$ à valeurs continues en une fonction $\hat{x}(t)$ discrète constituée par la suite des valeurs $x(t)$ aux instants d'échantillonnage $t = nT$ avec n un entier naturel et T est la période d'échantillonnage. Le choix de la fréquence d'échantillonnage n'est pas aléatoire car une petite fréquence nous donne une présentation pauvre du signal. Par contre une très grande fréquence nous donne des mêmes valeurs, redondance, de certains échantillons voisins donc il faut prélever suffisamment de valeurs pour ne pas perdre

l'information contenue dans $x(t)$. Le théorème suivant traite cette problématique : La fréquence d'échantillonnage assurant un non repliement du spectre doit être supérieure à 2 fois la fréquence haute du spectre du signal analogique : $F_e = 1/T = 2 \times F_{max}$ Par contre pour le signal audio (parole), on exige une bonne représentation du signal jusqu'à 20 kHz et l'on utilise des fréquences d'échantillonnage de 44.1 ou 48 kHz. Pour les applications multimédia, les fréquences sous-multiples de 44.1 kHz sont de plus en plus utilisées : 22.5 kHz, 11.25 kHz [6].

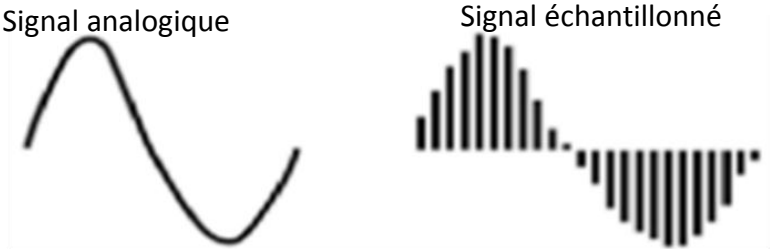

Figure 2.5 : Échantillonnage d'un signal audio.

Filtrage

Le coût d'un traitement numérique, filtrage, transmission, ou simplement enregistrement peut être réduit d'une façon notable si l'on accepte une limitation du spectre par un filtrage préalable. C'est le rôle du filtre de garde, dont la fréquence de coupure f_c est choisie en fonction de la fréquence d'échantillonnage retenue. Pour le signal audio (parole et musique), on exige une bonne représentation du signal jusque 20 kHz et l'on utilise des fréquences d'échantillonnage de 44.1 ou 48 kHz. Pour les applications multimédia, les fréquences sous-multiples de 44.1 kHz sont de plus en plus utilisées : 22.5 kHz, 11.25 kHz.

Fenêtrage

Généralement le découpage du signal dans le domaine temporel équivaut à multiplier le signal par une fonction rectangulaire, ce qui équivaut à une convolution dans le domaine fréquentiel entre le spectre du signal analysé et celui de la fenêtre. Dans la majorité des cas la fenêtre rectangulaire s'avère trop brutale. En effet il a été démontré [8] que toute variation rapide dans le domaine temporel correspond à des hautes fréquences dans le domaine fréquentiel qui se traduit par des ondulations sur le spectre. Alors on lui préfère d'autres fenêtres (voir Figure 2.6) plus douces. Parmi les fenêtres les plus utilisées on trouve :

Rectangulaire:

$$w(n) = \begin{cases} 1 \; pour \; 0 \leq n \leq N-1 \\ 0 \; ailleurs \end{cases}$$

Bartlett:

$$w(n) = \begin{cases} \dfrac{2n}{N-1} pour\ 0 \le n \le (N-1)/2 \\ 2 - \dfrac{2n}{(N-1)} pour\ (N-1)/2 \le n \le N-1 \\ 0\ ailleurs \end{cases}$$

Hanning :

$$w(n) = \begin{cases} 0.5 - 0.5\cos\left(\dfrac{2\pi n}{(N-1)}\right) pour\ 0 \le n \le N-1 \\ 0\ ailleurs \end{cases}$$

Hamming :

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{N-1}\right) pour\ 0 \le n \le N-1 \\ 0\ ailleurs \end{cases}$$

Blackman :

$$w(n) = \begin{cases} 0.42 - 0.5\ \cos\left(\dfrac{2\pi n}{N-1}\right) + 0.8\cos\left(\dfrac{4\pi n}{N-1}\right) pour\ 0 \le n \le N-1 \\ 0\ ailleurs \end{cases}$$

où N représente la longueur de la fenêtre, et n un échantillon du signal.

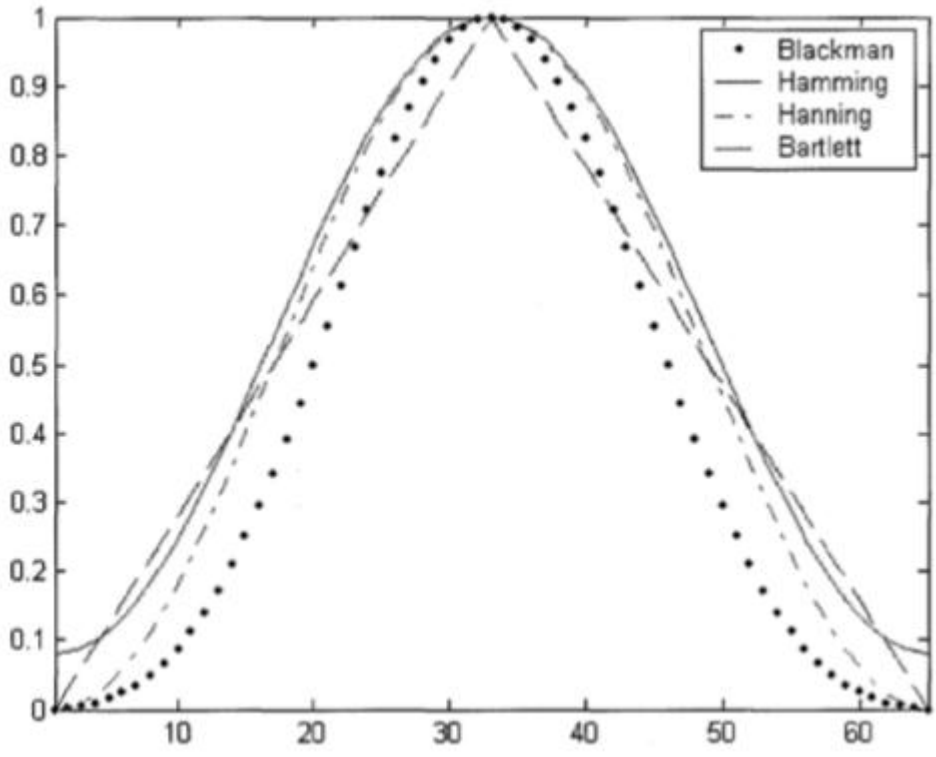

Figure 2.6 : Exemples de fenêtres de pondération.

En pratique la fenêtre de Hamming, est souvent la plus utilisée, la figure 2.7 est une illustration de son spectre et celui de la fenêtre rectangulaire. Dans cette figure on voit clairement que la fenêtre de Hamming permet une grande atténuation en dehors de la bande passante comparativement à la fenêtre rectangulaire d'où son avantage.

