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

DJERARDA THAMEUR

BOUDISSA AMEL

Intitulé

Détection de parole superposée par les modèles ensemblistes

Soutenu le 18/09/2024 :

Devant le Jury composé de :

Nom & Prénom	*Grade*	*Qualité*	*Etablissement*
Mme.Nora MELIZI	*M.C. A*	*Président*	*Univ-BBA*
Mr ABDENOUR HACINE GHARBI	*M.C. A*	*Examinateur*	*Univ-BBA*
Mr ASBAI NASSIM	*M.C. A*	*Encadrant*	*Univ-BBA*
Mme. BOUNAZOU HADJER	*Doctorante*	*Co-encadrante*	*Univ-BBA*

Année Universitaire 2023/2024

Dédicace

Je tiens à dédier ce modeste travail avec grand plaisir :

À mes chers parents pour leur soutien, leur patience et leurs encouragements durant mon parcours scolaire.

À l'ensemble des étudiants de la promotion Master Système Télécommunication de l'année 2023/2024 et à toutes les personnes qui occupent une place dans mon cœur.

À tous les membres de ma famille, à toutes les personnes portant le nom DJERARDA et BOUDISSA, ainsi qu'à tous ceux qui ont contribué à ma réussite.

Dédicace

Je dédie le fruit de mon travail à ma mère, qui m'a donné vie et espoir, et m'a appris à avancer avec sagesse et patience. Je remercie également mes sœurs Hakima et Celia pour leur soutien constant.

Ma profonde gratitude va à mon mari, qui a soutenu mon rêve avec dévouement et m'a offert une aide précieuse, ainsi que son inspiration.

À ceux qui se sont unis à moi et qui ont ouvert la voie au succès de notre carrière scientifique, Thameur DJerarda.

Enfin, je remercie ma famille, mes amis et toutes les personnes qui m'ont soutenue, de près ou de loin, lors de l'élaboration de ce travail.

Remerciements

Une personne n'a rien d'autre que ce à quoi elle aspire après des années d'efforts et de diligence. Aujourd'hui, nous avons été couronnés de lauriers d'excellence et de réussite, et ce qui était notre rêve d'hier est devenu réalité en un instant. Nous l'avons toujours voulu et avons travaillé dur pour y parvenir. Louange à Dieu, qui facilite les débuts et les fins avec excellence. Nous le remercions pour ses bienfaits.

Nous remercions sincèrement le Dr N. ASBAI, de l'Université BBA, d'abord pour avoir été le directeur de cette thèse, puis pour ses précieuses suggestions, ses encouragements constants et surtout pour son aide extrêmement utile tout au long de la réalisation de ce travail.

Nous remercions également HADJER BOUNAZOU, co-directrice de la thèse, pour ses encouragements et son soutien à cette modeste étude.

Nous remercions sincèrement nos chers parents pour leur soutien moral tout au long de notre parcours éducatif, et surtout durant les moments difficiles que nous avons traversés pour mener à bien ce travail.

Enfin, nous remercions les membres du jury d'avoir accepté d'évaluer notre travail.

Résumé : La détection de parole superposée est essentielle dans divers domaines tels que la reconnaissance vocale et la communication. Ce travail explore les techniques de modélisation statistique, en particulier les modèles ensemblistes (Arbre de décision, Random Forest et AdaBoost), pour améliorer la détection des voix simultanées dans les enregistrements audios. Nous analysons les différentes approches, de la collecte des données à l'évaluation des modèles, en mettant l'accent sur l'efficacité des modèles en environnements complexes. Les résultats montrent que la Forêt Aléatoire atteint une précision de 92%, ce qui est supérieur à celle de l'Arbre de décision (78%) et d'AdaBoost (85%). La Forêt Aléatoire se distingue par une capacité améliorée à différencier les classes, avec un taux de faux positifs réduit de 10% par rapport à AdaBoost et une meilleure gestion des bruits de fond dans des enregistrements à faible rapport signal/bruit (SNR).

Mots-clés : Détection de la parole superposée ; MFCC ; classificateur Adaboost ; classificateur Random Forest ; classificateur d'arbres de décision ; précision de détection.

