REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Université de Mohamed El-Bachir El-Ibrahimi - Bordj Bou Arreridj

Faculté *des Sciences et de la technologie*

Département d'Electronique

Mémoire

Présenté pour obtenir

LE DIPLOME DE MASTER

FILIERE : **Télécommunications.**

Spécialité : Systèmes des télécommunications.

Par

ROUAG KHALED DJIHED

KRAI RAMIA

Intitulé

Simulation et évaluation d'un système de segmentation en locuteurs d'un document audio

Évalué le : 03/07/2023

Par la commission d'évaluation composée de* :

Nom & Prénom	*Grade*	*Qualité*	*Etablissement*
S.SID.AHMED	**MCB**	**Président**	**Univ-BBA**
ASBAI NASSIM	**MCA**	**Encadreur**	**Univ-BBA**
Mme. BOUNAZOU HADJER	**Doctorante**	**Co-encadreur**	
A HACINE GHARBI	**MCA**	**Examinateur**	**Univ-BBA**

Année Universitaire 2022/2023

Dédicaces

A mes chers parents, pour tous leurs sacrifices, leurs amours, leurs tendresses, leurs soutiens, leurs confiances et leurs prières tout au long de mes études.

A Mon frère Charraf Eddine, mes sœurs MALAK et YASSMINE

A mes proches amis HALIS AYMEN et CHADI SALIM.

Mes amis d'enfance RAID et, MONAIM, ZIED

ROUAG KHALED DJIHED

Dédicaces

Je dédie ce travail

Pour maman et papa

A mes frères Samir, Amin, Rayan

A ma chère tante Widad

A mon ami Khaled DJIHED

À mes cousines

Pour toute la famille Krai

À tous ceux qui m'aiment

KRAI RAMIA

Remerciements

Loué soit Dieu, avec la grâce duquel les bonnes actions sont accomplies, qui nous a permis d'accomplir ce travail.

Un grand merci aux chers parents pour leurs soutiens morals tout au long de ce parcours académique.

Nous exprimons notre profonde gratitude à Dr. ASBAI Nassim, MCA en télécommunications, pour son encadrement et ses conseils scientifiques et méthodologiques.

Nous remercions également Mme HadjerBounazou pour son encadrement et son soutien, ainsi que les membres du jury qui ont accepté d'examiner et d'évaluer notre travail.

Enfin, nous remercions nos familles, nos amis et tous ceux qui nous ont soutenu de près ou de loin, lors du développement de cet travail.

Résumé

Ce travail est le résultat d'une simulation et évaluation d'un système de segmentation, la segmentation et le regroupement des locuteurs reposent sur l'utilisation du critère d'information bayésien (BIC). Ce processus consiste à découper l'enregistrement audio en segments de parole et à les regrouper en fonction des caractéristiques identifiées par le BIC.

Nous avons évalué sa précision sur des enregistrements audio courts en réalisant une étude comparative dans deux scénarios : le premier sans interférence entre locuteurs et le second avec interférence.

Pour cette évaluation, nous avons utilisé plusieurs métriques d'évaluation. Pour ces métriques on a trouves les résultats, notamment la pureté (92.47%), la couverture (92.47%) , la fausse alarme (FA) (0%), la détection manquée (MD) (5.26%) et le taux d'erreur de segmentation (DER) est (7.91%).

Ces résultats mettent en évidence l'importance de prendre en compte les métriques d'évaluation appropriées pour évaluer les performances des systèmes de segmentation de locuteurs. Ils soulignent également la nécessité de développer des approches plus robustes pour gérer les situations d'interférence, afin d'améliorer les performances et l'applicabilité de ces systèmes dans des conditions réelles.

Abstract

This work is the result of a simulation and evaluation of a segmentation system, speaker segmentation and grouping is based on the use of Bayesian Information Criterion (BIC). This process consists of slicing the audio recording into speech segments and grouping them according to the features identified by the BIC.

We evaluated its accuracy on short audio recordings by carrying out a comparative study in two scenarios: the first with no interference between speakers, and the second with interference.

For this evaluation, we used several evaluation metrics. Results for these metrics include purity (92.47%), coverage (92.47%), false alarm (FA) (0%), missed detection (MD) (5.26%) and segmentation error rate (DER) (7.91%).

These results highlight the importance of considering appropriate evaluation metrics when assessing the performance of speaker segmentation systems. They also highlight the need to develop more robust approaches to handling interference situations, in order to improve the performance and applicability of these systems under real-life conditions.

ملخص :

هذا العمل هو نتيجة محاكاة وتقييم نظام التجزئة , يستند تجزئة المتحدث وتجميعه إلى استخدام معيار المعلومات البايزي . تتضمن هذه العملية تقسيم التسجيل الصوتي إلى أجزاء كلام وتجميعها وفقًا للخصائص التي حددها معيار المعلومات البايزي .

قمنا بتقييم دقته على التسجيلات الصوتية القصيرة من خلال إجراء دراسة مقارنة في سيناريوهين: الأول دون تدخل بين المتحدثين والثاني مع التداخل .

في هذا التقييم، استخدمنا العديد من مقاييس التقييم. بالنسبة لهذه المقاييس، وجدنا النتائج، بما في ذلك النقاء (92.47٪)، والتغطية (92.47٪)، والإنذار الكاذب (0٪)، والكشف الفائت (5.26٪) ومعدل خطأ التجزئة هو (7.91٪).

وتبرز هذه النتائج أهمية مراعاة مقاييس التقييم المناسبة لتقييم أداء نظم تجزئة المتكلمين. كما أنها تسلط الضوء على الحاجة إلى تطوير نهج أكثر قوة لإدارة حالات التداخل، من أجل تحسين أداء هذه النظم وقابليتها للتطبيق في ظل ظروف العالم الحقيقي .

Table des matières

Listes des figures

Liste des tableaux

Abréviations

DERDiarizationerror rate

DAV/VADDétection d'activité vocale /voiceactivitydetection

DFTDiscret Fourrier transform

FFTFast Fourrier transform

DCTDiscret Cosinus transform

SRL Segmentation et regroupement en locuteurs

LPCCLinear Prédictioncepstral coefficients

LFCC Linear Frequency cepstral coefficients

MFCC Mel Frequency cepstral coefficients

HMM Hidden Morkov Models

VQVVoice Quality Value

SVMSupport vector machine

BICBayesian information criterion

BDBase de données

NIST National institue of standards and technology

FA False alarm

MD Missed detection

SPK ERROR Speaker error

Introduction générale

Introduction générale

Les sons que nous entendons sont le résultat de vibrations se propageant sous forme d'ondes. Parmi les paramètres acoustiques étudiés pour analyser ces ondes sonores, la fréquence joue un rôle essentiel et est mesurée en Hertz (Hz). Cependant, il est important de souligner que la fréquence de la voix humaine présente des variations en fonction de différents facteurs tels que l'individu, l'âge et le sexe.[1]

Depuis les années cinquante, la recherche sur la perception des sons de la parole a connu un développement intensif. De nombreux scientifiques se sont penchés sur cette question complexe, cherchant à comprendre comment les êtres humains perçoivent et interprètent les sons de la parole. Ces travaux ont permis de faire d'importants progrès dans notre connaissance des propriétés acoustiques de la parole.[2]

Dans ce cadre, La segmentation en locuteurs vise à obtenir des segments homogènes où chaque segment ne contient que la parole d'un seul locuteur, et ce, aussi longtemps que possible. Son objectif principal est de diviser un flux audio en portions distinctes pour identifier et isoler les paroles spécifiques de chaque locuteur.[3]

L'objectif de cette segmentation est de créer des segments cohérents et continus, afin de faciliter l'analyse ultérieure des données vocales. En éliminant le chevauchement entre les locuteurs, ainsi que les périodes de silence, de bruit ou de musique, la segmentation en locuteurs permet d'obtenir des segments clairs et distincts pour chaque locuteur.

Le but de cette étude est de pouvoir détecter le changement de locuteur et de corriger le groupement des syllabes lorsque les durées d'intervention sont courtes et également d'étudier l'effet des interventions qui se chevauchent sur les performances de notre système de segmentation.

Dans le cadre de notre travail, nous allons faire une Simulation et évaluation les performances d'un système de segmentation en locuteurs en utilisant deux types d'enregistrements vocaux, le premier est le cas parfait ou il y a aucune interférence entre les locuteurs(sans chevauchement) , tandis que le deuxième est le cas réel ou il y a un chevauchement entre locuteurs.

Chapitre 1

Notions Fondamentales et Traitement Numérique de la Parole

1.1. Introduction

La parole est un moyen de communication exclusif à l'être humain, qui a développé cette capacité unique pour transmettre des informations complexes et des émotions. Ce moyen de communication structuré est unique à l'espèce humaine il est le résultat d'une variation de la pression produite par l'émission d'un son par un locuteur [4]. Dans ce chapitre, Nous définissons brièvement certaines caractéristiques de ce signal ainsi que les différentes approches de traitement numérique utilisées pour l'analyser.

1.2. La parole

La parole se compose de séquences sonores et de silences qui servent à transmettre la pensée à travers un système de sons articulés [5].

La parole est un signal continu, aléatoire, redondant, non stationnaire et variable dans le temps, ce qui explique sa complexité.

1.2.1. Caractéristiques du Signal de Parole

Le signal de parole a des caractéristiques qui rendent son interprétation très difficile tels que :

- **Redondance et la Stationnarité**

Le signal de parole est redondant et non stationnaire mais il est peut-être considéré comme localement stationnaire. L'analyse du signal de parole se fait pendant ces périodes stationnaires dont la durée varie de 10 à 30ms. Cette durée correspond aussi la durée de stabilité de modèle de production [6].

- **la variabilité du signal**

La parole est caractérisée par une grande variabilité qui est influencée par divers facteurs, que ce soit pour un même locuteur ou pour plusieurs. Ces facteurs incluent notamment les perturbations liées au microphone (type, distance et orientation) ainsi que l'environnement acoustique (bruit, réverbération) [7].

- **la Continuité**

Il est continu dans le temps,ce qui requiert une discrétisation préliminaire du signal.

1.3. Prétraitement acoustique

Avant tout traitement de signal, Il faut effectuer certaines étapes (échantillonnage, préaccentuation et fenêtrage) pour réduire certains phénomènes qui apparaissent dans le signal vocal.

1.3.1 Echantillonnage

L'échantillonnage est une technique de traitement du signal permettant de numériser un signal analogique continu en un signal numérique discret. Cette technique consiste à prélever des échantillons à des intervalles réguliers de temps à partir du signal analogique. La fréquence d'échantillonnage doit être au moins deux fois plus élevée la fréquence maximale du signal, pour éviter la perte d'information due à l'échantillonnage, c'est le théorème de Shannon **[8]**.

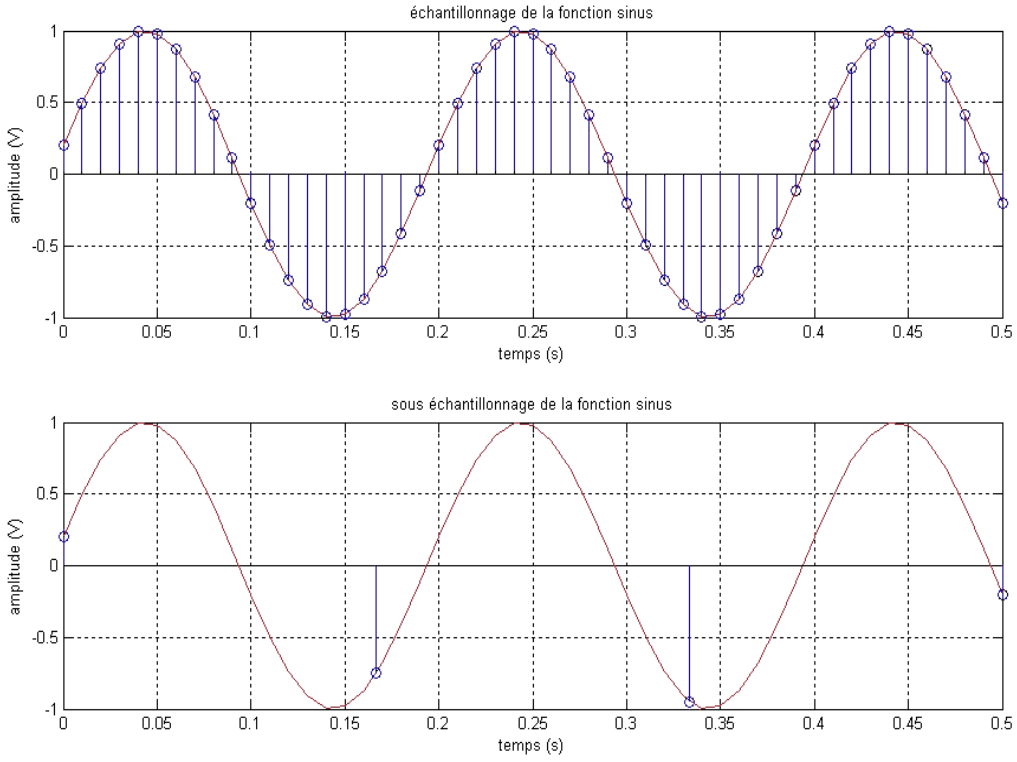


Figure 1.1 Echantillonnage

1.3.2. Préaccentuation et fenêtrage

1.3.2.1. Préaccentuation

L'amplitude des variations du signal à haute fréquence est moins importante que celle du signal à basse fréquence [8]. Les hautes fréquences véhiculent des informations tout aussi significatives que les basses fréquences, et l'amplitude est utile pour analyser le signal parlé.La préaccentuation est un exemple d'utilisation de connaissances sur la perception humaines elle consiste en un filtrage du signal de parole par le filtre :

$X(z) = 1 - \alpha z^{-1}$Avec $0.95 \leq \alpha \leq 1$(1.1)

1.3.2.2. Fenêtrage [8]

Cette technique est utilisée pour limiter l'effet des discontinuités du signal de parole. Une analyse à court terme montre que le signal vocal est quasi stationnaire sur des tranches temporelles de durées de 10 à 30 ms, cette analyse est effectuée à l'aide de fenêtres. Le choix de la fenêtre est très important parmi les fenêtres utilisées.On peut citer les fenêtres de hamming, hanning, blackman et de kaiser.

L'opération de fenêtrage consiste à multiplier le signal $x(n)$par un autre signal $h(k)$possédant N échantillons unités.