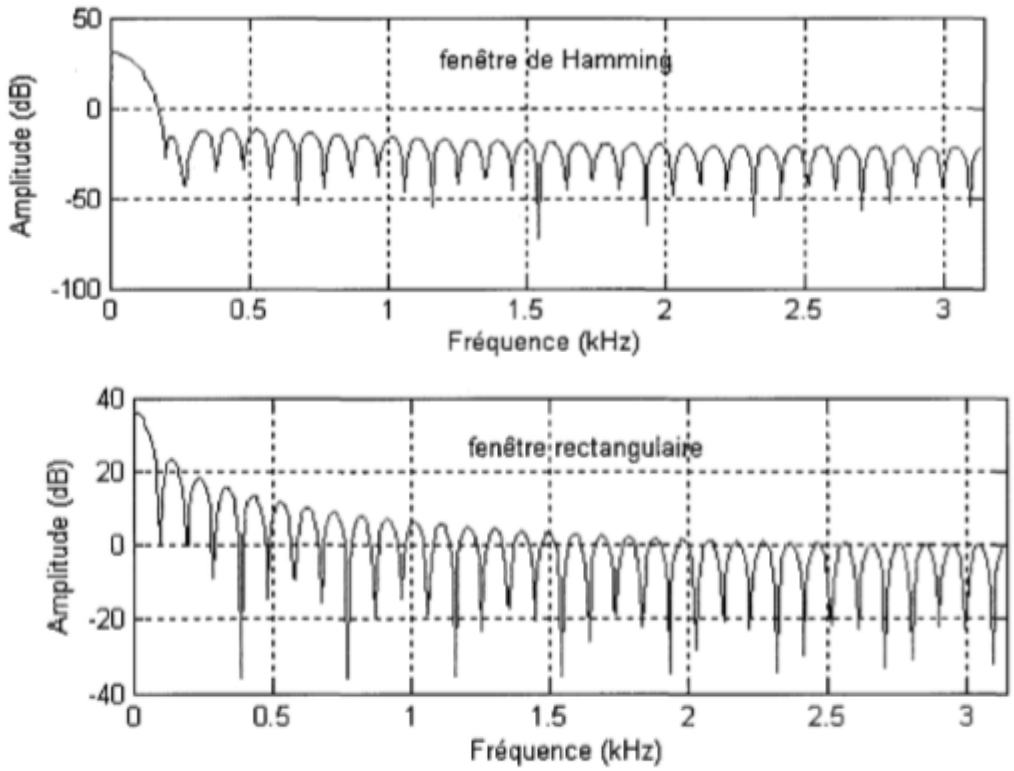

Figure 2.7 : Spectre des fenêtres de Hamming et Rectangulaire

Spectre

Il est souvent intéressant de représenter l'évolution temporelle du spectre à court terme d'un signal, sous la forme d'un spectrogramme. L'amplitude du spectre y apparaît sous la forme de niveaux de gris dans un diagramme en deux dimensions temps-fréquence. On parle de spectrogramme à large bande ou à bande étroite selon la durée de la fenêtre de pondération (Figure 2.8). Les spectrogrammes à bande large sont obtenus avec des fenêtres de pondération de faible durée (typiquement 10 ms); ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des formants. Les périodes voisées y apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite sont moins utilisés. Ils mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales [8].

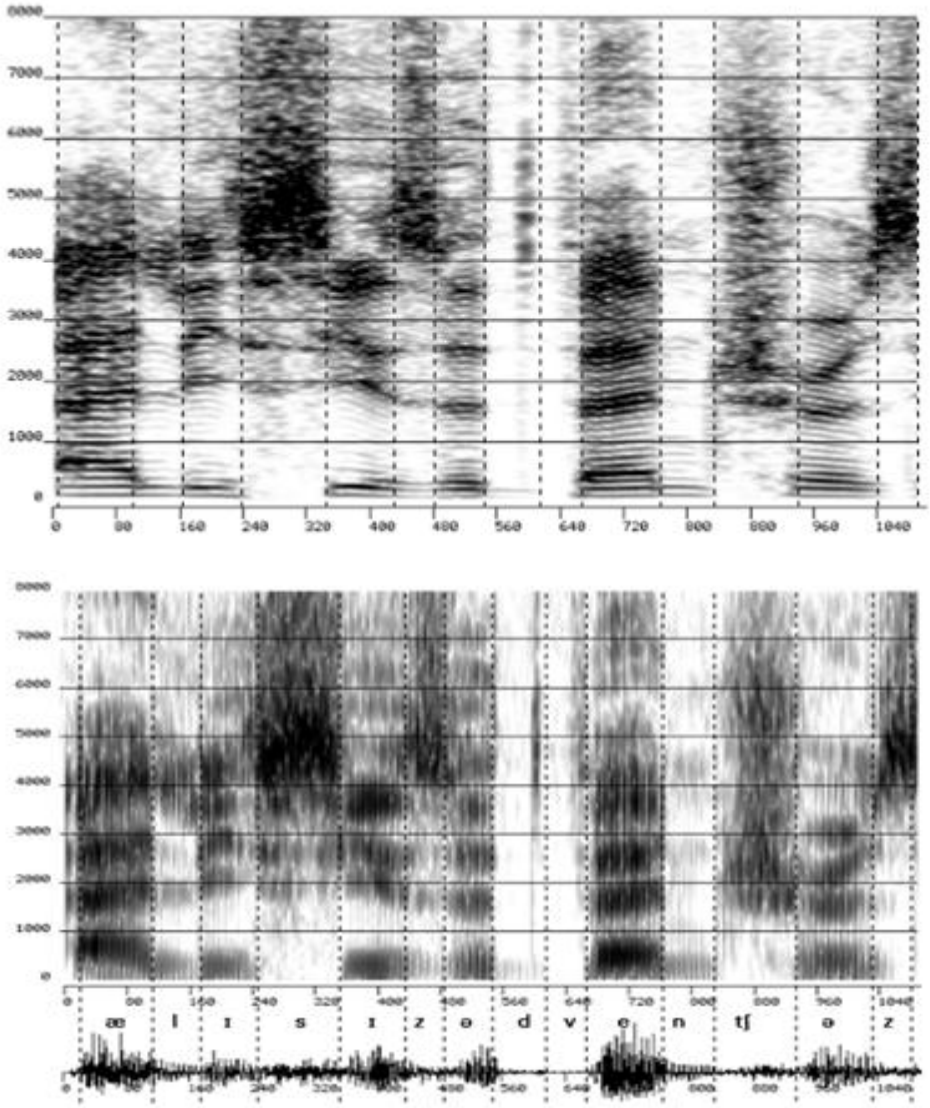

Figure 2.8 : Spectrogrammes à large bande (en bas), à bande étroite (en haut), et évolution temporelle de la phrase anglaise 'Alice's adventures', échantillonnée à 11.25 kHz (calcul avec fenêtres de Hamming de 10 et 30 ms respectivement).

Conclusion

Dans ce chapitre nous avons passé en revue le mécanisme de la production de la parole, le principe de son audition ainsi que les caractéristiques générales du signal vocal. Principalement on peut classer le signal vocal en deux catégories : les sons voisés, résultants de la vibration des cordes vocales, et les sons non voisés qui ne nécessitent pas l'intervention du larynx.

Chapitre 3 Extraction des caractéristiques d'un signal audio et classification par KNN

Introduction

Dans les chapitres précédents nous avons présenté les aspects généraux de la parole (production et audition), les différents outils nécessaires pour son traitement et sa paramétrisation, ainsi qu'un aperçu sur les principales approches de reconnaissance qu'on retrouve dans les différents systèmes de reconnaissance. Dans ce chapitre nous décrirons l'approche d'apprentissage automatique pour identifier des personnes à partir des caractéristiques extraites de la parole enregistrée.