Abstract: Overlapping speech detection is crucial in fields such as speech recognition and communication. This study explores statistical modeling techniques, particularly ensemble methods (Decision Tree, Random Forest and AdaBoost), to improve the detection of simultaneous voices in audio recordings. We analyze various approaches, from data collection to model evaluation, with a focus on model performance in complex environments. The results show that Random Forest achieves an accuracy of 92%, which is higher than that of Decision Tree (78%) and AdaBoost (85%). Random Forest is distinguished by an improved ability to differentiate classes, with a 10% lower false positive rate compared to AdaBoost and better handling of background noise in low signal-to-noise ratio (SNR) recordings.

Keywords: Detection of overlapped speech; MFCCs; Adaboost classifier; Random Forest classifier; Decision Tree Classifier; detection accuracy.

ملخص

تعدّ اكتشاف الكلام المتداخل أمًرا حيويًا في مجاالت مثل التعرف على الصوت واالتصاالت. تستكشف هذه الدراسة تقنيات النمذجة اإلحصائية، خاصة طرق التجّمع) الغابات العشوائية، AdaBoost)، لتحسين اكتشاف أألصوات المتداخلة في التسجيالت الصوتية. نقوم بتحليل طرق مختلفة، بدًء يز على أداء ا من جمع البيانات وصوًال إلى تقييم النماذج، مع الترك

النماذج في البيئات المعقدة.تظهر النتائج أن Random Forest أكثر قوة وتحقق دقة بنسبة 92%، وهي أعلى من دقة Decision Tree (78%) وAdaBoost (85%). تتميز Random Forest بقدرة محسنة على التمييز بين الفئات، مع معدل إيجابيات كاذبة أقل بنسبة 10% مقارنة بـ AdaBoost ومعالجة أفضل للضوضاء الخلفية في التسجيلات ذات نسبة الإشارة إلى الضوضاء المنخفضة (SNR).

الكلمات المفتاحية: : الكشف عن تراكب الكلام؛ مصنف MFCC؛ مصنف Adaboost؛ مصنف الغابة العشوائية؛ مصنف شجرة القرار؛ دقة الكشف.

Table des matières

Liste des abréviations

ADN : Acide Désoxyribonucléique.

ASR: Automatic Speech Recognition.

CART: Classification and Regression Tees.

DNN: Deep Neural Network

EER: Equal Error Rate.

FFT: Fast Fourrier transform.

F0: Fréquence fodndamentale.

HMM: Hidden Markov Model

IDCT: Inverse Discrete Cosinus Transform.

ID3: Iterative Dichotomiser 3.

MFCC: Mel Frequency cepstral coefficients.

OOB: Out-OF-Bag.

PAC: probablement approximativement correct.

RGPD : Règlement Général sur la Protection des Données.

RF: Random Forest.

STFT: Shor-Time Fourier Transform.

TPZ : Taux de passage par zéro.

TED : Text-Dependent

VAD : Détection de l'activité vocale.

Liste des figures

Liste des tableaux

Introduction Générale

INTRODUCTION GENERALE

L'évolution rapide des technologies de traitement du signal et de la parole a permis de grandes avancées dans la reconnaissance vocale et la compréhension des interactions humaines. L'un des défis majeurs de ce domaine est la **détection de parole superposée**, qui consiste à identifier et à isoler les segments de parole provenant de différents locuteurs dans un enregistrement audio. Cette tâche est cruciale dans divers contextes, tels que les réunions, les conversations téléphoniques, ou encore l'analyse de données vocales pour des applications de surveillance ou d'assistance vocale [1].

Le travail présenté dans cette étude explore les différentes techniques et approches utilisées pour améliorer la détection de parole superposée, en s'appuyant notamment sur les **modèles ensemblistes** [2]. Ces modèles, qui combinent les prédictions de plusieurs algorithmes individuels, permettent d'améliorer la précision des systèmes de détection, en prenant en compte les variabilités des signaux vocaux et des environnements dans lesquels ces signaux sont capturés [3]. Ils constituent une solution efficace pour surmonter les limites des méthodes traditionnelles de traitement de la parole.

Cette étude examine à la fois les aspects théoriques et pratiques de la détection de parole superposée. Elle inclut une analyse approfondie des techniques de prétraitement des données, de l'extraction des caractéristiques acoustiques, ainsi que de l'évaluation des performances des modèles utilisés. Les modèles comme la Forêt Aléatoire, les arbres de décision et AdaBoost sont particulièrement mis en avant pour leur capacité à capturer des informations complexes à partir de données vocales et à les exploiter pour améliorer la qualité de la détection.