$s(n) = \sum_{k=1}^{N} x(n).h(k)$(1.2)

Avec :

$s(n)$: signal résultant

$x(n)$: signal à fragmenter

$h(k)$: fenêtre de Hamming, $k = 1, \ldots\ldots, N$

$h(k) = \alpha + (1 - \alpha)\cos(\frac{2\pi k}{N})$(1.3)

$$\alpha = 0.54$$

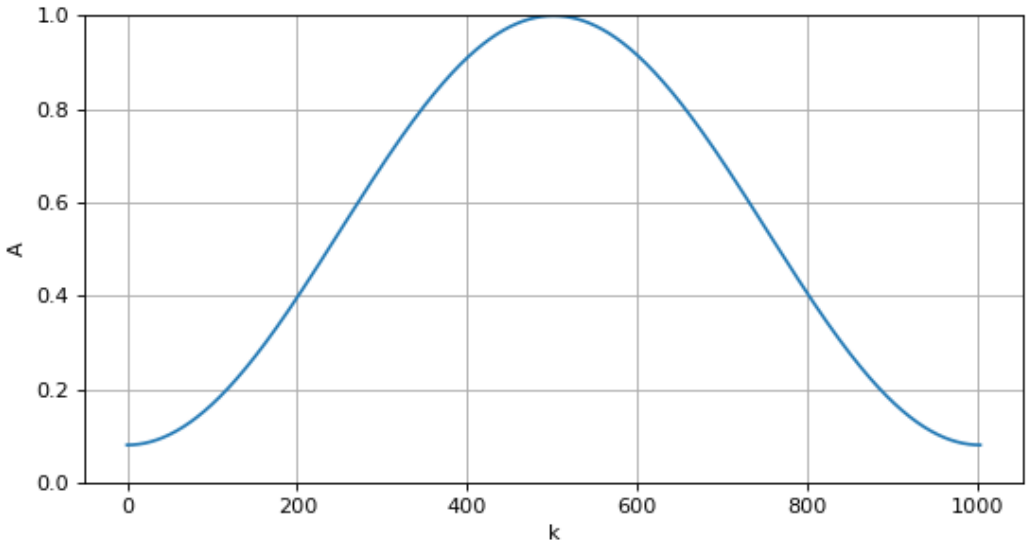


Figure 1.1 Fenêtre de Hamming au domaine temporel

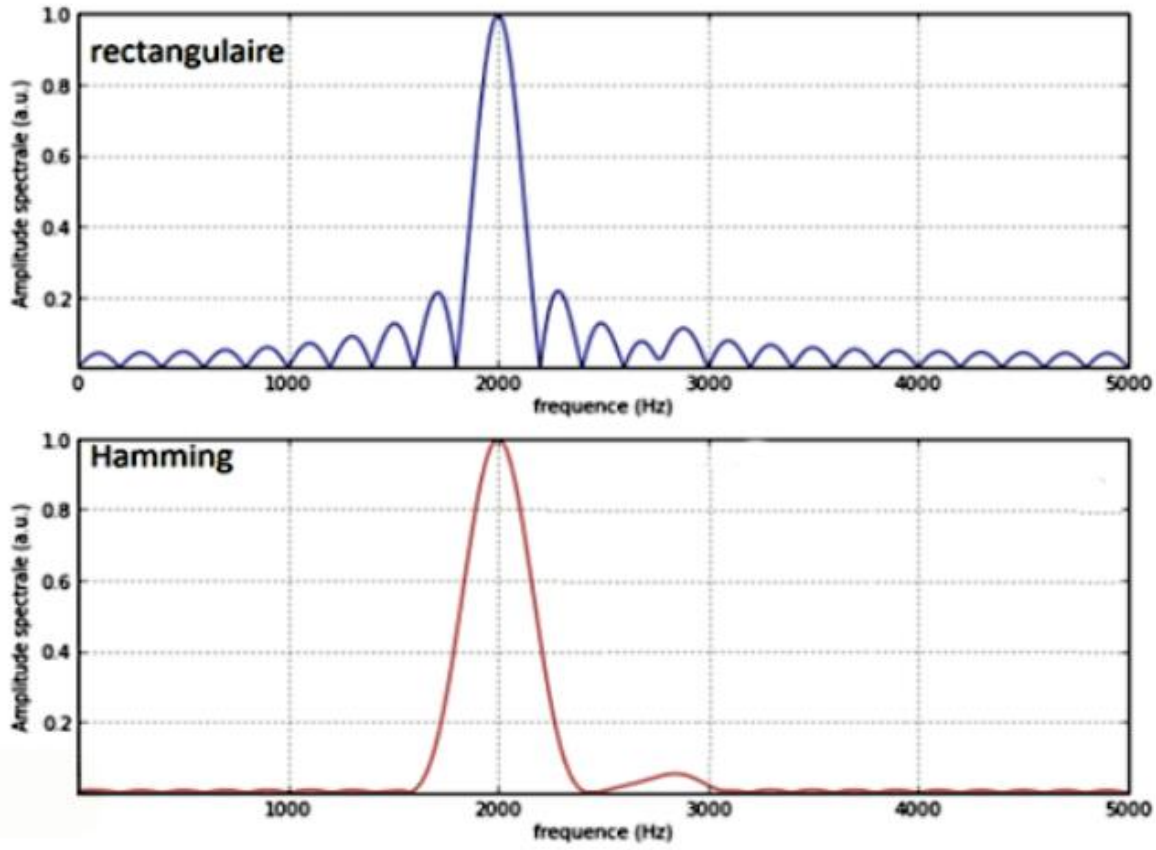


Figure 1.2 Un signal pondéré par deux fenêtres différentes (rectangulaire et Hamming), au domaine fréquentiel

1.4. Méthodes de traitement [9]

1.4.1. Analyse temporelle

Effectivement, l'analyse temporelle consiste à étudier le signal audio dans le domaine temporel, afin de mieux comprendre l'évolution du signal au fil du temps.

En utilisant ces techniques d'analyse temporelle, on peut mieux comprendre les événements qui se produisent dans le signal audio, et ainsi les isoler pour les étudier plus en détail. Cela peut être utile dans de nombreuses applications, comme la reconnaissance de la parole, la détection de défauts dans des machines, ou encore l'analyse de la qualité sonore des enregistrements audio. Parmi les principales techniques de l'analyse temporelle, on retrouve l'énergie et le taux de passage par zéro.

1.4.1.1.L'énergie

L'amplitude sonore est un paramètre acoustique important dans le traitement du signal audio car elle est liée à l'intensité du son et peut-être utilisée pour distinguer les sons voisés des sons non voisés, ainsi que pour séparer le signal utile des signaux gênants indésirables.

L'énergie est généralement plus forte pour les sons voisés que pour les sons non voisés, sa formule est donnée par :

$E = \sum_{n=1}^{N} s^2(n)$ (1.4)

$s(n)$: Lenième échantillon de la trame considérée.

N : Nombre d'échantillons de la fenêtre considéré.

1.4.1.2.Taux de passage par zéro

Le taux de passage par zéro (TPZ) est une mesure utilisée en traitement du signal pour caractériser le contenu spectral d'un signal de parole. Le TPZ peut être calculé à partir de la forme d'onde du signal de parole en comptant le nombre de passages par zéro pendant une période de temps donnée et en le divisant par la durée totale du signal.[10]

Le taux de passage par zéro est défini par l'expression suivante :

$TPZ = \frac{i.100}{N}\%$ (1.5)

i : Le nombre de passage par zéro

N : La taille de la fenêtre d'analyse

Le taux de passage par zéro des sons non voisés est supérieur à celui des sons voisés.

1.4.1.3.Détection d'activité vocale (DAV) [11]

Pour la reconnaissance(ou segmentation) du locuteur, il est important d'utiliser des paramètres acoustiques qui correspondent aux zones de parole plutôt qu'aux zones de silence ou de bruit. Ainsi, un algorithme de détection d'activité vocale (VAD; Voice ActivityDetection) est utilisé. Le VAD est un algorithme qui permet de distinguer les parties du signal vocal où il y a présence de parole de celles où il n'y en a pas. Cela est rendu possible grâce à un détecteur d'activité vocale. Le VAD est un élément clé dans diverses applications telles que la transmission de la parole (pour éviter le codage inutile ou la transmission de silence), la réduction de bruit dans un signal de parole et la reconnaissance de la parole ou du locuteur.

1.4.2. Analyse fréquentielle [12] [13] [14]

L'analyse spectrale est une méthode courante de traitement du signal qui consiste à décomposer une grandeur variant en fonction du temps en ses composantes fréquentielles. Elle permet de caractériser la répartition d'énergie ou de puissance d'un signal en fonction de la fréquence, ce qui est particulièrement utile pour rendre le signal de parole plus fidèle et proche du fonctionnement de l'oreille humaine. Les outils d'analyse fréquentielle les plus couramment utilisés incluent la transformation de Fourier et la transformation en cosinus discrète.

1.4.2.1 Transformée de Fourier discrète (DFT)

L'analyse fréquentielle du signal audio est basée sur les opérations de la transformée de Fourrier définie par l'équation suivante :

$S(f) = \int_{n=-\infty}^{+\infty} s(t).e^{-j2\pi ft}\, dt$(1.6)

Formule de série de Fourier :

Soit $s(x)$ une fonction définie sur l'intervalle $]-L, L[$, et $T = 2L$

$$S(f) = \frac{a_0}{2} + \sum_1^\infty (a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right) \quad (1.7)$$

Ou les coefficients a_n et $b_n \forall\, n \in N$ sont appelés coefficients de Fourier et valent :

$$a_n = \frac{1}{L}\int_{-L}^{L} s(x)\, cos\, \frac{n\pi x}{L}\, dx \quad (1.8)$$

$$b_n = \frac{1}{L}\int_{-L}^{L} s(x)\, sin\, \frac{n\pi x}{L}\, dx \quad (1.9)$$

La transformée de Fourier discrète (TFD) est une méthode couramment utilisée pour effectuer la transformation de Fourier d'un signal discret. Elle permet de représenter un signal discret périodique en termes de sa composante fréquentielle discrète correspondante. La TFD est largement utilisée en traitement de signal et en analyse spectrale.

On appelle Transformée de Fourier Discrète (TFD ou Discret Fourier Transform DFT) de N termes, la suite définie par :

$$S(k) = \sum_{n=0}^{N-1} s(n) e^{-j2\pi n \frac{k}{N}} \quad (1.10)$$

Avec une fréquence :

$$f = \frac{k}{N} \quad (1.11)$$

N : nombre d'échantillons

k, n : entre 0 et $N - 1$

1.4.2.2. Transformée en Cosinus Discrète (DCT) [15]

La Transformée en Cosinus Discrète (DCT) est une technique de traitement de signal qui permet de représenter un signal sous forme de coefficients de cosinus. Elle est utilisée notamment dans la compression d'images et de vidéos. La DCT est très similaire à la Transformée de Fourier Discrète (DFT), mais elle est plus adaptée aux signaux périodiques et symétriques.

la DCT s'exprime par :

$$C(u,v) = \frac{2}{N}\alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y).\cos\left(\frac{\pi(2x+1)u}{2N}\right).cos\left(\frac{\pi(2y+1)v}{2N}\right) \quad (1.12)$$

$$u,v : 0,1,\dots,N-1$$

$c()$ est définie par : $c(u) = \begin{cases} \frac{1}{\sqrt{2}} \; si \; u = 0 \\ 1 \quad si \; u \neq 0 \end{cases}$ (1.13)

1.5. Conclusion

Ce chapitre offre un aperçu approfondi des principes clés et des techniques utilisées dans le domaine du traitement de la parole. Nous avons exploré les différentes étapes du traitement de la parole, et quelques étapes de l'extraction des paramètres qui vont nous aider dans l'étude du deuxième chapitre

Chapitre 2

Architecture générale d'un système de segmentation en locuteurs

2.1. Introduction

Ce chapitre aborde en détail la tâche de segmentation et de regroupement automatique des locuteurs. Après avoir exposé le principe, l'importance et les défis de cette tâche, nous examinons les différentes composantes d'un système de segmentation et de regroupement en locuteurs. Enfin, nous présentons le système de référence qui sera utilisé dans nos expériences, offrant ainsi une méthodologie solide pour mener à bien cette tâche complexe.

La tâche de segmentation et regroupement en locuteurs (SRL), également connue sous le nom de diarisation des locuteurs, consiste à déterminer le nombre de locuteurs et leurs interventions dans un enregistrement audio. Cette tâche repose principalement sur la comparaison des voix des locuteurs. Pour ce faire, il est nécessaire d'obtenir des segments de parole homogènes qui ne contiennent que la voix d'un seul locuteur. Un segment de parole est une partie du signal audio qui contient de la parole et est délimitée par des frontières arbitraires. Il diffère d'un tour de parole, qui est une partie du signal audio contenant les paroles d'un seul locuteur et dont les frontières indiquent un changement de locuteur ou de type de données (musique, silence, etc.). Il peut arriver que plusieurs personnes parlent en même temps, créant ainsi une superposition de paroles. La tâche de SRL consiste donc à découper l'enregistrement audio en segments de parole et à les regrouper par locuteur.[33]

Traditionnellement, le processus de segmentation et de regroupement en locuteurs (SRL) est appliqué de manière individuelle à chaque enregistrement audio d'un corpus, sans utiliser de connaissances a priori sur les locuteurs. Le nombre de locuteurs ainsi que leur identité sont inconnus, et aucun modèle ou échantillon de leur voix n'est disponible. La plupart des systèmes de SRL proposés jusqu'à récemment suivent cette approche, où les émissions sont traitées et évaluées de manière individuelle (SRL d'émissions, single-show ,diarization).[34]

La segmentation en locuteurs a été récemment étudiée dans la littérature en tant qu'étape préliminaire à plusieurs tâches, telles que la transcription automatique de journaux télévisés, le regroupement automatique de messages et la poursuite de locuteurs.[32] Les premiers travaux sur la segmentation en locuteurs remontent aux années 1990, réalisés par Herbert Gish dans le but d'améliorer le trafic aérien dans un aéroport. Dans son article ,il a proposé une architecture qui a été ensuite adoptée par la plupart des systèmes de segmentation .Cette architecture se compose de plusieurs étapes.[35]

- Paramétrisation
- Pré-segmentation acoustique
- Détection de changements de locuteur
- Regroupement des segments.

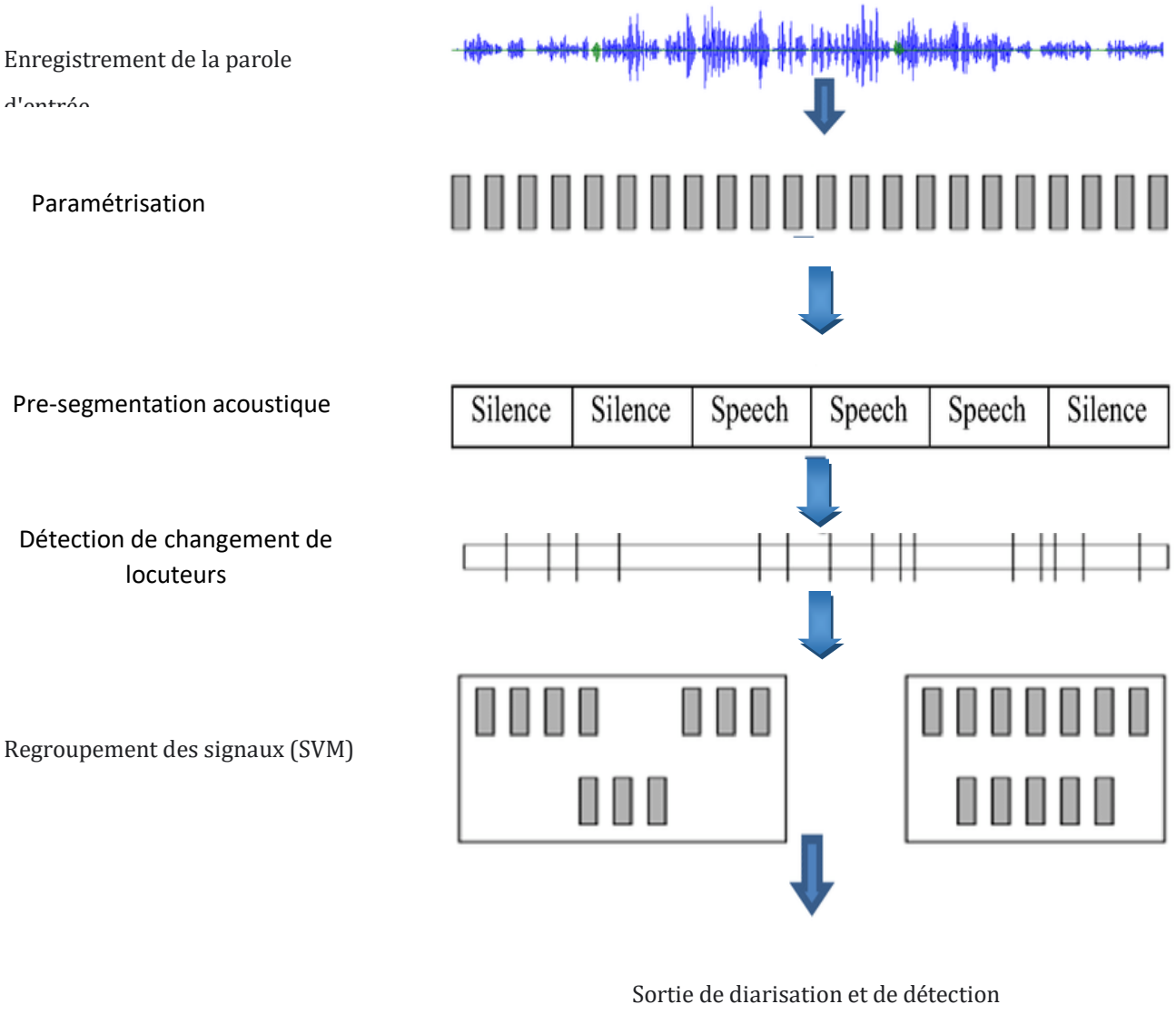


Figure 2.1 Architecture d'un système de segmentation

2.2. Paramétrisation

Les systèmes de segmentation et de transcription de l'audio requièrent un traitement préalable du signal acoustique en raison de sa complexité et de sa redondance. En effet, le signal vocal est caractérisé par une grande diversité d'informations, ce qui rend nécessaire une étape de prétraitement avant toute tentative de traitement ultérieur. Ainsi, le signal vocal est converti en une séquence de vecteurs de paramètres. Cette conversion permet de représenter de manière plus compacte et informative le contenu acoustique de l'audio, facilitant ainsi son analyse et son traitement par la suite. [16]

La paramétrisation acoustique du signal de la parole commence par la fragmentation du signal en séquences de taille fixe, généralement entre 10 ms et 30 ms. Ces séquences, appelées trames, sont réparties de manière uniforme le long du signal. Cette approche permet de considérer le signal de la parole comme pseudo stationnaire sur de très courts intervalles de temps, malgré son caractère continu et non stationnaire. [24],[35].