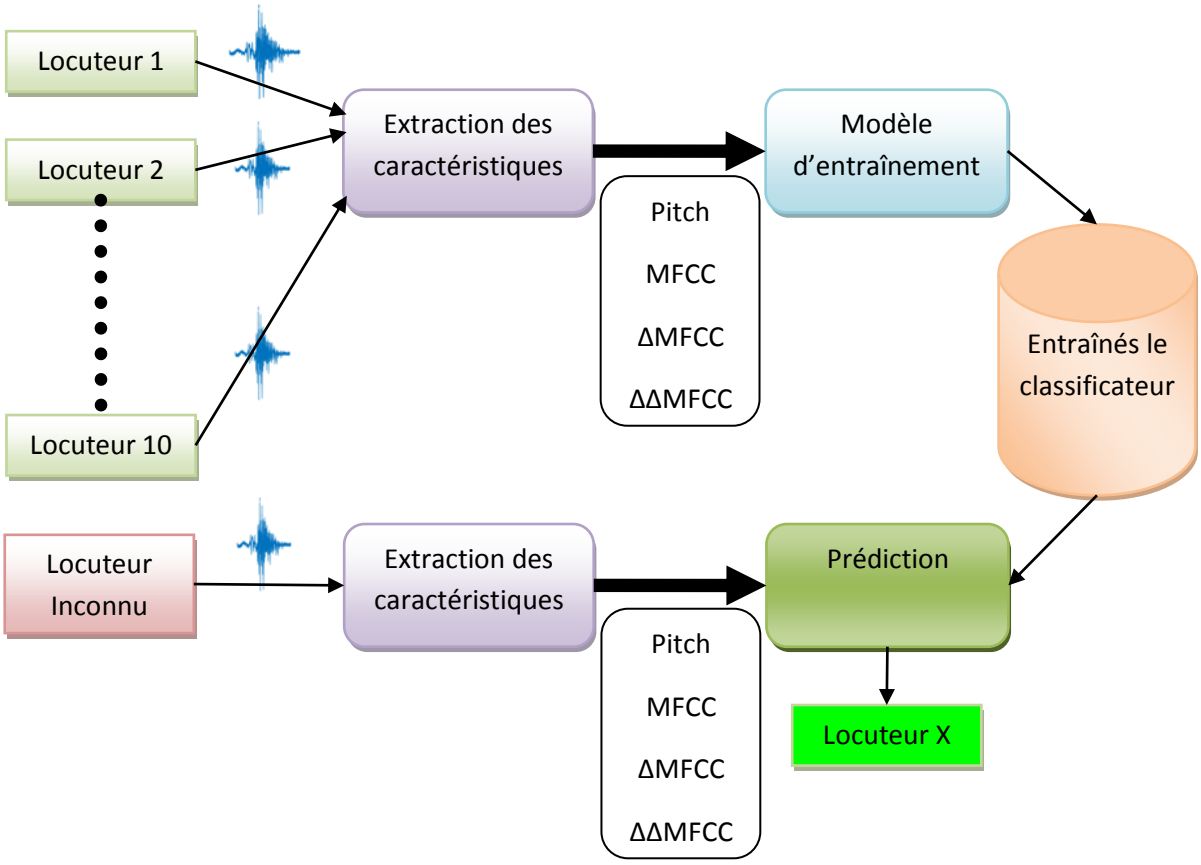

Figure 3.1 : Diagramme de la procédure d'identification du locuteur.

Les caractéristiques utilisées pour former le classificateur sont : le pitch, les MFCC, les ΔMFCC et les ΔΔMFCC. La démarche utilisée dans cette partie expérimentale pour l'identification du locuteur est illustrée dans le diagramme de la figure 3.1.
Le pitch, les MFCC, les ΔMFCC et les ΔΔMFCC sont extraits des signaux vocaux enregistrés pour 10 locuteurs. Ces caractéristiques sont utilisées pour former un classificateur K-Nearest Neighbor (KNN). Ensuite, les nouveaux signaux vocaux qui doivent être classifiés passent par la même extraction de caractéristiques. Le classificateur KNN formé prédit lequel des dix locuteurs est le plus proche.

Caractéristiques utilisées pour la classification

Cette section traite le pitch, les MFCCs et leurs drivées, le tout est utilisé pour classifier les locuteurs.

Pitch

La parole peut être classée en deux grandes catégories : voisée et non voisée. Dans le cas de la parole voisée, l'air des poumons est modulé par les cordes vocales et produit une excitation quasi périodique. Le son qui en résulte est dominé par une oscillation de fréquence relativement basse, appelée hauteur (pitch). Dans le cas d'une parole non voisée, l'air des poumons passe à travers une constriction du conduit vocal et devient une excitation turbulente et bruyante. Dans le modèle source-filtre de la parole, l'excitation est appelée la source, et le conduit vocal est appelé le filtre. La caractérisation de la source est une partie importante de la caractérisation du système vocal.

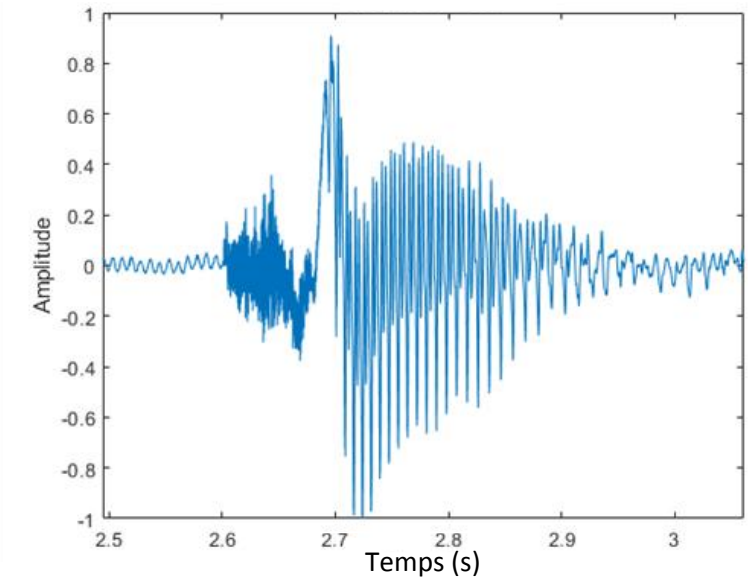

Figure 3.2 : Représentation dans le domaine temporel du mot anglais « two ».

Comme exemple de voisement ou non voisement, on considère une représentation dans le domaine temporel du mot en anglais "TWO" (/T UW/) (Figure 3.2). La consonne /T/ (parole non voisée) ressemble à du bruit, tandis que la voyelle /UW/ (parole voisée) est caractérisée par une fréquence fondamentale forte.
La méthode la plus simple pour faire la distinction entre le son voisée et non voisée est d'analyser le taux de passage par zéro. Un grand nombre de passages à zéro implique qu'il n'y a pas d'oscillation dominante dans les basses fréquences.

Une fois que nous avons isolé une région voisée, nous pouvons la caractériser en estimant le pitch. Dans cette expérience on utilise l'approche d'autocorrélation normalisée pour calculer le pitch. En appliquant l'algorithme d'autocorrélation sur le mot "two", on peut tracer le contour du pitch (voir figure ci-dessous).