En s'appuyant sur des expérimentations et des évaluations rigoureuses, cette recherche met en lumière les avantages et les limites des différentes approches utilisées, tout en proposant des pistes pour améliorer les performances futures des systèmes de détection de parole superposée. Ce travail constitue ainsi une contribution importante au domaine du traitement de la parole et de l'audio, offrant des perspectives prometteuses

Pour des applications variées.

Le premier chapitre représente les concepts de base du signal de parole ainsi que certaines méthodes de traitement pour l'analyse du signal, tandis que dans le deuxième chapitre, nous expliquerons en détail la structure du système de segmentation du locuteur.

Le troisième chapitre contient l ensembles des expériences effectuées et les résultats que nous obtenus.

Chapitre 1 fondements de la détection de parole superposée

1.1 Introduction

Dans ce chapitre, nous explorerons en détail les différentes approches et techniques utilisées dans la détection de parole superposée, en examinant les méthodes traditionnelles et les avancées récentes, ainsi que les défis persistants dans ce domaine.

Nous analyserons également les modèles ensemblistes comme une solution prometteuse pour améliorer la détection et la séparation des voix superposées dans des enregistrements audio complexes.

En combinant la robustesse et la précision de plusieurs modèles individuels, les modèles ensemblistes offrent un potentiel significatif pour améliorer la détection de parole superposée et répondre aux exigences croissantes des applications en temps réel.

1.2 Introduction à la détection de parole superposée

La détection de parole superposée consiste à identifier et isoler les segments de parole qui se produisent simultanément dans un enregistrement audio. Elle permet d'extraire des informations spécifiques à partir de signaux complexes, tels que des conversations téléphoniques ou des réunions [4].

Pour y parvenir, des outils de traitement du signal sont développés, notamment pour détecter la présence d'activité vocale mono et multi-locuteur. Ces outils permettent d'étudier la durée et le contenu linguistique des segments multi-locuteurs. En outre, des modèles neuronaux spécifiques [4] sont utilisés pour détecter les interruptions dans les conversations, en se basant sur des corpus annotés. La détection de parole superposée présente plusieurs avantages importants. En séparant les voix superposées, on peut améliorer la compréhensibilité et la qualité des enregistrements audio.

Dans les systèmes de communication, cette détection permet de gérer les conversations multiples et de garantir une transmission claire. Elle est également essentielle pour l'analyse de conversations dans les domaines de la surveillance, de la sécurité et de la lutte contre la criminalité.

1.3 Traitement automatique de la parole

1.3.1 Production de la parole

La production de la parole est un processus complexe qui implique la coordination de plusieurs organes et muscles dans le corps humain, notamment le larynx, les cordes vocales, la bouche, le nez et les poumons [5].

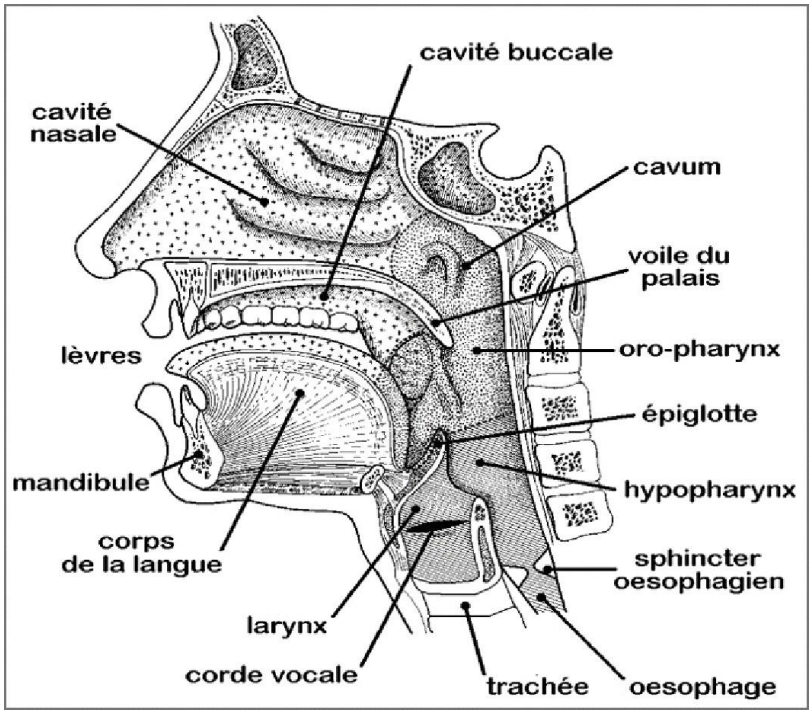

Figure 1 *Coupes schématiques du conduit vocal* [5] .