Différents types de paramètres peuvent être extraits du signal audio de la parole, et ils peuvent être classés en fonction de leur domaine d'origine : temporel, fréquentiel ou combinant à la fois le domaine temporel et fréquentiel. [16].

La segmentation en locuteurs repose sur l'utilisation de paramètres cepstraux, qui sont des descripteurs couramment utilisés en traitement de la parole. Ces paramètres permettent de séparer efficacement l'excitation glottique de la résonance induite par l'appareil vocal humain. Parmi les différents types de coefficients cepstraux, les coefficients MFCC (Mel FrequencyCepstral Coefficients) et les coefficients LPCC (LinearPredictionCepstral Coefficients) sont les plus répandus. [19].

2.2.1. Coefficients MFCC [8],[17] [18]

Les coefficients MFCC (Mel FrequencyCepstral Coefficients) sont calculés en utilisant l'échelle de fréquence perceptuelle Mel. Cette échelle a été conçue pour représenter la perception des fréquences sonores par l'ouïe humaine. Contrairement à l'échelle de fréquence linéaire, l'échelle de fréquence Mel est linéaire jusqu'à environ 1000 Hz, puis devient logarithmique au-delà de cette fréquence.

L'échelle Mel est divisée en un certain nombre de filtres triangulaires, uniformément répartis le long de l'échelle. Chaque filtre a une largeur d'environ 300 Mels, et la distance entre deux

filtres consécutifs est d'environ 150 Mels. Ces filtres sont utilisés pour analyser le signal de parole et capturer les caractéristiques phonétiques importantes.

L'utilisation de ces filtres présente plusieurs avantages. Tout d'abord, ils permettent de mieux représenter la perception humaine des fréquences, en accordant plus d'importance aux fréquences qui sont plus discriminantes pour la parole. De plus, les filtres Mel offrent une meilleure résolution temporelle pour les hautes fréquences, car la bande passante de chaque filtre est proportionnelle à la fréquence centrale du filtre. Cela signifie qu'il y a une plus grande densité de filtres dans les hautes fréquences, ce qui permet de mieux capturer les détails temporels dans ces gammes de fréquences.

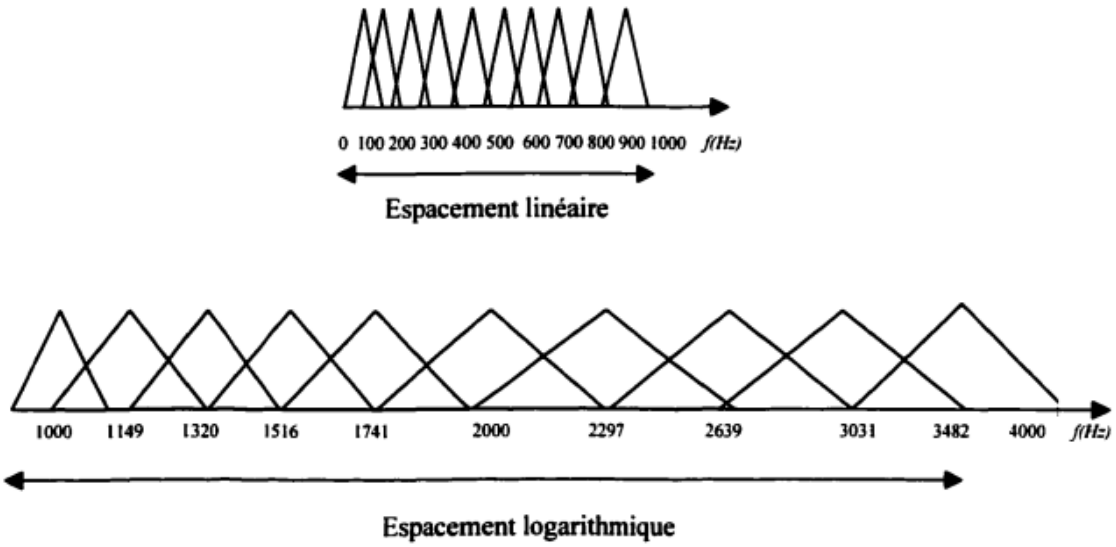


Figure 2.2 Banc de filtres à l'échelle Mel

La correspondance entre la fréquence en Hz et la fréquence Mel est définis par :

$$F_{Mel} = 2595.log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \quad (2.1)$$

Voici les principales étapes de l'extraction des coefficients MFCC reformulées :

1. Le signal vocal est pré-accentué pour réduire les effets des discontinuités du signal aux limites des trames. Ensuite, le signal est découpé en trames en utilisant une fenêtre de Hamming, avec un chevauchement de 50% entre les trames.

2. Pour chaque trame fenêtrée, une transformation de Fourier rapide (FFT) est appliquée. Ensuite, le module carré de la FFT est calculé, car il contient des informations pertinentes pour la segmentation en locuteurs.

3. Les trames sont ensuite filtrées à l'aide d'un banc de filtres triangulaires basés sur l'échelle de fréquence Mel. Cette étape permet de lisser le spectre et de réduire les informations à traiter.

4. Un logarithme est appliqué au spectre de chaque trame après le filtrage. Cela permet de réduire la dimension des coefficients.

5. Enfin, au lieu d'effectuer une transformation de Fourier inverse (FFT inverse), une transformée en cosinus discrète inverse (DCT inverse) est utilisée. La DCT inverse est préférée car elle décoré le mieux les coefficients que la FFT inverse. Cette étape permet d'obtenir les coefficients MFCC en résultat.

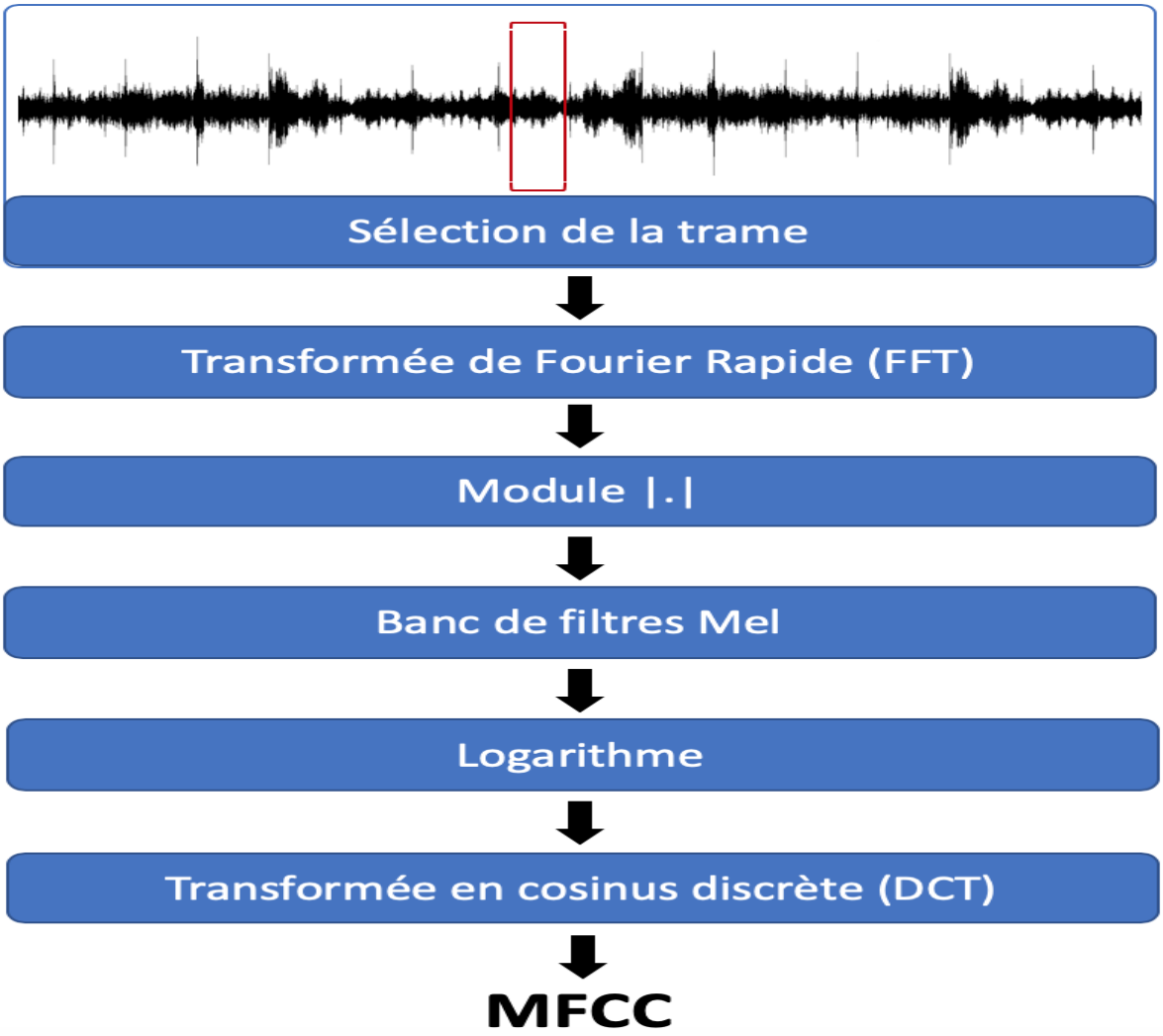


Figure 2.3 Les différentes étapes d'extractions des paramètres MFCC

Les coefficients MFCC sont calculés comme suivant :

$c_i = \sum_{l=1}^{k} \log(S_k) cos\left[n(l - \frac{1}{2})\frac{\pi}{k}\right]$ Pour $n = 1,2, \ldots . L(2.2)$

Ou L est le nombre de coefficients MFCC que l'on souhaite obtenir

Avec S_k : l'énergie a la sortie du filtre

k : le nombre de filtres

2.2.2. Coefficients différentiels et énergie [20][21]

Les coefficients différentiels et l'énergie sont souvent utilisés en complément des coefficients MFCC pour améliorer la représentation des caractéristiques acoustiques des signaux de parole. Les coefficients différentiels représentent la variation temporelle des coefficients MFCC et fournissent des informations sur la vitesse de variation des caractéristiques acoustiques, tandis que l'énergie représente la puissance acoustique totale du signal sur une période donnée.

En utilisant les coefficients différentiels et l'énergie en plus des coefficients MFCC, on peut améliorer la précision de la reconnaissance de la parole, ou de locuteur ou de segmentation en locuteurs, en particulier dans des environnements bruités ou avec une variabilité importante des locuteurs. Cependant, cela augmente également la complexité du modèle et peut nécessiter une plus grande quantité de données pour l'entraînement.

En revanche, l'énergie est fréquemment utilisée dans la segmentation car c'est la quantité intuitive qui caractérise le signal et parce qu'elle est fréquemment examinée sur un certain nombre de trames de signal ultérieures, ce qui facilite l'identification des silences, la mise en évidence des fluctuations et la distinction de la parole musique.

2.2.3. Moyenne de coefficients cepstraux(normalisation cepstrale)[22]

La normalisation cepstrale est une technique utilisée pour réduire les parasites dans le signal acoustique, causés par de nombreux facteurs tels que les conditions physiques et psychologiques du locuteur, la qualité du matériel et les conditions d'enregistrement. Cette technique consiste à soustraire les moyennes de coefficients cepstraux du signal afin de les centrer et les réduire. Le signal acoustique de la parole est influencé par de nombreux facteurs tels que les conditions physiques et psychologiques du locuteur, la qualité du matériel et les conditions d'enregistrement.

2.3. Modélisation du locuteur

Les modèles de caractérisation du locuteur sont basés sur des modèles statistiques utilisés dans la reconnaissance de formes. Parmi ces modèles, les mixtures de gaussiennes GMM (Gaussian Mixture Models) sont les plus couramment utilisées en reconnaissance du locuteur. Ces modèles ont été introduits en 1992 par Reynolds [23]. Les modèles de Markov cachés (HMM), introduits en 1975, sont largement utilisés dans la reconnaissance de la parole. Les

techniques de modélisation de locuteurs utilisant les GMM et les HMM ont démontré leursefficacités pour la reconnaissance de locuteurs. En plus de ces méthodes, il existe d'autres techniques telles que la quantification vectorielle (VQ), les machines à vecteurs de support (SVM) et les réseaux de neurones artificiels.[24]

2.3.1. Mixtures de gaussiennes

Les séquences de vecteurs acoustiques pour un locuteur donné peuvent être modélisées par une somme pondérée de distributions gaussiennes multidimensionnelles dans le cas des GMMs.[23]

Les paramètres du modèle de mélange gaussien comprennent les vecteurs moyens, les matrices de covariance et les poids de mélange pour toutes les densités de composants. L'ensemble complet de ces paramètres est représenté par une notation collective :

$$P(x) = \frac{1}{\sqrt{2\pi^d|\sum|}} exp\left(-\frac{1}{2}(x-\mu)^T \sum_{i=1}^{-1}(x-\mu)\right) (2.3)$$

Où *d* est la dimension d'un vecteur de paramètres noté *x*, μ le vecteur moyen et Σ la matrice de covariance estimée à partir de données $X = \{x_1, x_2, \dots x_N\}$selon les formules suivantes :

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (2.4)$$

$$\sum = (X-\bar{X})(X-\bar{X})^T (2.5)$$

Le modèle GMM est défini par un ensemble de Tdistributions gaussiennes, chacune ayant un poids wi associé, représentant la probabilité d'observer chaque distribution. Ainsi, pour un vecteur acoustique $x_i$donné, la probabilité qu'il appartienne à l'un des T composants est représentée par $P(x_i)$.Le nombre T est appelé l'ordre du modèle et est souvent déterminé par validation croisée.

$$P(x) = \sum_{i=1}^{T} w_i P_i(x) (2.6)$$

$$\sum_{i=1}^{T} w_i = 1 (2.7)$$

Les GMM sont largement utilisés comme modèles paramétriques de la distribution de probabilité des mesures ou des caractéristiques continues dans les systèmes biométriques, en raison de leur capacité à représenter une grande variété de distributions d'échantillons. Toutefois, Une étape essentielle pour générer des modèles de locuteurs robustes consiste à

déterminer les valeurs optimales de ces paramètres. Parmi les méthodes d'estimation couramment utilisées, on trouve le calcul de la probabilité, les tests d'hypothèse de Bayes, etc.[25]

2.3.1.1.Calcul de probabilité [26]

Afin de déterminer si une séquence de vecteurs acoustiques $X = x_1, x_2, \dots x_N$correspond à un locuteur donné, il est nécessaire de calculer la probabilité que cette séquence ait été prononcée ou non par le locuteur en question. Pour cela, il est courant d'utiliser la vraisemblance de chaque observation$P(x_i \backslash \lambda)$, qui est calculée à l'aide de la formule suivante dans le cas des GMM :

$$P(x_i \backslash \lambda) = \sum_{j=1}^{T} w_j \frac{1}{\sqrt{(2\pi)^d |\sum_j|}} exp\left(-\frac{1}{2}(x_i - \mu_j)^T \sum_j^{-1}(x_i - \mu_j)\right) (2.8)$$

$$P(X \backslash \lambda) = \sqrt[N]{\prod_{i=1}^{N} P(x_i \backslash \lambda)} (2.9)$$

La vraisemblance de la séquence est obtenue en prenant la moyenne géométrique des vraisemblances des observations. Cette moyenne est nécessaire pour réduire l'impact de la durée de la séquence de test sur la vraisemblance$P(X \backslash \lambda)$.Cette méthode repose sur l'hypothèse que les vecteurs acoustiques sont indépendants les uns des autres.