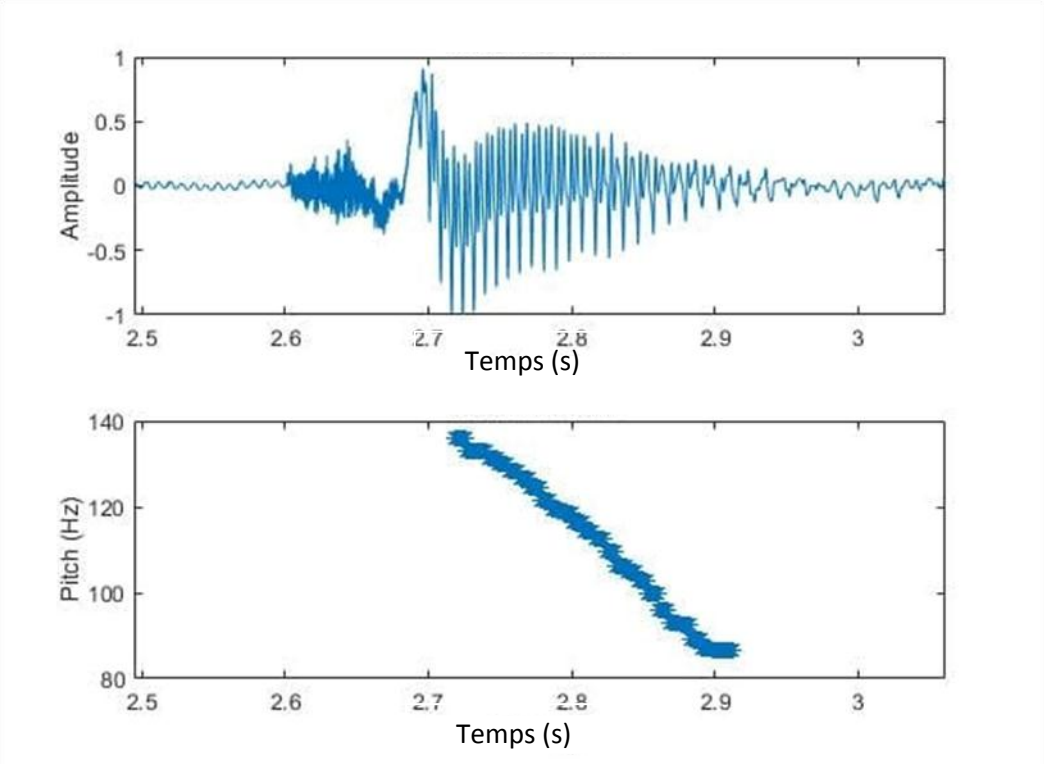

Figure 3.3 : Pitch.

MFCC

Les MFCCs sont des caractéristiques populaires extraites des signaux vocaux pour l'utilisation dans les tâches de reconnaissance. Dans le modèle source-filtre de la parole, les MFCCs représentent le filtre (conduit vocal). La réponse en fréquence de l'appareil vocal est relativement lisse, alors que la source de la voix peut être modélisée comme un train d'impulsions. Le résultat est que le conduit vocal peut être estimé par l'enveloppe spectrale d'un segment de parole.

L'idée motivante du MFCC est de comprimer l'information sur le conduit vocal (spectre lissé) en un petit nombre de coefficients basés sur une compréhension de la cochlée.

Bien qu'il n'existe pas de norme stricte pour le calcul du MFCC, les étapes de base sont décrites dans le chapitre 1.

Un signal vocal est de nature dynamique et change avec le temps. On suppose que les signaux vocaux sont stationnaires sur des échelles de temps courtes et que leur traitement s'effectue dans des fenêtres de 20 à 40 ms (cf. chapitre 1). Cet exemple utilise une fenêtre de 30 ms avec un chevauchement de 75%.

Base de données

Dans notre travail, nous avons utilisé la base de données Census (également appelée base de données AN4) du groupe CMU Robust Speech Recognition [9].

Cette base de données alphanumériques AN4 a été enregistrée à l'Université Carnegie Mellon vers 1991. Elle est décrite en détail dans [10].

La base de données utilisée contient 1018 énoncés pour apprentissage et 140 énoncés pour test, tandis que la base de données fournie ici contient 948 énoncés de formation et 130 énoncés pour test.

Toutes les données sont échantillonnées à 16 kHz (échantillonnage linéaire 16 bits). Tous les enregistrements ont été réalisés avec un microphone parlant rapproché.

Chacun des fichiers tar compressés contient des données dans l'un des formats suivants :

*.raw : fichiers audio au format raw (PCM linéaire, sans en-tête).

*.sph : fichiers audio au format Sphere du NIST.

*.mfc : fichiers audio codés en coefficients mélographiques.

L'ensemble de données contient des enregistrements de sujets masculins et féminins parlant des mots et des chiffres. Les fichiers de parole sont partitionnés en sous-répertoires en fonction des étiquettes correspondant aux locuteurs.

Le stockage des données est divisé en deux parties. 80 % des données de chaque étiquette sont utilisées pour l'entrainement et les 20 % restants sont utilisés pour les tests. L'étiquette identifie le locuteur.

Entraînement du classificateur

Après la collecte des caractéristiques des dix locuteurs, nous pouvons former un classificateur basé sur ces dernières. Nous utilisons un classificateur de type KNN. Le KNN est une technique de classification naturellement adaptée à la classification multi-classe. Les hyperparamètres du classificateur du voisin le plus proche comprennent le nombre de voisins les plus proches, la mesure de distance utilisée pour calculer la distance aux voisins et le poids de la mesure de distance. Les hyperparamètres sont sélectionnés pour optimiser la précision de validation et les performances sur l'ensemble de test. Dans cet exemple, le nombre de voisins est fixé à 5 et la métrique de la distance choisie est la distance Euclidienne pondérée inversement carrée.

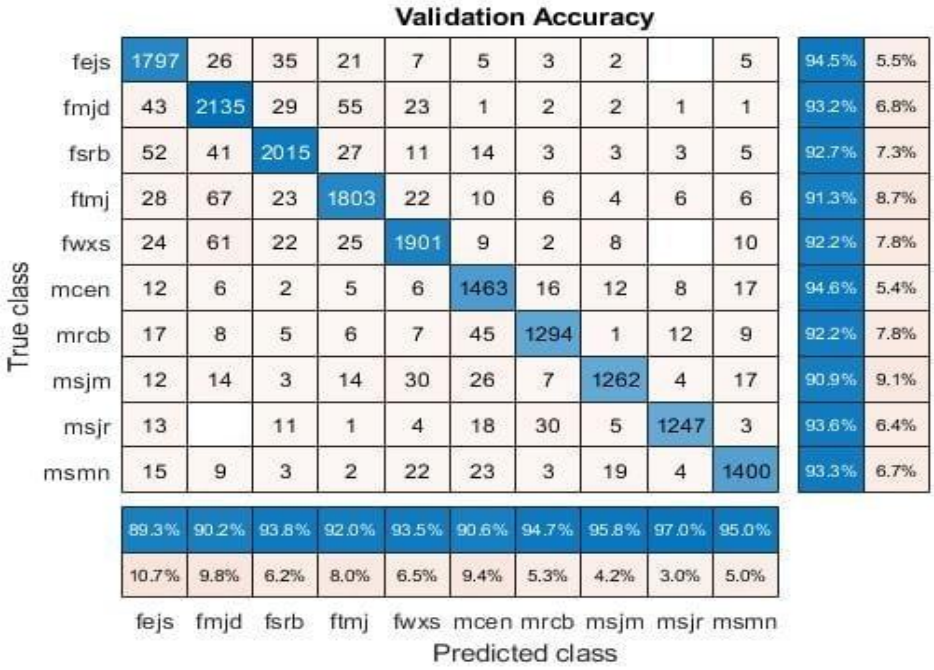

Figure 3.4 : Matrice de confusion

3.1 Résultat de simulation

À travers la simulation que nous avons faite avec MATLAB en utilisant la base de données AN4, nous avons à chaque fois choisie une combinaison entre coefficients: MFCC, ΔMFCC et ΔΔMFCC, ainsi que le pitch. Aussi, nous avons joué sur le nombre des ces coefficients. Tout cela a été résumé à travers des histogrammes extraits par l'excel.