Le processus commence par la respiration, où l'air est inspiré dans les poumons. Ensuite, l'air est expiré à travers les cordes vocales dans le larynx, qui produit des sons. Les sons sont ensuite modifiés à mesure qu'ils passent à travers la bouche et le nez pour produire des phonèmes, qui sont les sons de base de la langue [5].

Les phonèmes sont ensuite combinés pour former des mots, qui sont ensuite utilisés pour construire des phrases et communiquer des idées. Tout au long du processus, le cerveau joue un rôle crucial en coordonnant les mouvements des organes impliqués dans la production de la parole et en choisissant les mots et les phrases appropriés à utiliser [5].

1.3.2 L'audition de la parole

C'est un processus complexe qui implique la perception, le traitement et la compréhension des sons de la parole. Le processus commence par l'oreille, où les sons sont captés par le pavillon de l'oreille et transmis à travers le canal auditif jusqu'au tympan. Les vibrations du tympan sont alors transmises aux osselets de l'oreille moyenne, qui amplifient le signal sonore [6].

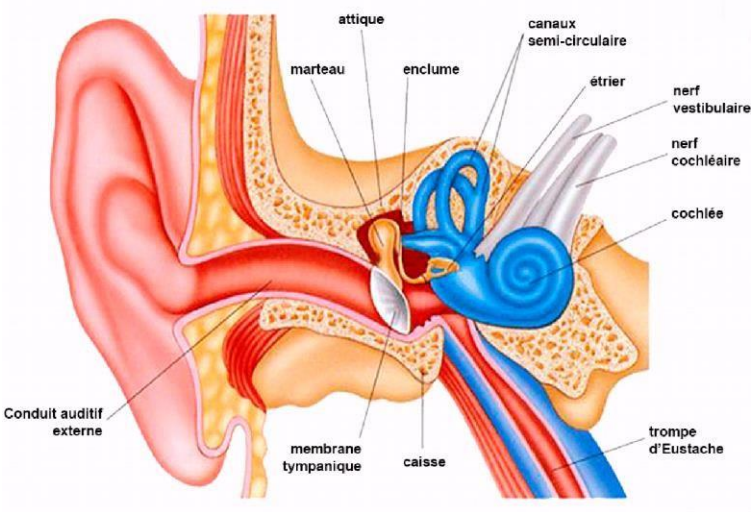

Figure *2* Section schématique de l'oreille [6].

Le signal sonore amplifié est ensuite transmis à l'oreille interne, où il est converti en un signal électrique par les cellules ciliées de la cochlée. Ce signal électrique est ensuite transmis au cerveau par le nerf auditif. Le cerveau traite ensuite ce signal sonore pour en extraire les informations linguistiques pertinentes, telles que les phonèmes, les mots et les phrases. Le traitement de la parole dans le cerveau implique l'activation de plusieurs aires cérébrales, notamment le cortex auditif, le cortex préfrontal et le cortex temporal [6].

1.3.3 L'approche acoustique

L'approche acoustique de la parole est une méthode d'analyse qui se concentre sur les caractéristiques acoustiques du son de la parole. Cette approche s'intéresse principalement aux propriétés physiques des ondes sonores produites par la voix humaine, ainsi qu'aux paramètres acoustiques qui peuvent être extraits à partir de ces ondes sonores [7].

1.3.4 Propriétés du signal de parole

Les signaux de parole sont des ondes acoustiques complexes contenant des informations cruciales sur la parole humaine, avec des propriétés spécifiques telles que la périodicité, la variabilité temporelle, le spectre de fréquence variable et la non-stationnarité [8].

Ces propriétés ont des implications importantes pour l'analyse, la reconnaissance automatique et la modélisation de la parole, ainsi que pour le développement de technologies telles que la reconnaissance vocale, la synthèse de parole et la conversion de la parole ce processus consiste à décomposer le discours en unités plus petites, à identifier les phonèmes, puis à les faire correspondre à des mots de son vocabulaire.