2.4. Segmentation de locuteurs

En matière diarisation automatique par locuteurs, la segmentation vise à repérer les moments où il y a un changement de locuteur, c'est-à-dire les points de rupture.**[27]**

L'objectif de la segmentation dans le domaine de diarisation automatique par locuteurs est de diviser le flux audio en segments homogènes, où chaque segment contient uniquement la parole d'un seul locuteur. De plus, la segmentation doit garantir que les segments adjacents appartiennent toujours à des locuteurs différents.

2.4.1. la Pré-segmentation acoustique

La pré-segmentation acoustique est une étape essentielle dans les systèmes de segmentation en locuteurs. Elle consiste à découper le signal audio en segments acoustiques distincts, qui peuvent correspondre à des tours de parole ou à des zones de silence, de musique ou de bruit.

L'objectif de la pré-segmentation acoustique est de fournir des unités de base pour l'analyse ultérieure du signal et faciliter l'identification des changements de locuteurs.
Différentes techniques sont utilisées pour la pré-segmentation acoustique. Parmi elles, on trouve la détection de pauses, qui repose sur la détection des périodes de silence entre les tours de parole, et la détection d'énergie, qui se base sur l'analyse de l'énergie du signal pour identifier les zones de parole et de silence. D'autres approches plus avancées peuvent également être utilisées, telles que l'utilisation de modèles acoustiques ou la détection de changements de propriétés spectrales.

2.4.2. Détection du changement de locuteurs [28],[29]

La deuxième étaped'un système de segmentation consiste en la détection de changement de locuteurs, après la modélisation de locuteurs. Elle est réalisée sur chaque segment de parole afin de déterminer les moments où il y a un changement de locuteur. L'objectif est de repérer les points de changements entre les différents locuteurs et d'obtenir des segments ne contenant de la parole que d'un seul locuteur idéalement.

Lorsque le flux audio n'est pas segmenté, une détection acoustique plus générale doit être effectuée afin de détecter le changement de locuteurs, ainsi que les différentes classes de parole telles que le silence, la musique, le bruit. En revanche, si le flux audio est déjà segmenté, on ne détecte que les changements de locuteurs.

Il n'est pas nécessaire d'étiqueter (annoter) les segments en fonction des locuteurs lors de l'étape de détection de changements de locuteurs, car une étape de regroupement ultérieure permettra de réunir les segments appartenant au même locuteur.

La détection des changements de locuteurs est une tâche difficile car la durée de certains tours de locuteurs peut être très courte. Plusieurs techniques ont été proposées dans la littérature, regroupées en trois catégories :

- Détection de changements de locuteur par détection de silences.
- Détection de changements de locuteur par utilisation d'une distance.
- Détection de changements de locuteur par identification de la nature des segments.

Dans notre travail, nous nous intéressons à la détection de changement de locuteurs par l'utilisation d'une distance.

2.4.2.1. La détection de changement de locuteurs par l'utilisation d'une distance

La détection de changement de locuteurs par l'utilisation d'une distance est une méthode qui repose sur l'utilisation de mesures de distances pour détecter les changements de locuteurs dans un flux audio. Cette méthode peut être appliquée à différentes caractéristiques audio, telles que la fréquence fondamentale, le spectre, ou encore les coefficients cepstraux.[30]

En effet, la méthode de détection de changement de locuteurs par l'utilisation d'une distance repose sur le calcul d'une mesure de distance entre deux vecteurs de caractéristiques audio successifs. Les vecteurs de caractéristiques audio sont extraits à partir de segments de parole contenant un seul locuteur. Lorsque la distance entre deux vecteurs est supérieure à un seuil prédéfini, un changement de locuteur est détecté.

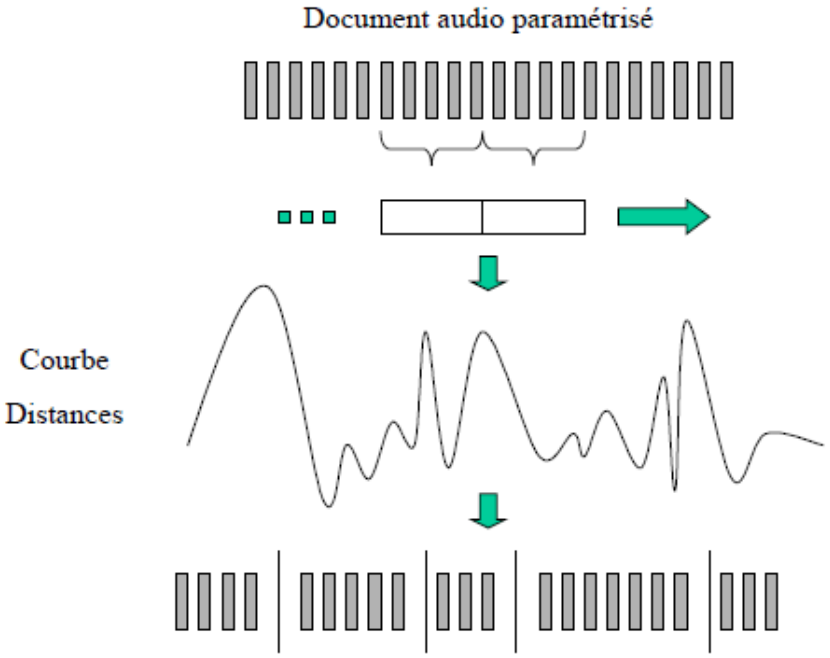


Figure 2.4 Utilisation de deux fenêtres adjacentes pour le calcul d'une distance

2.5. Regroupement des segments

Après avoir obtenu des segments contenant uniquement les paroles d'un locuteur unique, la dernière étape d'un système de segmentation de locuteurs est le regroupement des segments, qui revêt une grande importance dans certains scénarios. L'objectif de cette étape est de regrouper tous les segments appartenant au même locuteur dans un cluster ou une classe. Chaque cluster doit correspondre à un locuteur spécifique et ne contenir que des segments provenant de ce locuteur.

Le regroupement des segments est un problème de classification non supervisée, où nous devons regrouper une collection de segments en classes, sans disposer d'informations a priori sur la nature des classes ou leur nombre.

L'étape de regroupement n'est pas limitée à la segmentation de locuteurs et peut être utilisée dans d'autres applications, telles que l'extraction de messages téléphoniques déposés par une même personne sur une messagerie vocale.

La plupart des systèmes de regroupement de locuteurs présentés dans la littérature utilisent une approche hiérarchique. Il existe deux méthodes couramment utilisées pour le regroupement : l'approche ascendante (bottom-up clustering) par agglomération, et l'approche descendante (top-down clustering) par division.

L'approche ascendante consiste à commencer par considérer chaque segment comme un cluster individuel, puis fusionner progressivement les clusters similaires jusqu'à ce que tous les segments soient regroupés en un seul cluster global.

L'approche descendante, quant à elle, part d'un seul cluster global et le divise en sous-clusters plus petits à chaque étape, en utilisant des critères de similarité spécifiques.

Ces méthodes de regroupement sont adaptées pour traiter la segmentation de locuteurs, car elles permettent d'identifier les similarités entre les segments et de regrouper ceux provenant du même locuteur. Cependant, le choix de la méthode de regroupement dépend des spécificités de la tâche et des caractéristiques des données audio.

En résumé, l'étape de regroupement des segments dans un système de segmentation de locuteurs est cruciale pour rassembler les segments appartenant au même locuteur dans des clusters distincts. Les approches ascendante (buttom-up clustering)et descendante(top-down clustering) sont souvent utilisées dans la littérature pour accomplir cette tâche, offrant différentes perspectives sur la manière de regrouper les segments.

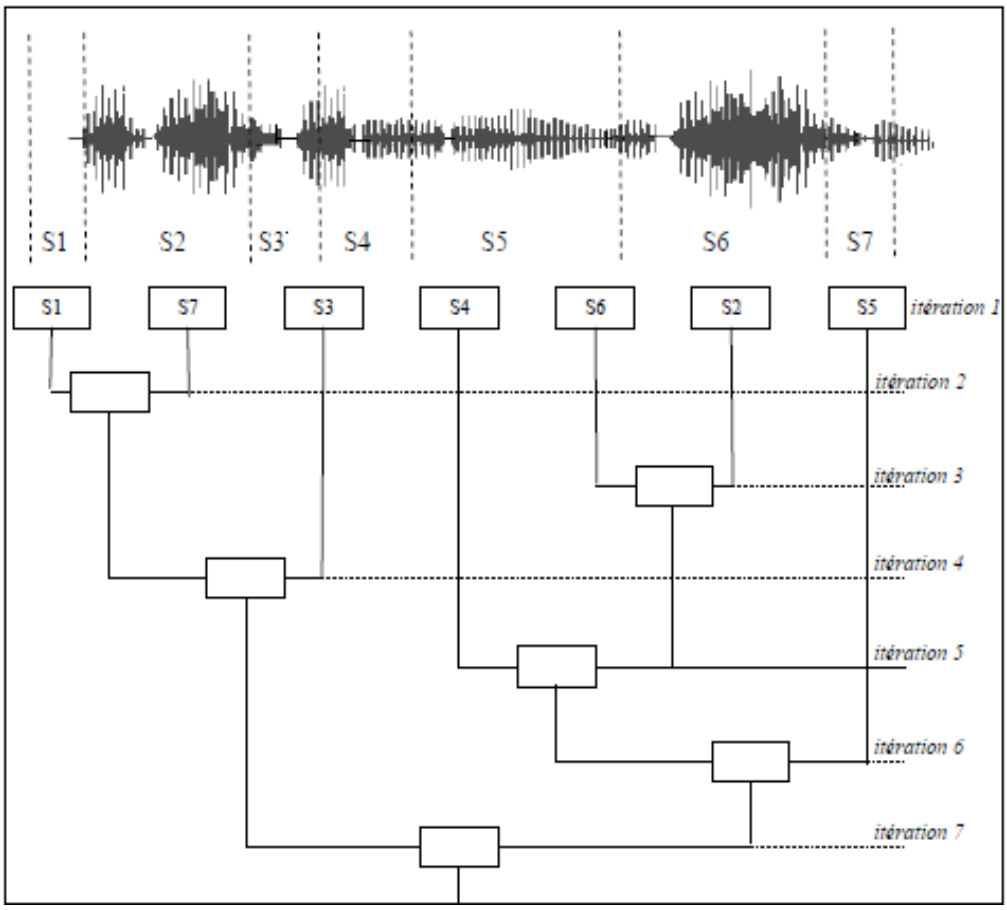


Figure 2.5 Exemple de regroupement par agglomération.

2.5.1. Regroupement hiérarchique agglomératif [31]

Le regroupement hiérarchique agglomératif est une méthode de classification non supervisée qui vise à regrouper les données en formant une hiérarchie de clusters. Dans cette approche, chaque donnée est initialement considérée comme un cluster individuel, puis les clusters les plus proches sont successivement fusionnés jusqu'à ce qu'un seul cluster global soit formé.

L'algorithme de regroupement hiérarchique agglomératif peut être résumé en quelques étapes clés :

1. Initialisation : Chaque donnée est assignée à son propre cluster.
2. Calcul de la similarité : Une mesure de similarité (par exemple, distance euclidienne) est utilisée pour calculer la similarité entre les clusters.

3. Fusion : Les deux clusters les plus similaires sont fusionnés pour former un nouveau cluster.
4. Mise à jour de la matrice de similarité : La matrice de similarité est mise à jour pour refléter la similarité entre les nouveaux clusters et les autres clusters.
5. Répétition : Les étapes 2 à 4 sont répétées jusqu'à ce qu'un seul cluster global soit obtenu.

Le regroupement hiérarchique agglomératif génère une structure de dendrogramme qui représente la hiérarchie des clusters. Ce dendrogramme peut être utilisé pour déterminer le nombre optimal de clusters en fonction de la structure des données.

Il convient de noter que le choix de la mesure de similarité et de la méthode de fusion des clusters peut avoir un impact significatif sur les résultats du regroupement hiérarchique agglomératif. Différentes mesures de similarité (par exemple, distance euclidienne, distance de Manhattan, corrélation) et différentes méthodes de fusion (par exemple, liaison simple, liaison complète, liaison moyenne) peuvent être utilisées en fonction des caractéristiques et des objectifs spécifiques de l'application.

Les mesures introduites dans le regroupement sont utilisées pour évaluer la similitude, par exemple :

- La distance BIC
- La distance de KullBach-Liebler
- Le rapport de vraisemblance croisé

La distance BIC (Bayesian Information Criterion) est connue pour produire les meilleurs résultats en termes de précision de regroupement dans la détection de changement de locuteurs. Cependant, cette distance est également la plus coûteuse en termes de temps de calcul, car elle nécessite l'estimation d'un modèle statistique pour chaque paire de locuteurs.

Le rapport de vraisemblance croisée est une mesure utilisée pour évaluer la similitude entre deux locuteurs (i, j). Il est calculé en évaluant la vraisemblance $P(X|\lambda)$ de l'ensemble des trames acoustiques X par rapport au modèle λ associé à chaque locuteur. Plus le rapport de vraisemblance croisée est élevé, plus les locuteurs i et j sont considérés comme similaires en termes de caractéristiques acoustiques.