La précision ou accuracy en anglais est choisie comme critère d'évaluation dans nos expériences. Cette dernière est exprimée en pourcentage %.

Remarque : dans ce qui suit, nous avons opté pour l'appellation anglaise Accuracy.

1er cas

Calcul de l'accuracy pour 13, 26 et 41 MFCC avec $\boldsymbol{k = 1, \dots, 5}$ pour le classifieur.

Tableau 3.1 : Accuracy (%) pour différents nombres des MFCC.

Nombre des coefficients	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
13 premiers	92.8	89.6	90.2	88.9	88.7
26 premiers	96.4	94.3	94.4	93.6	93.1
41 premiers	**96.6**	94	94.1	93.3	92.8

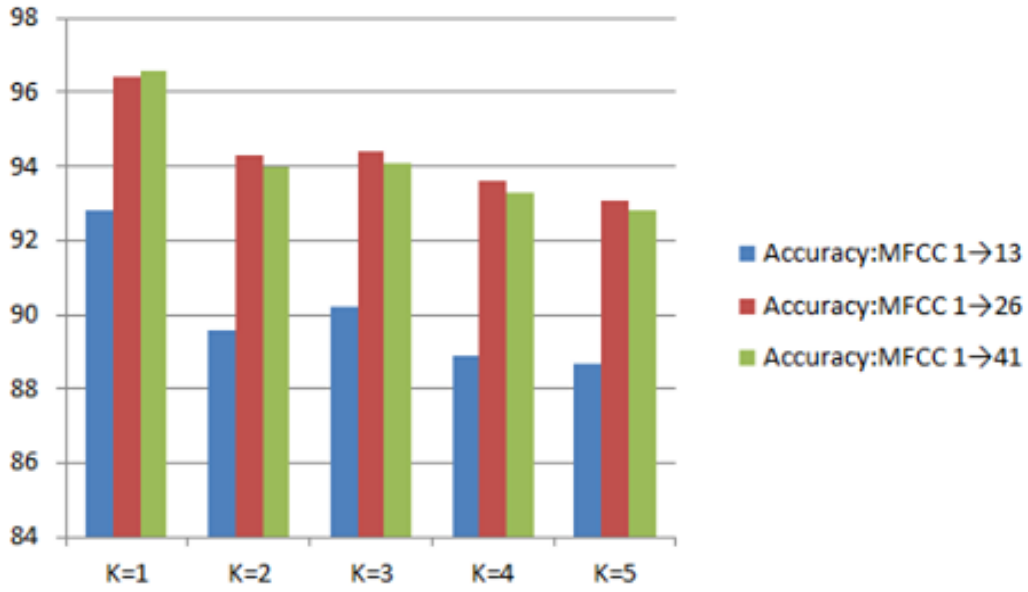

Figure 3.5 : Histogramme des accuracy avec 13,26 et 41 MFCC.

On observe que plus on augmente le nombre des MFCC et plus la précision (accuracy) est meilleure. Aussi, plus le k augmente, la précision diminue (exemple : pour MFCC de 1 à 13 et $k = 5$ l'accuracy = 88.7% et pour MFCC de 1 à 41 et $k = 1$ l'accuracy = 96.6% donc plus le nombre des MFCCs est grand et k petit le système plus précis.

2ème cas

Calcul de l'accuracy pour 13, 26, 41 ΔMFCC avec $k = 1, \dots 5$.

Tableau 3.2 : Accuracy (%) pour différents nombres des ΔMFCC.

Nombre des coefficients	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
13 premiers	51.9	43.7	43.2	42.6	41.5
26 premiers	77.7	64.7	63.6	60.1	57
41 premiers	85.5	73.5	71	66.9	62.8

On remarque ici que l'accuracy a diminuée par rapport au premier cas ce qui montre l'importance du MFCC sur le système pour la reconnaissance.

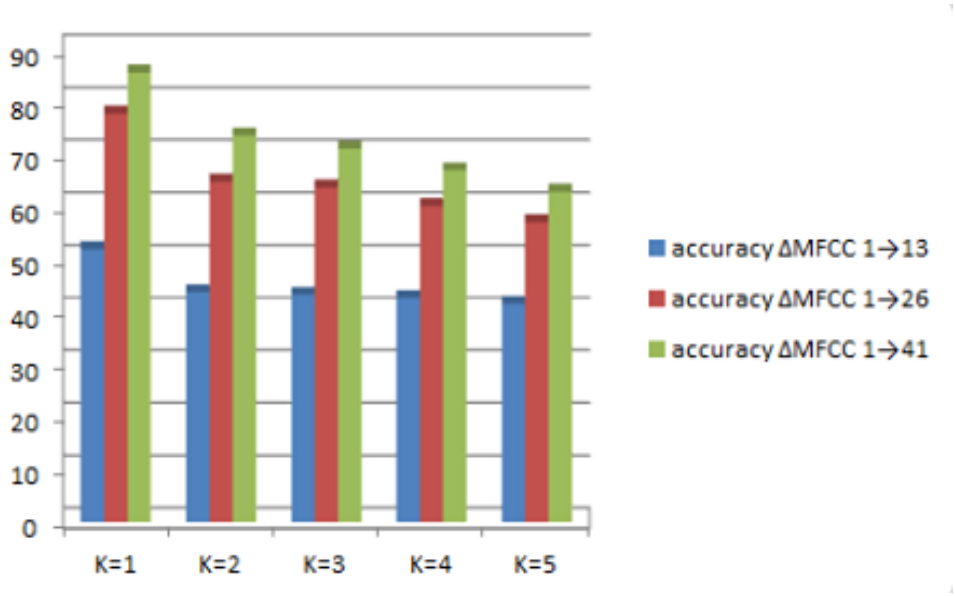

Figure 3.6 : Histogramme des acuuracy avec 13,26 et 41 ΔMFCC.

3ème cas

Calcul de l'accuracy pour 13, 26 et 41 ΔΔMFCC avec $k = 1, \dots 5$.

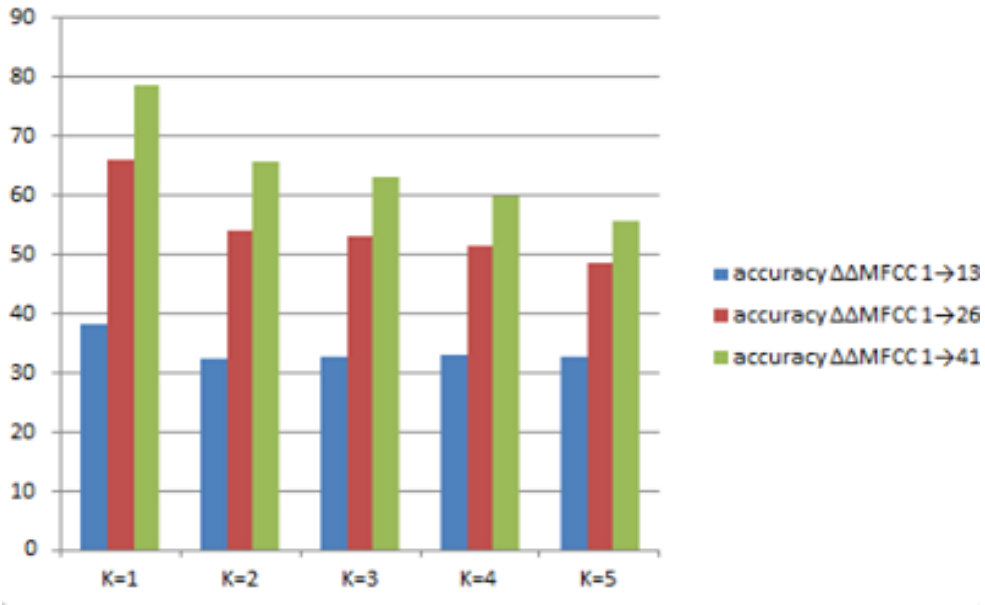

Figure 3.7 : Histogramme acuuracy avec 13,26 et 41ΔΔMFCC.