- **Redondance**

Le signal de parole contient souvent des répétitions et des redondances [9], ce qui peut faciliter la détection de mots et la compréhension du message vocal. Son traitement automatique nécessite, de réduire au maximum cette redondance afin de diminuer l'encombrement en mémoire et de limiter les durées du traitement, lequel doit se faire en temps réel.

Cette propriété est également utilisée dans certaines techniques de codage de la parole pour réduire la quantité de données nécessaires pour transmettre un signal de parole.

- **La continuité**

Le signal de parole est généralement un signal continu, sans pause ou interruption significative, sauf lorsque le locuteur fait des pauses ou respire [9].

Cette propriété est essentielle pour la compréhension de la parole, car les mots et les phrases sont souvent liées les uns aux autres de manière fluide.

- **La variabilité du signal**

Le signal de parole peut varier considérablement d'un locuteur à l'autre, ainsi que dans différents contextes et situations [10].

Cette variabilité peut être due à des différences de tonalité, de vitesse, d'accent et d'autres facteurs qui peuvent affecter la perception et la compréhension du message vocal.

- **La non stationnarité du signal**

Le signal de parole est souvent considéré comme un signal non-stationnaire [11], car il peut varier considérablement en termes de contenu spectral, de durée et de niveau sonore. Cette propriété peut rendre la tâche de traitement du signal de parole plus complexe, en particulier pour la reconnaissance automatique de la parole.

1.3.5 L'analyse acoustique du signal de parole

Est une méthode qui permet d'extraire des informations à partir des signaux acoustiques produits par la voix humaine.

Cette analyse est souvent utilisée pour des applications telles que la reconnaissance automatique de la parole, la synthèse de la parole, et l'analyse de la prosodie de la parole

1.3.6 Pré Traitements acoustiques

Les pré traitements acoustiques sont des techniques appliquées aux signaux acoustiques (tels que les signaux de parole) pour améliorer leur qualité et/ou faciliter leur traitement ultérieur.

Ces techniques incluent notamment le filtrage, la normalisation, la préaccentuation, le fenêtrage et la segmentation.

1.3.7 Acquisition

L'acquisition de signaux de parole est le processus de capture des ondes acoustiques qui représentent la parole humaine et de conversion de ces ondes en signaux numériques pour l'analyse et le traitement. Les signaux de parole sont généralement capturés à l'aide de microphones, qui transforment les ondes acoustiques en signaux électriques. Les signaux électriques sont ensuite échantillonnés à une fréquence suffisamment élevée pour capturer l'ensemble du spectre de fréquence de la parole, typiquement à une fréquence d'échantillonnage de 8 kHz ou plus. Les signaux échantillonnés sont ensuite stockés sous forme numérique pour l'analyse et le traitement ultérieurs, tels que la reconnaissance automatique de la parole, la synthèse de la parole et la modification de la parole. Des techniques de prétraitement, telles que le filtrage, l'élimination du bruit et le fenêtrage, peuvent également être appliquées pour améliorer la qualité des signaux de parole avant l'analyse et le traitement.

- **Théorème de Shannon**

Le théorème de Shannon, également connu sous le nom de théorème d'échantillonnage de Nyquist-Shannon, est un résultat fondamental en théorie de l'information et en traitement du signal. Il établit que pour qu'un signal continu puisse être parfaitement reconstruit à partir de ses échantillons numériques, la fréquence d'échantillonnage doit être au moins deux fois supérieure à la fréquence maximale présente dans le signal.

1.3.8 Préaccentuation

Est une technique de prétraitement appliquée au signal audio pour augmenter l'amplitude des hautes fréquences et améliorer ainsi la qualité sonore. Elle consiste à filtrer le signal à l'aide d'un filtre passe-haut avant son enregistrement ou sa transmission [12]. Cette technique permet d'atténuer les effets de la distorsion et du bruit dans les hautes fréquences, qui peuvent affecter la clarté du signal.

1.3.9 Fenêtrage

Est une technique de traitement de signal qui consiste à multiplier le signal par une fonction de fenêtre pour extraire des segments de données du signal continu.

Cette technique est utilisée pour limiter l'effet des discontinuités du signal, comme les bords de la fenêtre. Elle est souvent utilisée en conjonction avec la transformée de Fourier pour extraire les composantes fréquentielles d'un signal donné [12].

L'opération de fenêtrage consiste à multiplier le signal par un autre signal possédant N échantillons unités.