Ces mesures sont couramment utilisées dans les systèmes de détection de changement de locuteurs pour évaluer la distance et la similitude entre les locuteurs. Cependant, il est important de noter que le choix de la mesure dépend des objectifs spécifiques de l'application et des ressources disponibles en termes de temps de calcul et de modélisation des locuteurs.[18]

La distance $d_{scl}(i,j)$ s'écrit :

$$d_{scl}(i,j) = \frac{1}{n_i} log \frac{P(X_i \backslash \lambda_{ubm})}{P(X_i \backslash \lambda_j)} + \frac{1}{n_j} log \frac{P(X_j \backslash \lambda_{ubm})}{P(X_j \backslash \lambda_i)} (2.10)$$

n_i, n_j:nombre de tramesacoustique

λ_{ubm}: Modèle du monde

La distance de KullBach-Liebler mesure la distance de probabilité entre deux locuteurs (i,j), elle est définis par l'équation suivante :

$$d_{kl}(i,j) = log \frac{P(X_i \backslash \lambda)}{P(X_i \backslash \lambda_j)} + \frac{P(X_j \backslash \lambda)}{P(X_j \backslash \lambda_i)} (2.11)$$

2.5.2. Critère d'arrêt du regroupement hiérarchique

Le critère d'arrêt du regroupement hiérarchique est une condition qui détermine quand arrêter le processus de regroupement et finaliser la segmentation en locuteurs. Il existe différents critères d'arrêt qui peuvent être utilisés, en voici quelques-uns :

1. Nombre fixe de classes : Le critère d'arrêt consiste à spécifier un nombre prédéfini de classes souhaité. Une fois que ce nombre est atteint, le regroupement s'arrête et les classes obtenues sont considérées comme les locuteurs identifiés.
2. Variation de la distance : Le critère d'arrêt est basé sur la variation de la distance entre les clusters lors des regroupements successifs. Si la variation de la distance entre les clusters est inférieure à un seuil prédéfini, cela indique que la similarité entre les segments n'augmente plus de manière significative, ce qui peut être interprété comme un arrêt du regroupement.
3. Taille minimale des clusters : Le critère d'arrêt consiste à définir une taille minimale pour les clusters. Si la taille d'un cluster devient inférieure à ce seuil, le regroupement s'arrête car il est considéré comme non significatif en termes de nombre d'échantillons ou de durée.

4. Critère de qualité du regroupement : Le critère d'arrêt est basé sur une mesure de qualité du regroupement, telle que l'indice de silhouette ou l'inertie intra-cluster. Lorsque la qualité du regroupement atteint un maximum ou dépasse un seuil spécifié, le regroupement s'arrête.

Ces critères d'arrêt peuvent être utilisés individuellement ou en combinaison pour déterminer quand arrêter le regroupement hiérarchique. Le choix du critère d'arrêt dépendra des caractéristiques du problème spécifique et des objectifs de la segmentation en locuteurs.

2.6. Conclusion

Dans ce chapitre, nous avons présenté les étapesd'un système de segmentation en locuteurs. Tout d'abord, nous avons utilisé la paramétrisation pour extraire les coefficients cepstraux MFCC, qui caractérisent la distribution spectrale d'énergie du signal acoustique. Ensuite, nous avons procédé à la modélisation des locuteurs à l'aide de modèles tels que les HMM (Hidden Markov Models) ou les GMM (Gaussian Mixture Models). Les GMM sont souvent privilégiés en raison de leur capacité à représenter différentes distributions d'échantillons.

Les deux dernières étapes consistent à détecter les changements de locuteurs et à regrouper les segments correspondants. Pour la détection de changement de locuteurs, nous avons utilisé la distance BIC (Bayesian Information Criterion), qui permet de détecter les changements proches les uns des autres. Cette distance nécessite l'estimation de modèles spécifiques pour chaque paire de locuteurs. Ensuite, nous avons appliqué une approche de regroupement hiérarchique en utilisant des critères tels que le rapport de vraisemblance croisé ou la distance BIC pour évaluer la similarité entre les locuteurs. Le critère d'arrêt nous permet de déterminer le nombre final de classes, garantissant ainsi une segmentation précise.

Chapitre 3

Simulation et évaluation d'un système de segmentation en locuteurs d'un document audio

3.1.Introduction

L'objectif de ce chapitre est d'analyser les résultats obtenus en utilisant des mesures d'évaluation telles que le taux d'erreur de segmentation (DER), la précision de la segmentation et la pureté. La segmentation en locuteurs est une tâche importante dans le domaine du traitement automatique de la parole, qui consiste à diviser un enregistrement audio en segments correspondant à chaque locuteur présent dans le document.
Dans le cadre de la segmentation d'un locuteur spécifique dans une conversation téléphonique multi-locuteurs, avec ou sans chevauchement entre les locuteurs, Ces mesures nous permettent d'évaluer les performances du système de segmentation en locuteurs dans différentes conditions.

3.2.Base de Données

La base de données (BD) que nous avons utilisée dans nos tests, est un sous ensemble extrait d'une BD téléphonique réelle (NIST2005), constituée de 6 Locuteurs dont 3 femmes et 3 hommes (indiqués par loc1, loc2, loc3, loc4, loc5,loc6).Chaque locuteur est représenté par 02 fichiers (répétitions R1 et R2) de parole téléphoniques échantillonnée à 8kHz et de durée 8 secondes chacun, au total la BD contient12 fichiers de 8s chacun. Cette BD est structurée pour avoir deux scénarios possibles : un premier scénario de conversations téléphoniques sans interférences entre les intervenants, et un deuxième scénario de conversations téléphoniques, cette fois-ci avec interférences entre les intervenants (locuteurs).

Le détail des deux scénarios est résumé dans les deux tableaux suivants :

Tableau3.1. Différents tests de segmentation en locuteurs des conversations téléphoniques sans interférences entre les intervenants

Tests	**Flux audio**	**Nombre de Locuteurs intervenants**	**Durée totale (secondes)**
test 1	Loc1_R1- Loc2_R1- Loc1_R2- Loc2_R2- Loc3_R1	3	40
test 2	Loc3_R1- Loc3_R2- Loc4_R1- Loc2_R2- Loc1_R1	4	40
test 3	Loc6_R1- Loc2_R1- Loc1_R1- Loc1_R2- Loc3_R1	4	40
test 4	Loc1_R1- Loc2_R1- Loc3_R2- Loc4_R1- Loc5_R1	5	40
test 5	Loc5_R1- Loc2_R1- Loc5_R2- Loc6_R2- Loc4_R1	4	40
test 6	Loc1_R1- Loc2_R1- Loc3_R2- Loc3_R1- Loc4_R2- Loc5_R1	5	48
test 7	Loc1_R1- Loc6_R1- Loc3_R2- Loc4_R1- Loc4_R2- Loc5_R1	5	48
test 8	Loc1_R1- Loc2_R1- Loc3_R2- Loc4_R1- Loc5_R2- Loc6_R1	6	48
test 9	Loc1_R1- Loc1_R2- Loc3_R2- Loc3_R1- Loc4_R2- Loc4_R1	3	48
test 10	Loc3_R1- Loc3_R2- Loc1_R2- Loc1_R1- Loc4_R2	3	40

Tableau 3.2 Différents tests de segmentation en locuteurs des conversations téléphoniques avec interférences entre les intervenants

Tests	**Intervalles d'interférences**	**Nombre d'interférences**	**Durée totale (secondes)**
test 1	[6s-8s]	1	38
test 2	[14s-16s],[28s-30s]	2	36
test 3	[6s-8s],[28s-30s]	2	36
test 4	[30s-32s]	1	38
test 5	[6s-8s],[12s-14s],[18s-20s]	3	34
test 6	[6s-8s],[12s-14s],[18s-20s]	3	42
test 7	[14s-16s],[36s-38s]	2	44
test 8	[14s-16s],[20s-22s]	2	44
test 9	[14s-16s],[28s-30s]	2	44
test 10	[14s-16s],[28s-30s]	2	36

3.3.Protocol expérimental

Dans la phase de prétraitement acoustique, nous avons mis en œuvre la méthode de détection d'activité vocale (VAD) pour détecter les segments vocaux et éliminer les périodes de silence dans le flux de parole. Le module de paramétrage a été utilisé pour extraire des vecteurs caractéristiques composés de 14 coefficients MFCC (Mel FrequencyCepstral Coefficients) toutes les 15 millisecondes, en utilisant une fenêtre de Hamming d'une durée de 25 millisecondes. Afin d'évaluer les performances de la segmentation des locuteurs dans les deux scénarios proposés, nous utilisons plusieurs métriques d'évaluation.

Voici un exemple typique de pipeline de diarisation d'un locuteur :

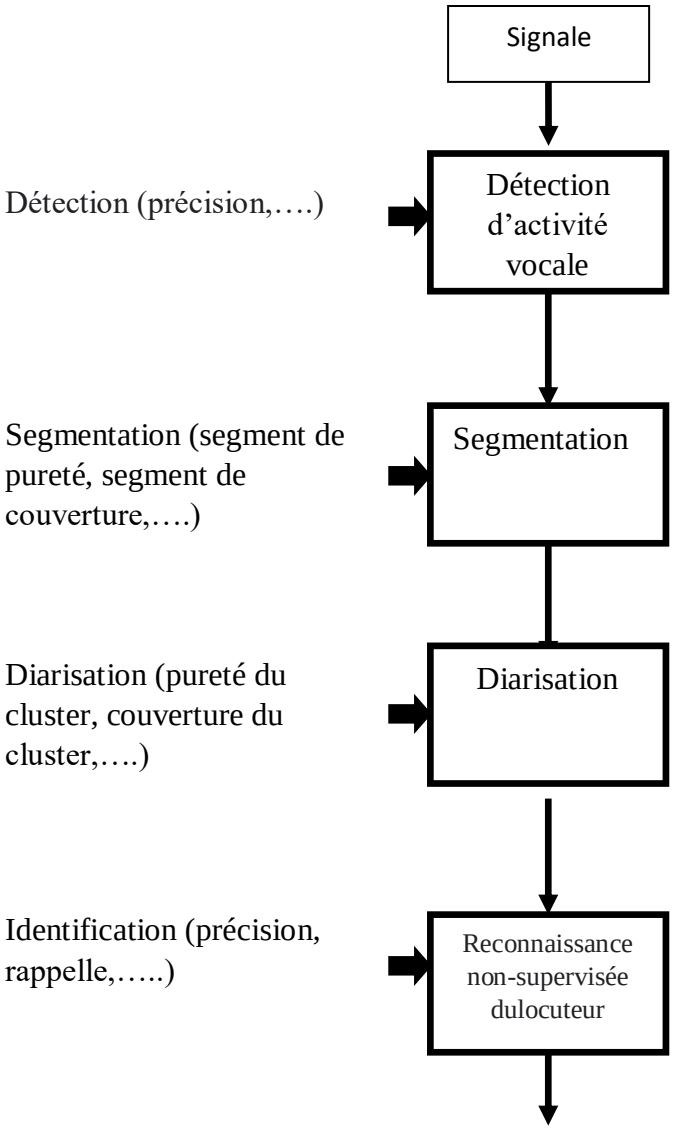


Figure 3.1.pipeline de segmentationen locuteurs d'un flux audio

3.4.Métriques d'évaluation

Chaque étape du système de diarisation a sa propre mesure d'évaluation.

- **La détection d'activité vocale (VAD)**

VAD est généralement évaluée en fonction de l'erreur de fausse alarme et l'erreur de la détection manquée, qui sont deux éléments importants de la DER.

- **Détection de changements de locuteur**

Le système de détection des changements de locuteurs est généralement évalué en fonction de la couverture et de la pureté.

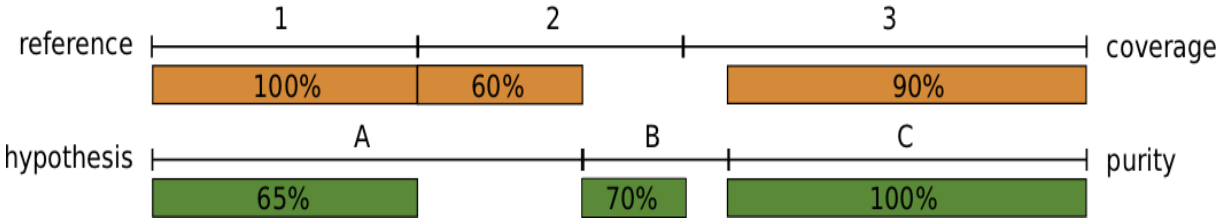


Figure 3.2 Les modules de détection de changement avec pureté et couverture.

Couverture (**Coverage**) :La couverture par segment est calculée pour chaque segment de référence comme le rapport entre la durée de l'intersection avec le segment d'hypothèse le plus co-occurrent et la durée du segment de référence. Par exemple, la couverture du segment de référence 1 est de 100 % car il est entièrement couvert par le segment d'hypothèse A.

La pureté (**Purity**) :est la double mesure qui indique le degré de pureté des segments d'hypothèse. Par exemple, le segment A n'est pur qu'à 65 % car il est couvert à 65 % par le segment 1 et à 35 % par le segment 2.

La pureté et la couverture ont été introduites pour mesurer la qualité des classes, mais elles peuvent également être adaptées à la tâche de détection des points de changement du locuteur.

$$\text{Purity} = \frac{\sum_{\text{cluster}} \max_{\text{speaker}} |\text{cluster} \cap \text{speaker}|}{\sum_{\text{cluster}} |\text{cluster}|} \quad (3.1)$$

$$\text{coverage} = \frac{\sum_{\text{cluster}} \max_{\text{speaker}} |\text{speaker} \cap \text{cluster}|}{\sum_{\text{cluster}} |\text{speaker}|} \quad (3.2)$$

- **|speaker| :**estla durée de la parole de locuteur de référence.
- **|cluster| :** est la durée de la parole de locuteur d'hypothèse.

- **| cluster∩speaker | :**est la durée de leur intersection

➢ **Diarisation**

Erreur de segmentation (DiarizationError Rate) DER est introduit par le NIST comme la fraction de temps de parole qui n'est pas attribuée au bon locuteur, en utilisant une correspondance optimale entre l'annotation des locuteurs des références et des hypothèses. Il est défini comme suit :

$$DER = \frac{\text{False alarm} + \text{Missed detection} + \text{confusion}}{\text{total}} \quad (3.3)$$

- Confusion **:**est la durée de la confusion des locuteurs.
- FA : parole mal-classée (False Alarm), non-parole classéeincorrectement comme parole.
- MD : la détection manquée (MissedDetection) parole classéeincorrectementcommenon-parole (silence).

3.5.Résultats expérimentaux

- **Segmentations en locuteurs sans chevauchement**

Les résultats donnés dans le tableau ci-dessous, présentent les résultats d'évaluation de notre système en se basant sur les trois métriques : MD%, FA%, VAD ERROR%.