Lorsqu'on augmente les ΔΔMFCC de 13 jusqu'à 41 ΔΔMFCC l'accuracy est améliorée de 38% à 78,4%. En contre partie, lorsque k augment accuracy se diminue.

4ème cas

Calcul de l'accuracy pour 13, 26 et 41 MFCC + 13, 26,41 ΔMFCC avec $k = 1, \dots 5$.

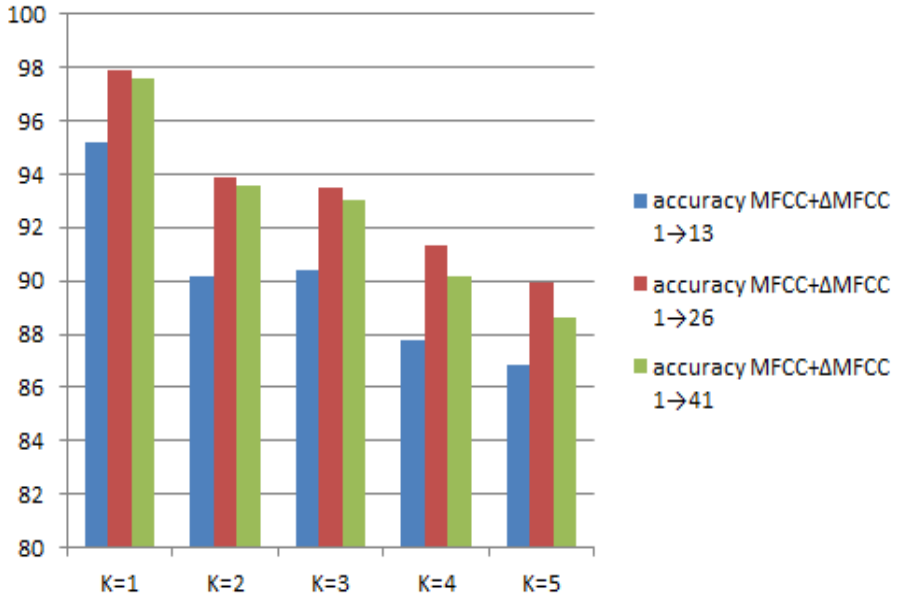

Figure 3.8 : Histogramme des acuuracy avec 13,26 et 41 MFCC et ΔMFCC.

Lorsque on augmente les ΔMFCC de 13 jusqu'à 41 MFCC, l'accuracy est améliorée de 95.2% à 97.6 %. Lorsque k augmente l'accuracy diminue.

Remarque : L'accuracy est meilleure avec 26 MFCC + 26 ΔMFCC = 97.9%.

5ème cas

Calcul de l'accuracy pour 13, 26, 41 MFCC + 13, 26, 41 ΔΔMFCC avec $k = 1, \dots 5$.

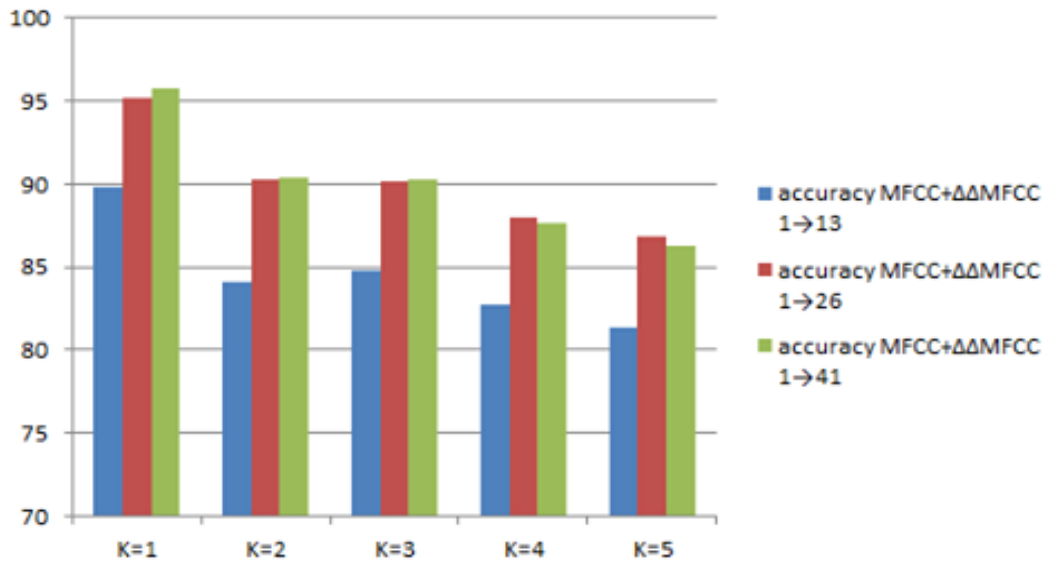

Figure 3.9 : Histogramme des acuuracy avec 13, 26 et 41 MFCC + ΔΔMFCC.

Lorsque on augmente les ΔMFCC de 13 jusqu'à 41 MFCC, accuracy s'améliore de 89.8% à 95.8%. Lorsque k augmente accuracy se diminue.

6ème cas

Calcul de l'accuracy pour 13, 26, 41 ΔMFCC + 13, 26, 41 ΔΔMFCC avec différents k.

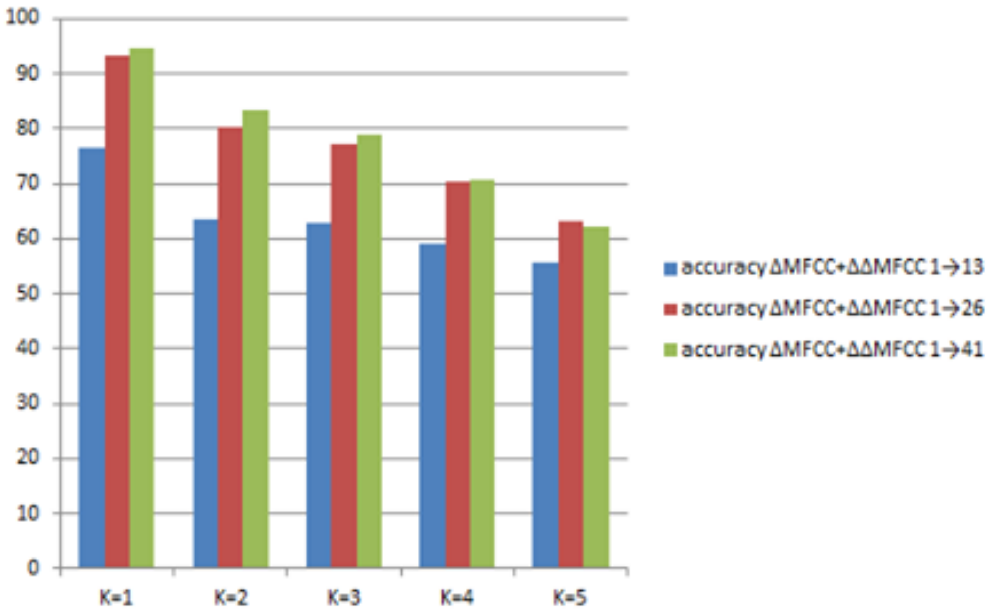

Figure 3.10 : Histogramme des acuuracy avec 13,26 et 41 ΔMFCC + ΔΔMFCC.