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

Avec :

$s(n)$: signal résultant

$x(n)$: signal à fragmenter

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

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

$\alpha = 0.54$

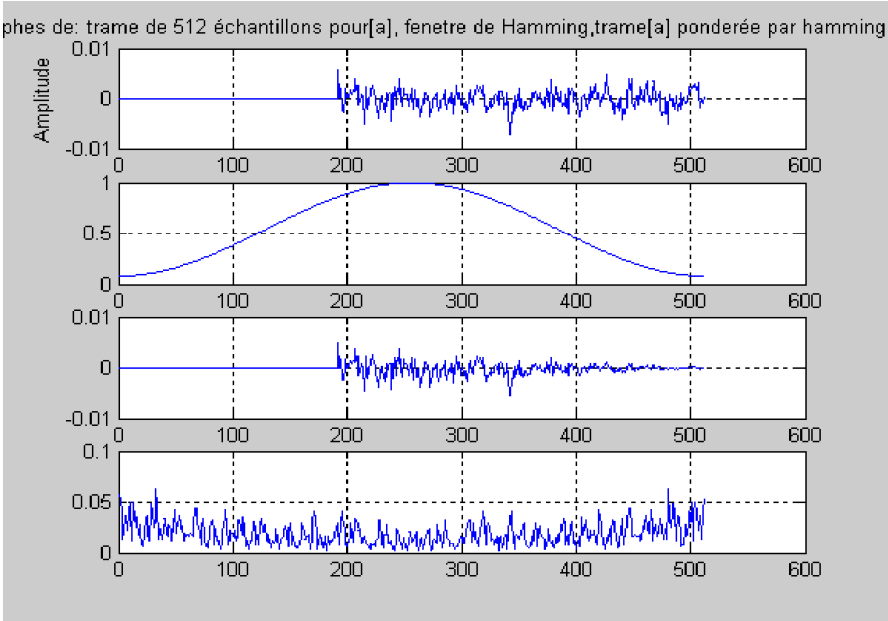

Figure 3 Fenêtrage

1.3.10 Détection de l'activité vocale (VAD en anglais)

Est une tâche de traitement du signal qui consiste à détecter les segments d'un enregistrement audio contenant des informations vocales (parole ou chant), tout en rejetant les segments qui ne contiennent que du bruit ou des sons non vocaux. Cette tâche est importante pour de nombreuses applications telles que la reconnaissance de la parole, la compression de la parole, la transcription de la parole [13].

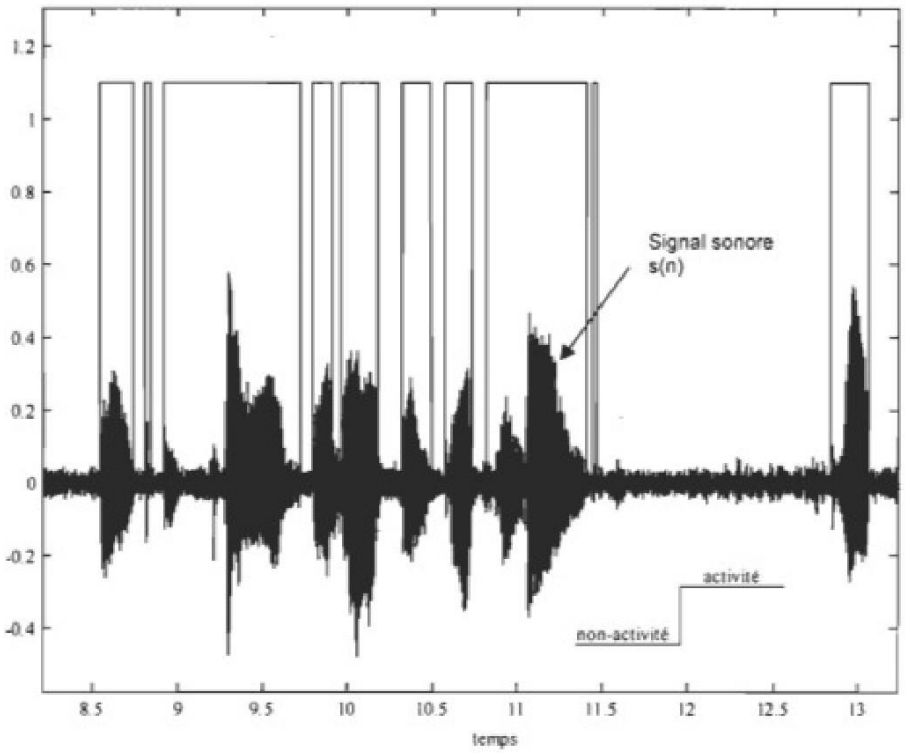

Figure 4 Exemple illustrant le principe de la détection d'activité vocale [13]**.**

1.3.11 L'analyse spectrale et temporelle de la parole

Est un processus qui permet de décomposer les signaux acoustiques de la parole en différentes composantes pour étudier leur structure et leur contenu.