Tableau 3.3résultats de diarisation sans chevauchement

Tests	MD%	FA%	VAD Error%	Method of Clustering
test 1	0	0	0	gmm/ac
test 2	0	0	0	gmm/ac
test 3	0	0	0	gmm/ac
test 4	0	0	0	gmm/ac
test 5	0	0	0	gmm/ac
test 6	0	0	0	gmm/ac
test 7	0	0	0	gmm/ac
test 8	0	0	0	gmm/ac
test 9	0	0	0	gmm/ac
test 10	0	0	0	gmm/ac

Les figures ci dessue présente les histogrammes de pureté et couverturedans l'étape de segmentation

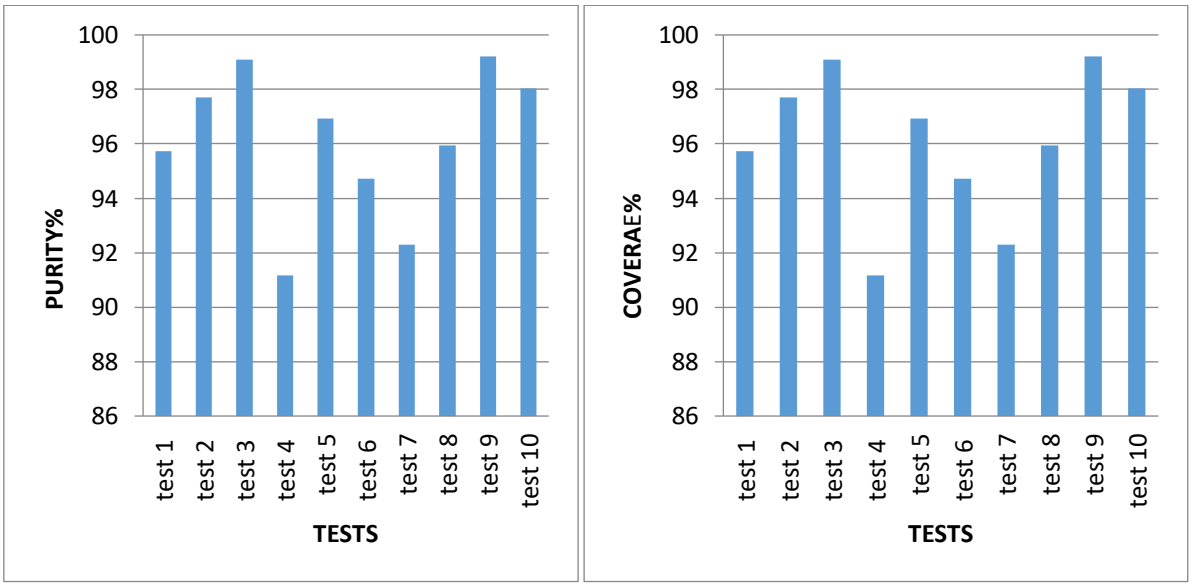


Figure 3. Histogramme des taux de couverture obtenus pour les différents tests

❖ **Discussion**

En regardant l'histogrammesdes figures (3.2et 3.3), nous pouvons observer que la majorité de nos résultats se situe dans la plage des 95% à 99%, ce qui suggère une cohérence élevée dans les performances. Les valeurs les plus fréquentes semblent se concentrer autour de 96% à 99%.Cependant, nous remarquons également quelques résultats légèrement inférieurs dans la plage des 91% à 94%, qui pourraient indiquer des performances relativement moins élevées dans ces cas spécifiques.

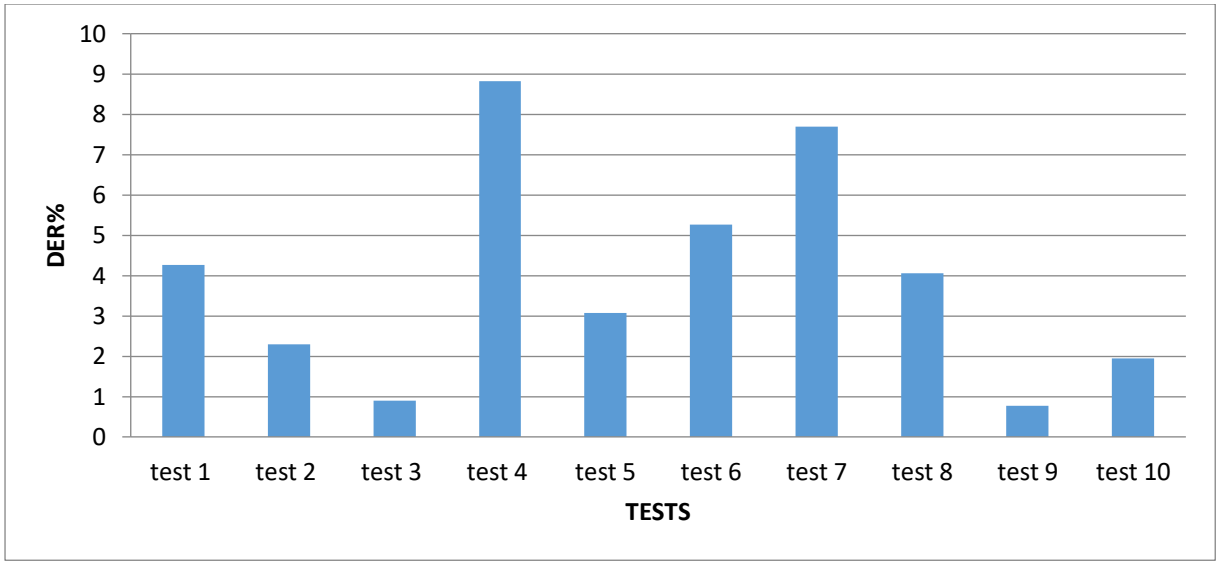


Figure 3.4 Histogramme des taux de DER obtenus pour les différents tests

❖ **Discussion**

En observant les valeurs de la figure 3.4, nous pouvons constater que la plupart des valeurs se situent dans une plage relativement élevée, allant de 2.3 à 8.82. Cependant, il y a deux valeurs qui se distinguent par leur faible niveau de DER, à savoir 0.89 et 0.77.

La valeur de 0.77 représente la plus basse du groupe, ce qui signifie qu'elle indique le taux d'erreur le plus faible parmi les valeurs données. Par conséquent, le meilleur résultat en termes de DER dans cet histogramme est de 0.77.

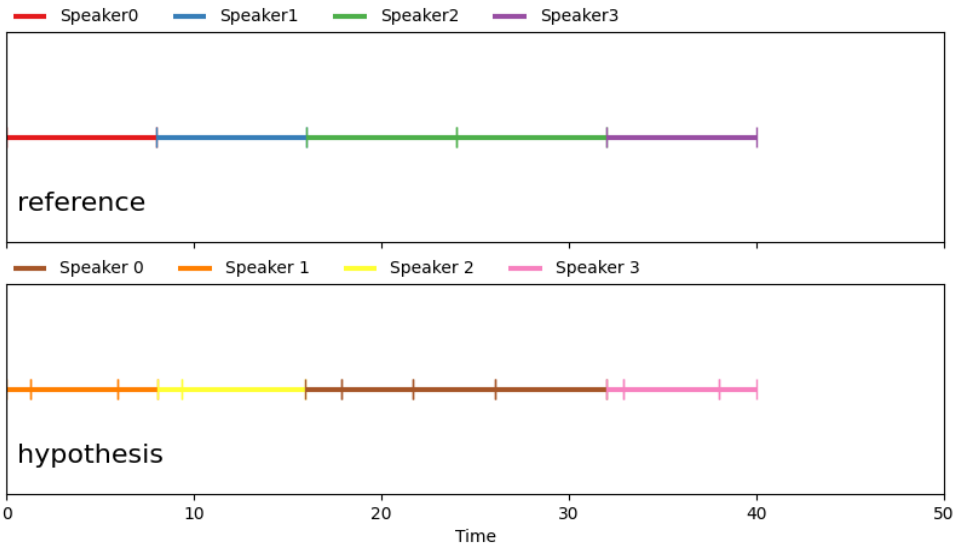


Figure 3.5. Résultats de diarisation sans chevauchement d'un des tests

❖ **Discussion**

Pour ce test spécifique, nous avons comparé les références (speaker 0, speaker 1, speaker 2, speaker 3) avec les hypothèses (speaker 1, speaker 2, speaker 0, speaker 3).

Nous pouvons observer que les locuteurs 1, 2, et 3 sont correctement identifiés et attribués dans les hypothèses, ce qui indique que le système de diarisation a réussi à détecter ces locuteurs avec précision.

Cependant, il y a une différence dans l'attribution du locuteur 0. Dans la référence, il est identifié comme le premier locuteur, alors que dans les hypothèses, il est attribué en tant que le premier et une petite parie de deuxième locuteur. Cela suggère une inversion dans l'ordre des locuteurs détectés.

En conclusion, bien que l'ordre des locuteurs puisse différer entre les références et les hypothèses, la diarisation sans chevauchement a montré une efficacité globale dans l'identification précise des locuteurs dans ce test spécifique.

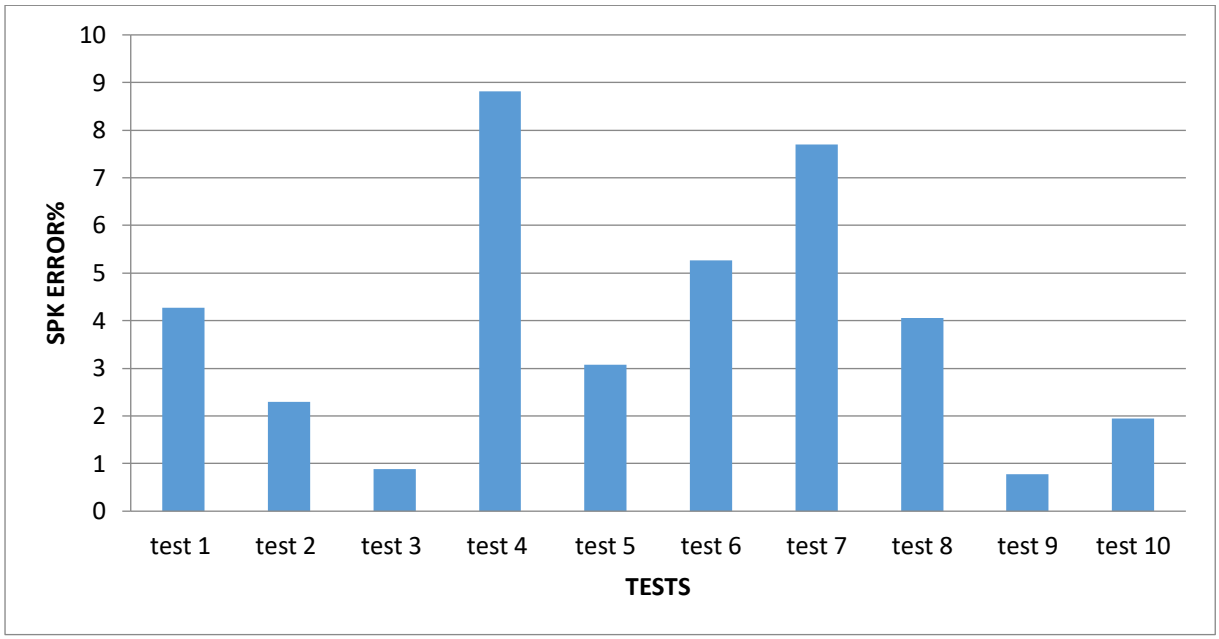


Figure 3.6. Histogramme des taux de spkerror obtenus pour les différents tests

❖ **Discussion**

Dans le contexte de l'histogramme que nous avons fourni, une valeur plus basse de Spk_error indique généralement une meilleure précision et une meilleure performance du système de détection des locuteurs, nous pouvons constater qu'elles représentent des taux d'erreur pour la détection ou l'identification des locuteurs. Plus précisément, elles indiquent le pourcentage d'erreurs commises lors de la tâche de détection des locuteurs.

Parmi les valeurs données, nous remarquons que la plus faible est de 0.77, ce qui suggère un taux d'erreur relativement bas. Cela indique une meilleure performance en termes de détection des locuteurs par rapport aux autres valeurs.

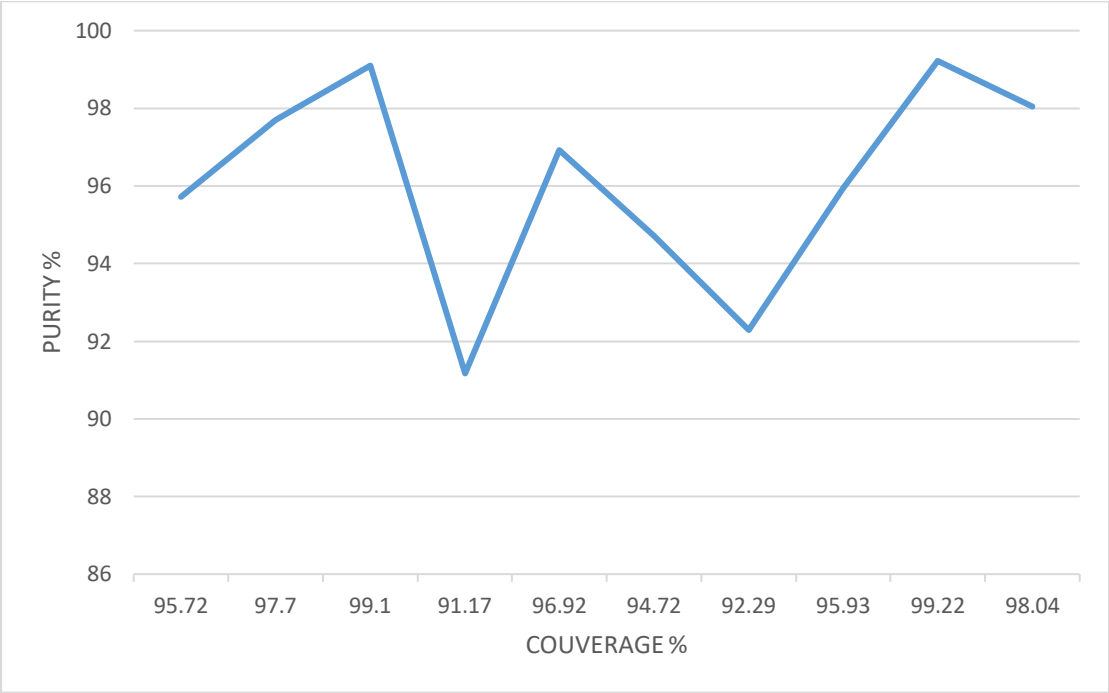


Figure3.7courbe de taux de pureté en fonction de couverture obtenus pour les différents tests

❖ **Discussion**

En examinant la courbe de la figure (3.5), La pureté mesure la proportion d'échantillons correctement classés par rapport à l'ensemble des échantillons. Les résultats de pureté varient entre 91.17% et 99.22%. Ces valeurs indiquent une performance globalement élevée du modèle dans la classification correcte des échantillons.

La couverture évalue la proportion d'échantillons couverts par rapport à l'ensemble des échantillons réels. Les valeurs de couverture se situent également entre 91.17% et 99.22%, ce qui suggère une capacité du modèle à capturer la majorité des échantillons réels.

Cette courbe montre une performance solide du modèle, avec des valeurs élevées de pureté et de couverture pour chaque mesure. Cela suggère que le modèle est capable de classifier avec précision un pourcentage élevé d'échantillons tout en capturant la majorité des échantillons réels.

- **Segmentations en locuteurs avec chevauchement**

Les résultats donnés dans le tableau ci-dessous, présentent les résultats du pourcentage de chevauchement et de FA.

Tableau 3.4. Résultats de diarisationavec chevauchement

Tests	**% of overlap**	**FA %**	**Method of Clustering**
test 1	5.20	0	gmm/ac
test 2	11.11	0	gmm/ac
test 3	11.11	0	gmm/ac
test 4	5.20	0	gmm/ac
test 5	17.64	0	gmm/ac
test 6	14.28	0	gmm/ac
test 7	9.09	0	gmm/ac
test 8	9.09	0	gmm/ac
test 9	9.09	0	gmm/ac
test 10	11.11	0	gmm/ac
All tests	10.3	0	gmm/ac

❖ **Discussion**

En analysant les résultats, plusieurs observations peuvent être faites :

1. Pourcentage de chevauchement : Les tests 5 et 6 ont les pourcentages de chevauchement les plus élevés, ce qui signifie que ces enregistrements audio ont une plus grande présence de chevauchement entre les locuteurs. Cela peut rendre la tâche de diarisation plus difficile.

2. Erreurs de fausse alarme (FA%) : La plupart des tests ont un faible taux d'erreurs de fausse alarme, ce qui est positif, car cela indique que le système n'attribue pas de manière incorrecte des segments de parole à des locuteurs non présents.