Lorsque on augmente les ΔMFCC + ΔΔMFCC de 13 jusqu'à 41 accuracy est amélioré de 76.4% à 94.5 %. Lorsque k augment accuracy se diminue.

7ème cas

Calcul de l'accuracy pour 13, 26, 41 MFCC + pitch, $k = 1, \dots 5$.

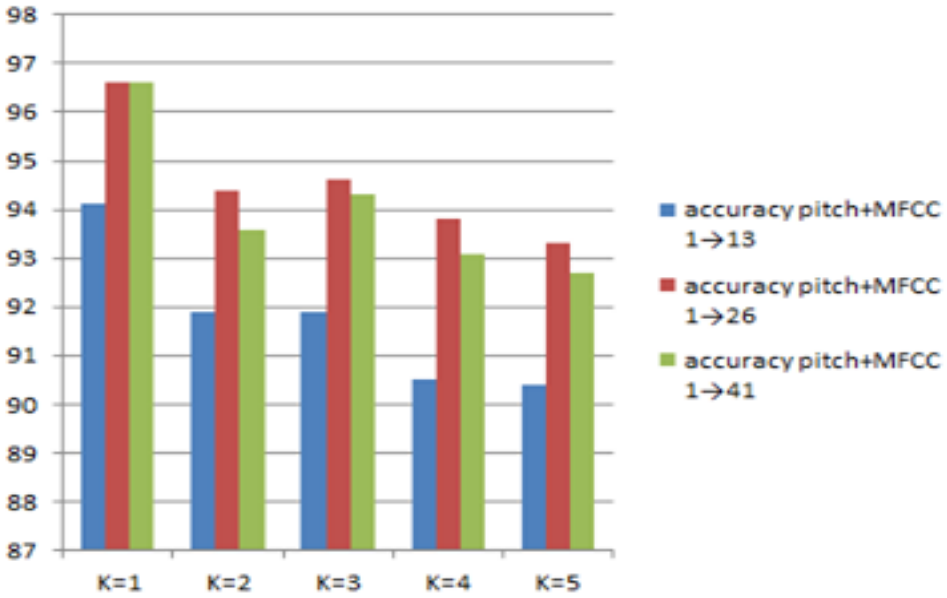

Figure 3.11 : Histogramme des acuuracy avec 13, 26 et 41 MFCC + pitch.

Lorsque on augmente le nombre des MFCC de 13 jusqu'à 41 MFCC + pitch, l'accuracy est améliorée de 94.1% à 96.6 %. Lorsque k augment accuracy se diminue.

Remarque : On obtient la même Accuracy obtenu pour 26 et 41 MFCC + pitch.

8ème cas

Calcul de l'accuracy pour 13 , 26, 41 ΔMFCC + pitch avec $k = 1, 2, \ldots 5$.

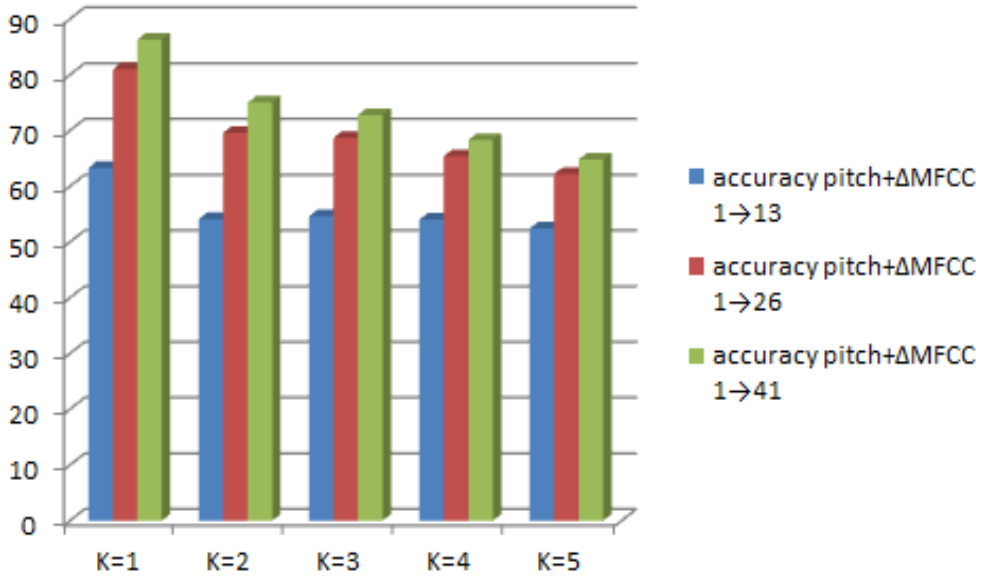

Figure 3.12 : Histogrammedes accuracy avec 13, 26, 41 ΔMFCC + pitch.

Lorsque on augmente les ΔMFCC de 13 jusqu'à 41 ΔMFCC, l'accuracy est améliorée de 63.3% à 86.3 %. Lorsque k augment accuracy se diminue.

9ème cas

Calcul de l'accuracy pour 13, 26, 41 ΔΔMFCC + pitch avec différents $k = 1, \ldots .5$.

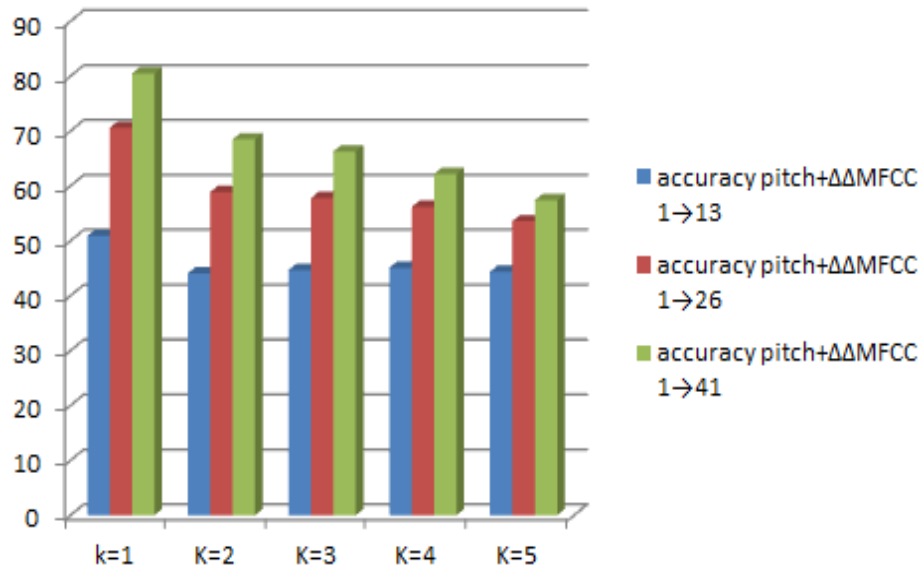

Figure 3.13 : Histogrammedes accuracy avec 13, 26, 41 ΔΔMFCC+ pitch.

Lorsque on augmente les ΔMFCC de 13 jusqu'à 41 ΔΔMFCC+ pitch, l'accuracy est améliorée de 50.9% à 80.5 %. Lorsque k augment accuracy se diminue.

10^{éme} cas

Calcul de l'accuracy pour 13, 26, 41 (ΔMFCC + ΔΔMFCC) + pitch avec différents $k = 1,2, \dots 5$.

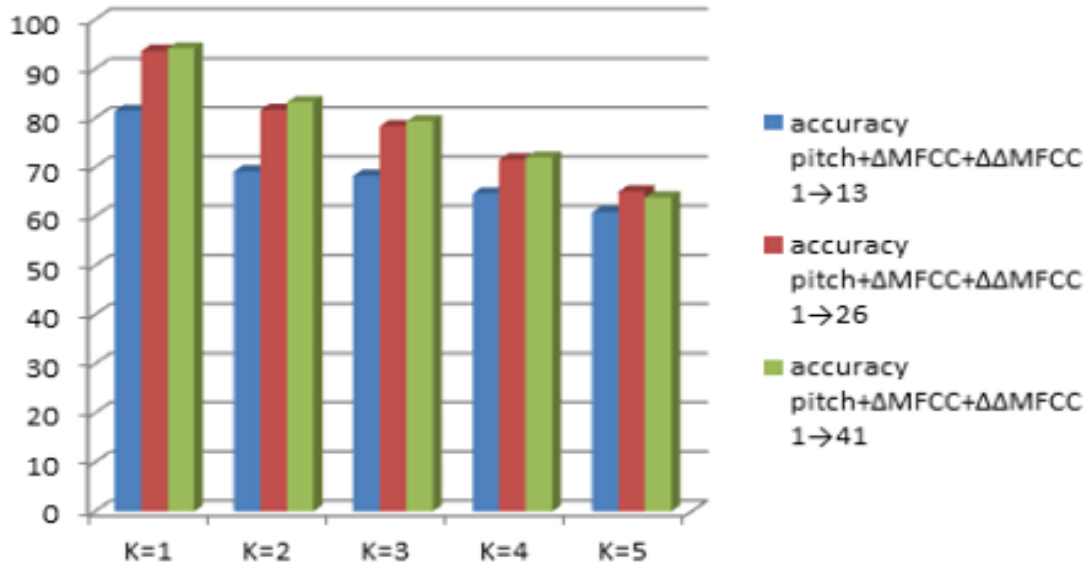

Figure 3.14 : Histogramme des accuracy avec 13, 26, 41 (ΔMFCC+ΔΔMFCC)+ pitch.

Lorsque on augmente les (ΔMFCC+ΔΔMFCC) de 13 jusqu'à 41 + pitch, l'accuracy est améliorée de 80.9% à 94.2%. Lorsque k augment accuracy se diminue.

11ème cas :

Calcul de l'accuracy pour 13, 26, 41 (MFCC + ΔMFCC + ΔΔMFCC) + pitch avec différents $k = 1,2, \dots 5$.

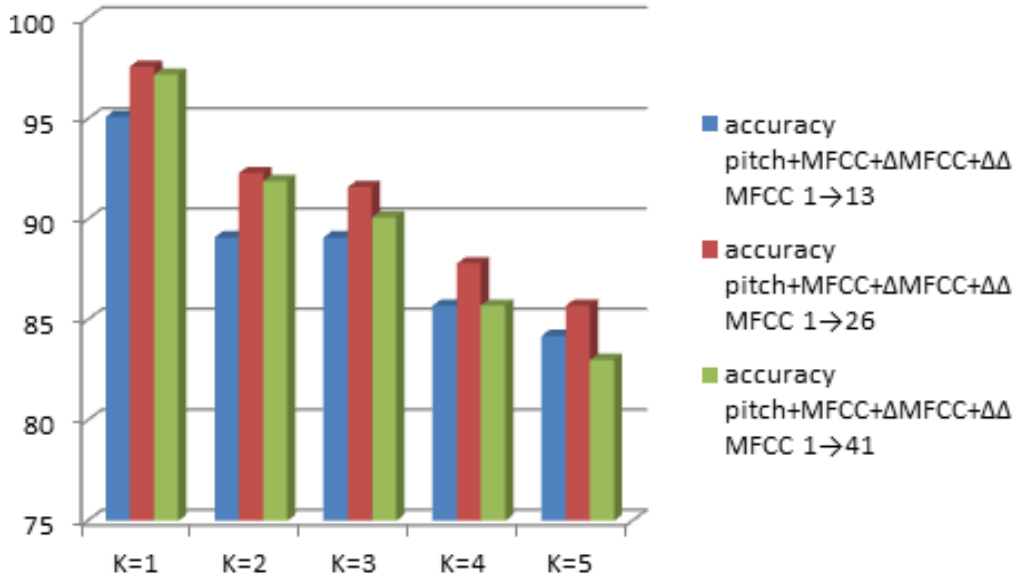

Figure 3.15 : Histogramme des accuracy avec 13, 26, 41 (MFCC+ΔΔMFCC+ΔMFCC)+ pitch.

Lorsque les fonctions MFCC et ΔMFCC et ΔΔMFCC sont utilisées, les performances sont généralement 20% plus grand à celles des fonctions MFCC, car elles fournissent des informations plus riches sur le contexte des trames.

Remarque : La meilleure accuracy est obtenu lorsque en combine 26 (MFCC+ ΔMFCC+ ΔΔMFCC) = **97.9%**. Et lorsque en utilise tous les paramètres l'accuracy = 97.6%.

Conclusion générale

Ce projet de fin d'étude traite de l'identification vocale dans le but de faire une authentification des personnes. Motivée par l'amélioration de la précision de la l'identification vocale de personne, ce travail se concentre sur l'exploitation de techniques de l'intelligence artificielle. Pour développer un tel système d'identification, nous avons opté pour le modèle K-plus proches Voisins k-ppv (k-Nearest Neighbor ou kNN, en anglais). Ce dernier a fait ses preuves dans le traitement de la parole.

Le système complet d'identification se base essentiellement sur une approche d'apprentissage automatique pour identifier des personnes à partir des caractéristiques extraites du signal audio. Les caractéristiques ainsi obtenues servent à former un classificateur qui permet l'identification vocale d'un locuteur. L'objectif principal de ce projet était de trouver les meilleurs paramètres (attribues) qui permettent d'avoir un taux élevé de la reconnaissance, à savoir une meilleure précision.

Nous avons débuté ce projet de fin d'étude par un chapitre expliquant reconnaissance automatique du locuteur ainsi que les outils utilisés pour effectuer une telle tâche. Puis, nous avons décrit au chapitre deux la production de la parole et son traitement, et finalement dans le chapitre trois nous avons testé notre système en utilisant la base de donnée AN4 par l'extraction de plusieurs paramètres vocaux (MFCC, pitch ….).

Des résultats expérimentaux ont permis de trouver la meilleure combinaison des paramètres qui améliore la précession de la reconnaissance (presque 98%).

Bibliographie

[1] J.P. Haton, C. cerisara, D. Fohr, Y. Laprie, and K. Smaili, "Reconnaissance automatique de la parole: du signal à son interprétation". Paris: Dunod, 2006.

[2] A. Harrag, "Extraction des données d'une base: Application à l'extraction des traits du locuteur", thèse doctorat, 2011.

[3] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4-37, January 2000.

[4] T. Dutoit, "Introduction au Traitement Automatique de la Parole", 2000.

[5] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 34, no. 1, pp. 52–59, February 1986.

[6] D. Jurafsky, J. H. Martin, "Speech and Language Processing", International Edition Paperback – 29 Apr 2008.

[7] Calliope, G. Fant, "La parole et son traitement automatique", décembre 1989.

[8] M. Benidir,"Théorie et traitement du signal", Collection : Sciences Sup, Dunod, juillet 2004.

[9] http://www.speech.cs.cmu.edu/databases/an4/

[10] A. Acero, "Acoustical and environmental robustness in automatic speech recognition", Kluwer Academic Publishers, 1993.