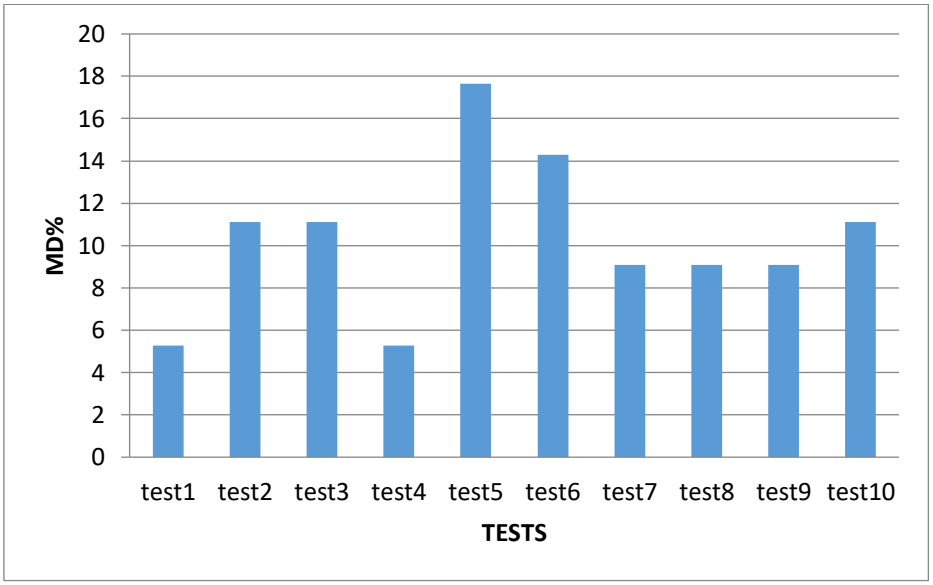


Figure 3.8 Histogramme des taux de MD obtenus pour les différents tests avec chevauchement

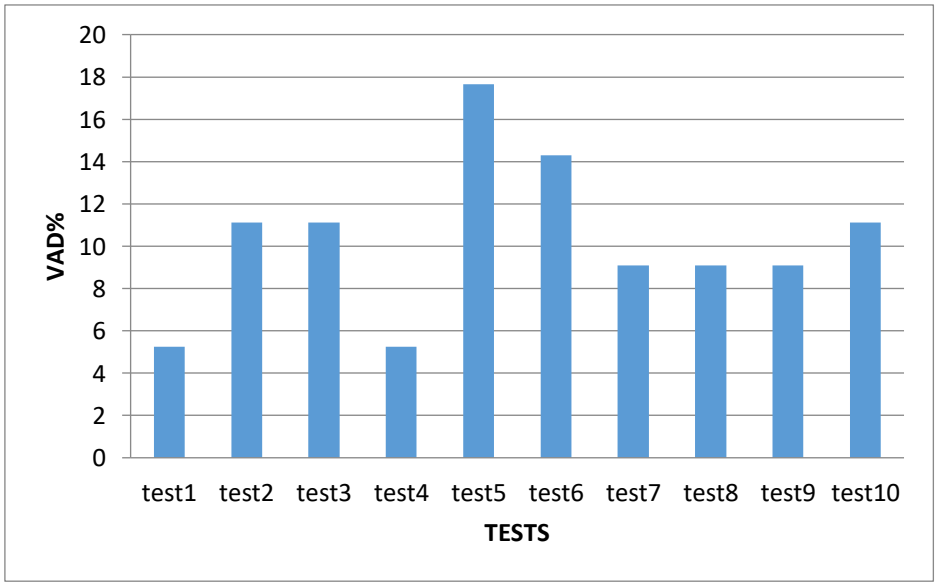


Figure 3.9histogramme des taux de VAD obtenus pour les différents tests avec chevauchement

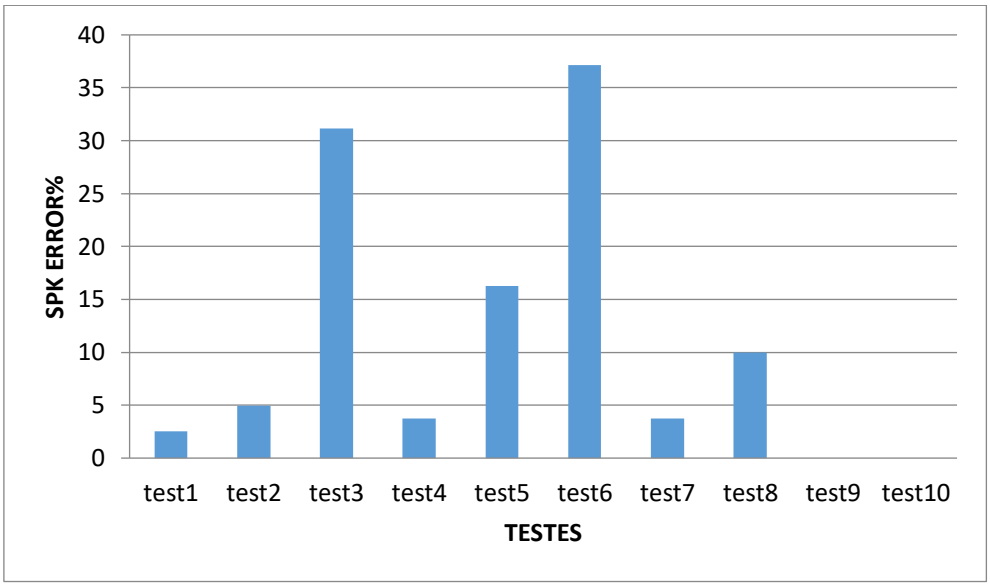


Figure 3.10 Histogramme des taux de spk_error obtenus pour les différents tests avec chevauchement

❖ **Discussion**

Dans cette discussion, nous pouvons observer les meilleures valeurs pour chaque mesure dans les figures 3.8 et 3.9 et 3.10 :

- En ce qui concerne le % de discours multiple (MD%), les tests 1, 4 et 7 obtiennent les meilleures valeurs avec 5.26% de discours multiple, ce qui suggère une meilleure séparation des locuteurs dans ces tests.
- Pour le taux d'erreur de locuteur (Spk_error%), le test 9 avec 0% présente la meilleure performance, indiquant une identification précise des locuteurs dans ce test.
- Le taux d'erreur (Error%) le plus faible est également obtenu par le test 9 avec 9.09%, ce qui indique une précision élevée dans la diarisation.
- En ce qui concerne la détection d'activité vocale (VAD), les tests 1, 4, 7 et 9 obtiennent tous la meilleure valeur de 5.26%, ce qui indique une bonne identification des régions de parole dans le signal.

- En général, un chevauchement plus élevé peut rendre la tâche de diarisation plus difficile et entraîner des performances plus faibles.

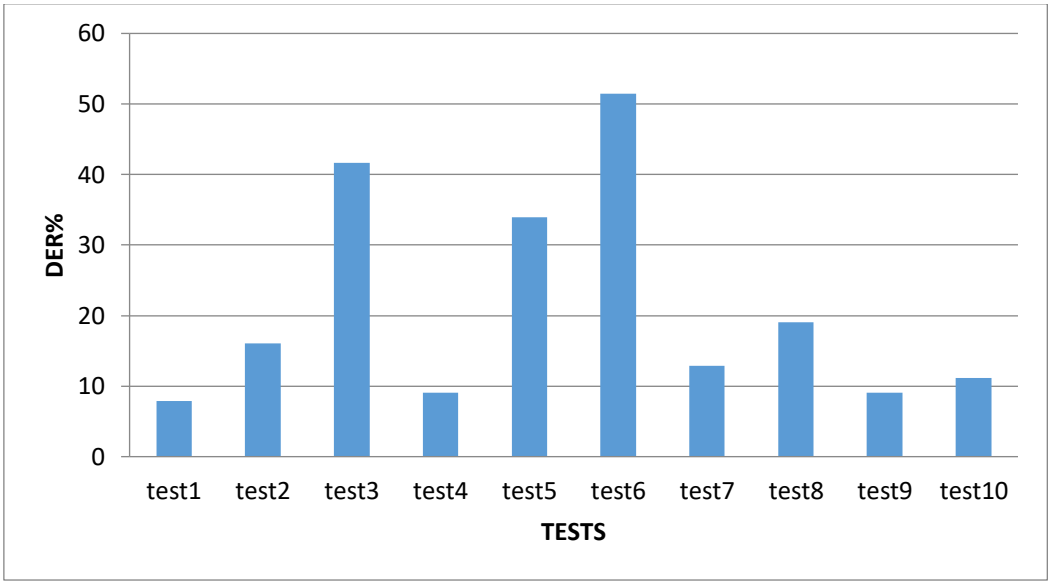


Figure 3.11. Histogramme des taux de DER obtenus pour les différents tests avec chevauchement

❖ **Discussion**

Les tests 1, 4 et 9 ont les taux d'erreur de diarisation les plus bas, avec des valeurs de 7.91%, 9.02% et 9.09% respectivement Les tests 3 et 6 présentent les taux d'erreur de diarisation les plus élevés, avec des valeurs de 41.61% et 51.42% respectivement. Cela suggère une moins bonne performance de la diarisation dans ces tests, avec un nombre plus élevé d'erreurs globales.

Les autres tests ont des taux d'erreur de diarisation compris entre 11.11% et 33.90%. Ces valeurs représentent une gamme de performances intermédiaires en termes de précision de la diarisation.

Le test ayant le taux d'erreur de diarisation (DER%) le plus bas est le test 1 avec une valeur de 7.91%. Il affiche la meilleure performance globale en termes de segmentation et d'attribution précise des locuteurs dans l'enregistrement audio analysé.

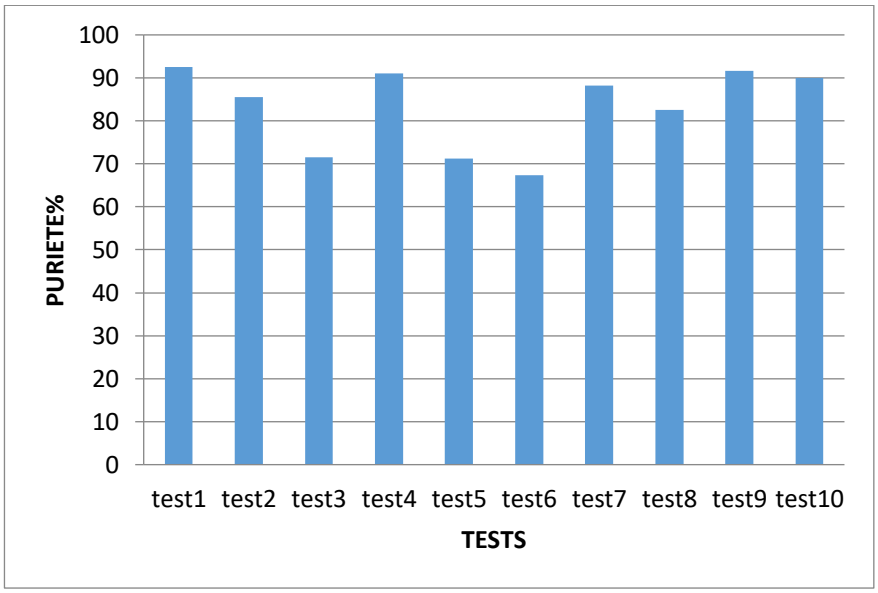


Figure 3.12. Histogramme des taux de pureté obtenus pour les différents tests avec chevauchement

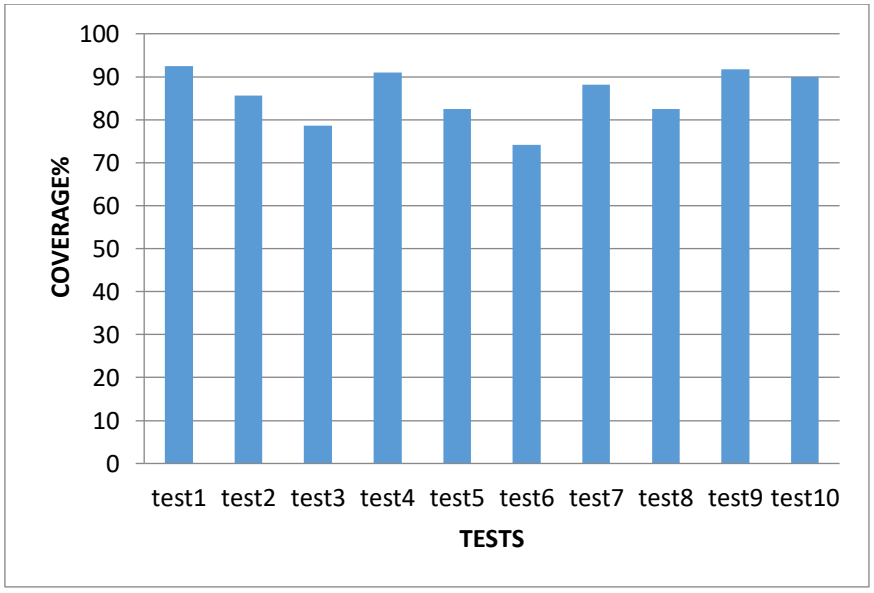


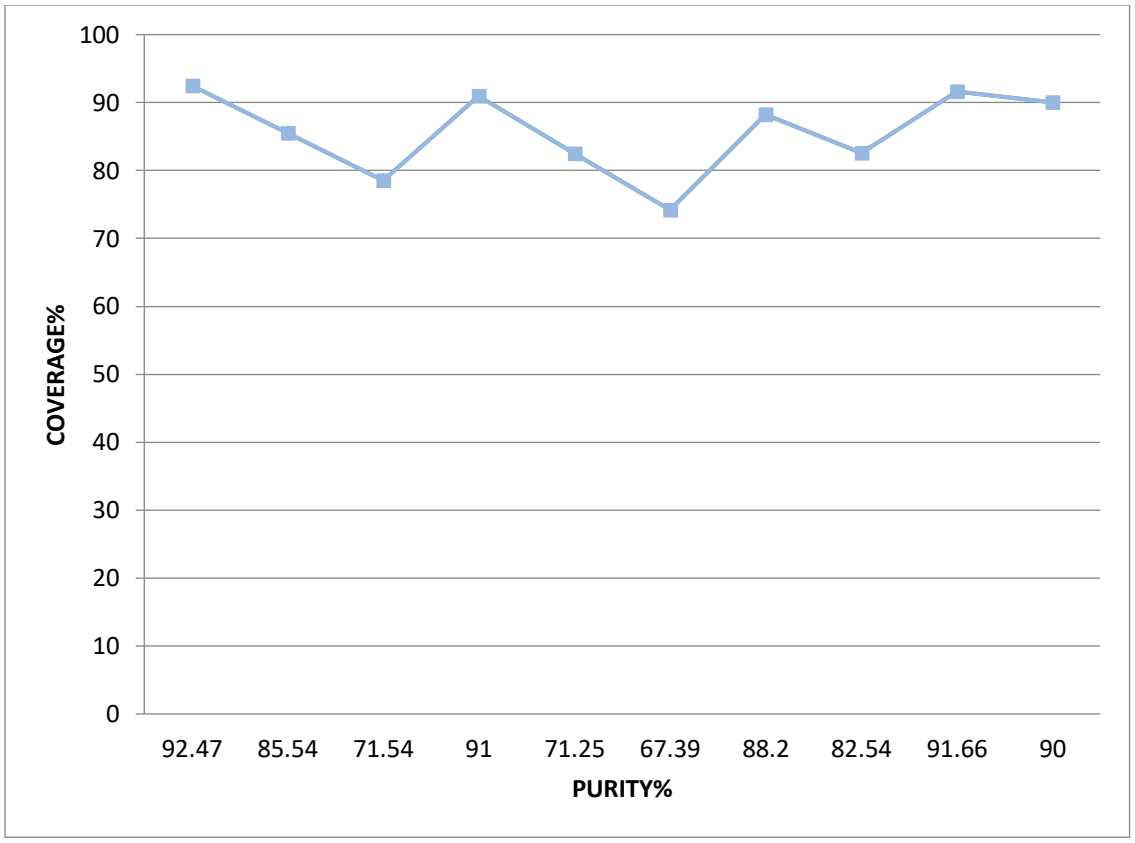
Figure 3.13. Histogramme des taux de couverture obtenus pour les différents tests avec chevauchement

❖ **Discussion**

En analysant les résultats des figures (3.12 et 3.13), nous pouvons observer que les tests 1 et 4 ont les meilleures valeurs pour Pureté% et Couverture% avec des scores élevés de 92.47% et

91% respectivement. Cela indique une bonne précision et couverture de la diarisation dans ces tests.

Les tests de diarisation présentent des variations dans les mesures de Pureté% et Couverture%, avec certains tests montrant une meilleure précision et couverture que d'autres. Les tests 1 et 4 se distinguent en affichant les meilleurs résultats dans ces deux mesures.



.

Figure 3.14.courbe de taux de pureté en fonction de couverture obtenus pour les différents tests avec chevauchement

❖ **Discussion**

En analysant la courbe, nous pouvons noter les observations suivantes :

- Les tests 1 et 4 ont les valeurs les plus élevées de Purity% et Coverage%, indiquant une diarisation précise et une couverture complète du signal de parole.

- Les tests 6 et 7 ont des valeurs relativement faibles de Purity% et Coverage%, suggérant des performances inférieures en termes de précision et de couverture.
- Les tests 3, 5 et 9 présentent des valeurs de Coverage% plus élevées par rapport à leur Purity%, ce qui peut indiquer une meilleure couverture globale, mais une précision légèrement inférieure en termes de regroupement des locuteurs.
- Les tests 2 et 8 ont des valeurs relativement équilibrées de Purity% et Coverage%, montrant une performance globalement cohérente en termes de précision et de couverture.

Ces tests (test 1 et test 4) ont obtenu les meilleures performances en termes de Purity% et Coverage%, ce qui indique une diarisation précise et une couverture complète du signal de parole.

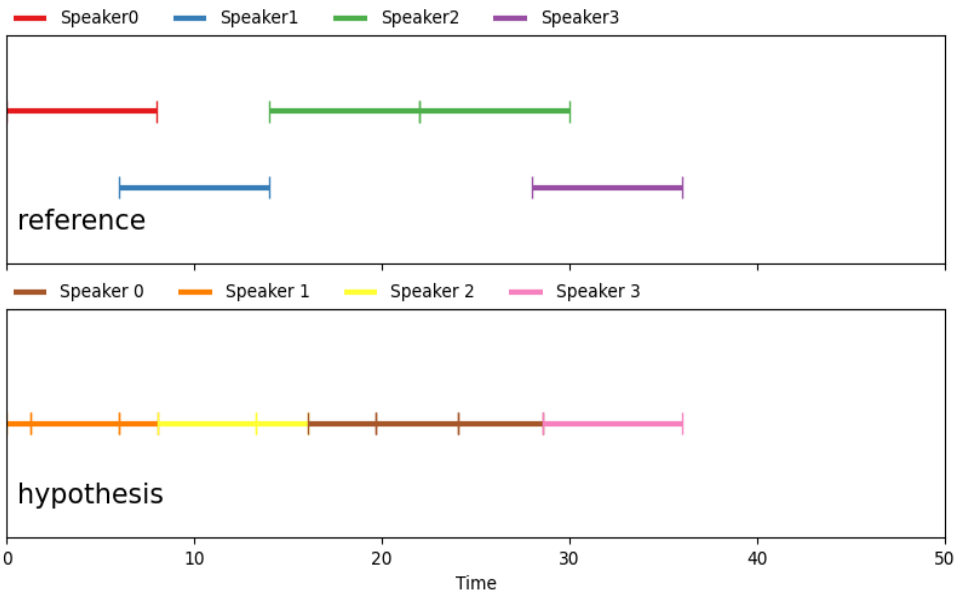


Figure 3.15.résultat de diarisation avec chevauchement d'un des tests

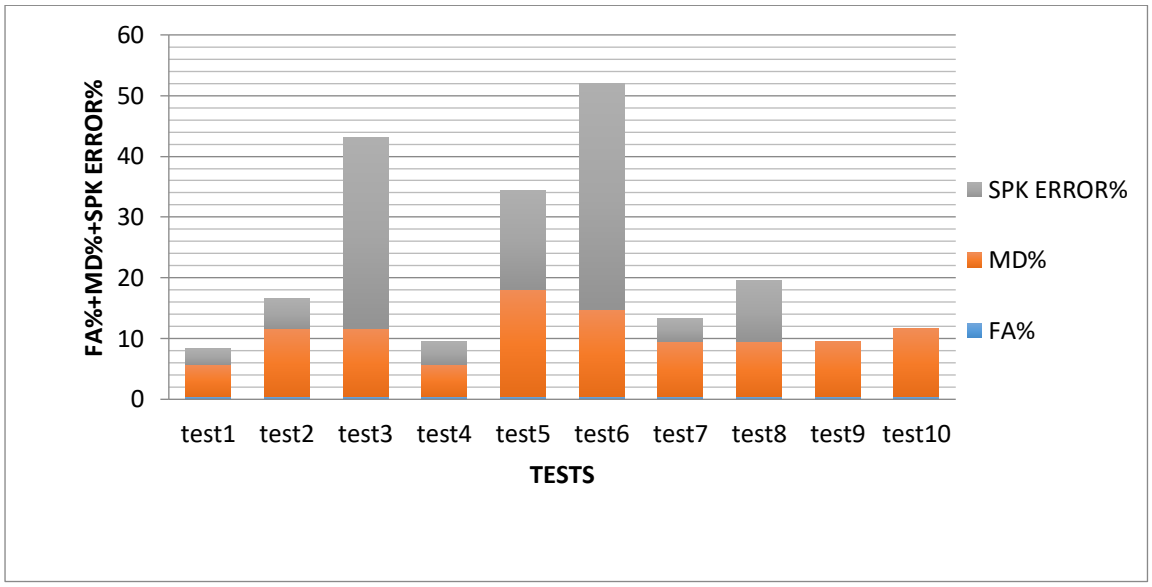


Figure 3.16histogramme des taux deDER et SPK ERROR et MD obtenus pour les différents tests avec chevauchement

❖ **Discussion**

Les valeurs données dans les colonnes MD%, FA% et Spk_error% sont des taux d'erreur de segmentation des locuteurs (DER - DiarizationError Rate) pour différentes conversations téléphoniques multi-locuteurs. La meilleure valeur serait celle qui présente le taux d'erreur le plus bas, indiquant une meilleure performance de la segmentation des locuteurs.

En examinant les valeurs fournies, il semble que la conversation ayant le taux d'erreur de segmentation des locuteurs le plus bas est la suivante :

MD% : 9.09 FA% : 0 Spk_error% : 0

Cela signifie que dans cette conversation spécifique, il y a eu un taux d'erreur de 9.09% pour la détection manquée (MissedDetection), aucun pourcentage d'erreurs de fausse détection (False Alarms) et un taux d'erreur total de 0%.

3.6.Conclusion

Dans ce chapitre, nous avons présenté en détail les expériences que nous avons menées et les résultats que nous avons obtenus au cours de notre travail. Nous avons simulé deux scénarios pour notre système afin de l'évaluer dans des conditions réalistes, en le comparant à un scénario idéal.

Nos résultats démontrent de manière convaincante que notre système fonctionne de manière excellente dans un environnement propre et sans interférences. Malgré les courtes durées des interventions dans les flux de conversations (entre 8 et 16 secondes), notre système parvient à maintenir des performances élevées. Cela témoigne de l'efficacité de toutes les étapes que nous avons soigneusement développées, notamment l'extraction des paramètres, la détection des changements de locuteurs et le regroupement des segments.

Cependant, nous avons également constaté une forte dégradation des performances de notre système lorsque des interférences entre les locuteurs se produisent dans les flux de conversations. Cette dégradation était prévisible, car il est bien connu dans la littérature que le chevauchement des paroles a un impact négatif sur la robustesse et la précision du système de segmentation en locuteurs.

En somme, nos expérimentations ont confirmé la fiabilité et la justesse de notre système dans des conditions idéales, tout en soulignant les défis que représentent les interférences entre les locuteurs. Ces résultats nous incitent à poursuivre nos recherches pour améliorer la résistance de notre système face à de telles situations, afin de le rendre encore plus performant et adaptable à des environnements réels complexes.

Conclusion générale

Conclusion générale

Le domaine du traitement du signal est un vaste domaine et ce travail s'inscrit dans ce domaine, en particulier le traitement de la parole. Nous appliquons également à la segmentation en locuteurs d'un document audio, plus précisément des conversations téléphoniques, dans le but de faire une détection de changement de locuteur et un regroupement de segments corrigés dans le cas où les durées des interventions sont courtes et interférées entre eux.

Pour cela, nous avons simulé et évalué un système de segmentation à l'aide de tests sur des flux audio, où il y a interférence entre locuteurs, qui affiche l'état réel lorsqu'il y a interférence entre locuteurs.

A travers les expériences que nous avons menées et les résultats que nous avons obtenus au cours de nos travaux. Nous avons simulé deux scénarios pour notre système afin de l'évaluer dans des conditions réelles et l'avons comparé à un scénario idéal.

Nos résultats montrent que notre système fonctionne parfaitement dans un environnement propre et sans interférence. Malgré les courtes durées des intrusions dans les flux de conversation (entre 8 et 16 secondes), notre système est capable de maintenir des performances élevées. Cela montre l'efficacité de toutes les étapes que nous avons soigneusement développées, y compris l'extraction des paramètres, la détection du changement de locuteur et le regroupement des segments.

Cependant, nous avons également remarqué une forte dégradation des performances de notre système lorsque des interférences se produisaient entre les locuteurs dans les flux de conversation.

En résumé, nos expériences ont confirmé la fiabilité et la précision de notre système dans des conditions idéales, tout en soulignant les défis posés par les interférences des locuteurs. Ces résultats nous encouragent à poursuivre nos recherches pour améliorer la résistance de notre système à de telles situations, afin de le rendre plus efficace et adaptable à des environnements réels complexes.

Références

Références

[1] Laboratoires Unisson. (2022, 22 juillet). *Quelles est la fréquence de la voix humaine ? - Laboratoires Unisson*. https://www.laboratoires-unisson.com/faq/technique/quelle-est-la-frequence-de-la-voix-humaine

[2]Lecocq, Pierre. *Lexique 8/L'accès lexical*. Vol. 8. Presses Univ. Septentrion, 1989.

[3] Delacourt, Perrine, and Christian J. Wellekens. "Segmentation en locuteurs d'un document audio." CORESA.

[4]Taylor, P., Black, A. W., &Caley, R. (1998). The architecture of the Festival speech synthesis system. In *The third ESCA/COCOSDA workshop (ETRW) on speech synthesis*.

[5]Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain* (No. 202). WW Norton & Company.

[6]Gold, B., Morgan, N., & Ellis, D. (2011). *Speech and audio signal processing: processing and perception of speech and music*. John Wiley& Sons.

[7] DEBILOU Chaima/ BOUDAOUD Samiha .2011. Amélioration d'un synthétiseur de la parole par concaténation.

[8]Koriba, M. (2010). *Reconnaissance automatique de la parole par LPC, MFCC et PLP. Application aux signaux GSM* (Doctoral dissertation, Alger).

[9]McAulay, R., &Quatieri, T. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *34*(4), 744-754.

[10]O'shaughnessy, D. (1987). *Speech communication: human and machine*. Universities press.

[11] Rabiner, L., &Juang, B. H. (1993). *Fundamentals of speech recognition*. Prentice-Hall, Inc..

[12] TALEB, O., Nabil, K. T., & Yassine, P. S. (2022). Conference Proceedings: NewMat'21–1st International Conférence: New Trends on Innovative Construction Materials-ESSA-Tlemcen (Algeria)–22, 23 March, 2022.

[13] Fréderic, C. (2013). Théorie Et Traitement Des Signaux. Presses polytechniques et universitaires romandes. Suisse

[14]Burrus, T. V., Burrus, C., Narasimhan, K., Guo, Y., & Li, C. Introduction To Wavelets And Wavelet Transforms-A Primer, Brrus CS, 1998.

[15]"Digital Image Processing" par Rafael C. Gonzalez et Richard E. Woods (2018),: "Image Enhancement in the Frequency Domain", Section 4.3 : "The Discrete Cosine Transform (DCT)".

[16]Marine, C. A. M. P. E. D. E. L., & Pierre, H. O. O. G. S. T. O. Ë. L. (2011). *Sémantique et multimodalité en analyse de l'information*. Lavoisier.

[17]RICHARD, G. (2012). Eléments de Reconnaissance de la Parole pour PACT Télécom ParisTech Extraits du poly de cours de l'UE SI340.

[18] B.Tinhinane (2020) *Segmentation en locuteurs d'un document audio* (mémoire de fin d étude ,Université de Mohamed El-Bachir El-Ibrahimi - Bordj Bou Arreridj ,Faculté des Sciences et de la technologie ,Département d'Electronique).

[19]Zaabi, K. (2004). *Implémentation d'une méthode de reconnaissance de la parole sur le processeur de traitement numérique du signal TMS320C6711* (Doctoral dissertation, École de technologie supérieure).

[20]Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., ...& Woodland, P. (2002). The HTK book. *Cambridge university engineering department*, *3*(175), 12.

[21] Harba, R. (2012). Sélection de parametres acoustiques pertinents pour la reconnaissance de la parole. *These de doctorat, Université d'Orléans (France)*.

[22]Dupuy, G. (2015). *Les collections volumineuses de documents audiovisuels: segmentation et regroupement en locuteurs* (Doctoral dissertation, Université du Maine).

[23] Reynolds, D. A. (1992). *A Gaussian mixture modeling approach to text-independent speaker identification*. Georgia Institute of Technology.

[24]Moraru, D. (2004). *Segmentation en locuteurs de documents audio et audiovisuels: application à la recherche d'information multimédia* (Doctoral dissertation, Grenoble INPG).

[25] Dunn, R. B., Reynolds, D. A., &Quatieri, T. F. (2000). Approaches to speaker detection and tracking in conversational speech. *Digital signal processing*, *10*(1-3), 93-112.

[26] Reynolds, D., Campbell, J., Dunn, B., Jones, D., Sturim, D., &Quatieri, T. (2002, May). Mitlincoln laboratory system: 1sp, 2sp and segmentation. In *Proc of NIST SpRec 2002 Workshop, Vienna, VA*.

[27] S. Ouamour, H. Sayoud, & M. Boudraa, Application of Statistical Measures for the detection of speaker transitions, AMAM'03, Nice (France), 10-13 Février 2003. http://acm.emath.fr/amam/.

[28] Zhu, X. (2007). *Structuration automatique en locuteurs par approche acoustique* (Doctoral dissertation, Université Paris Sud-Paris XI).

[29]Delacourt, P. SEGMENTATION ET INDEXATION PAR LOCUTEURS D'UN DOCUMENT AUDIO.

[30]"Speaker Change Detection Using the Normalized Cut Distance and Spectral Clustering" de N. Ryant, J. W. Pittonet M. J. Collier, publié en 2008

[31] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100-108.

[32]Delacourt, P., &Wellekens, C. J. Segmentation en locuteurs d'un document audio. CORESA.

[33] Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., &Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on audio, speech, and language processing*, *20*(2), 356-370.

[34] Barras, C., Zhu, X., Meignier, S., &Gauvain, J. L. (2006). Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(5), 1505-1512.

[35] DUPUY, Grégor. *Les collections volumineuses de documents audiovisuels: segmentation et regroupement en locuteurs*. 2015. Thèse de doctorat. Université du Maine.