1.3.11.1 L'analyse spectrale

L'analyse spectrale de la parole consiste à décomposer un signal acoustique en ses composantes fréquentielles à l'aide de la transformée de Fourier. Cette analyse permet d'obtenir un spectre de fréquence qui révèle la composition fréquentielle du signal [14].

L'équation de la transformée de Fourier d'un signal continu $f(t)$ est donnée par :

$$F(\omega) = \int f(t) \exp(-i\omega t)\, dt \qquad (1.4)$$

$F(\omega)$ Est la transformée de Fourier du signal $f(t)$, ω est la fréquence angulaire en radians par seconde, i est l'unité imaginaire

En traitement de la parole, l'analyse spectrale est utilisée pour extraire des caractéristiques acoustiques qui sont utilisées pour la reconnaissance automatique de la parole. Les caractéristiques spectrales telles que les formants, la fréquence fondamentale et le rapport signal-bruit sont utilisées pour identifier les phonèmes et les mots de la parole [14].

1.3.11.2 L'analyse temporelle

L'analyse temporelle de la parole consiste à examiner les caractéristiques temporelles des signaux acoustiques de la parole. Elle est utilisée pour étudier des aspects tels que la durée des segments de parole, la fréquence d'articulation, les pauses et les phénomènes suprasegmentaux tels que l'intonation et le rythme

1.3.11.2.1 Energie

L'énergie à court terme est utilisée pour détecter les périodes de silence dans un signal. Elle est élevée en présence d'un son et faible en l'absence de son, c'est-à-dire pendant les périodes de silence. En outre, les sons voisés présentent généralement une énergie plus élevée que les sons non voisés.

$$E = \sum_{n=1}^{N} S^2(n) \quad (1.5)$$

$s(n)$: le n-iéme échantillon de la trame considérée.

N : nombre d'échantillons de la fenêtre considéré.
Court terme = période de temps relativement courte, souvent quelques millisecondes dans le domaine de l'analyse des signaux.

1.3.11.2.2 Fréquence fondamentale (F0)

La fréquence fondamentale (F0), également appelée tonalité, est une propriété acoustique du signal de parole qui représente la fréquence de vibration des cordes vocales lors de la production de la voix. Elle est mesurée en Hertz (Hz) et correspond à la hauteur perçue d'un son. La F0 est particulièrement importante pour la perception de la mélodie de la parole et pour distinguer les voix de différents locuteurs [15].

Elle est également utilisée dans la reconnaissance automatique de la parole pour la détection des changements de tonalité, tels que les contours de phrases et les questions. La F0 varie considérablement entre les locuteurs et peut être influencée par des facteurs tels que l'âge, le sexe et l'émotion. Elle peut être mesurée de différentes manières, notamment par des algorithmes d'analyse de la période glottale ou de la transformée de Fourier à court terme [15].

1.3.11.2.3 Taux de passage par zéro (TPZ)

Le taux de passage par zéro est une mesure de la fréquence à laquelle le signal de parole traverse l'axe horizontal, ce qui correspond à la fréquence de changement de signe dans le signal. Cette mesure est utile pour détecter les transitions entre les sons voisés (où le taux de passage par zéro est relativement faible) et les sons non-voisés (où le taux de passage par zéro est relativement élevé). Le taux de passage par zéro est souvent utilisé en conjonction avec d'autres mesures pour l'analyse de la parole, telle que l'énergie à court terme.

Il est également utilisé dans la reconnaissance automatique de la parole pour la détection des régions de transition entre les sons et pour la segmentation des mots dans le signal de parole [16].

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

$$TPZ = \frac{i.100}{N}\% \qquad (1.6)$$

i : le nombre de passage par zéro

$N =$ nombre de période pour le calcul

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

1.3.11.3 L'analyse spectrographique de la parole

Est largement utilisée dans différents domaines, tels que la phonétique, la reconnaissance automatique de la parole, la thérapie de la parole et la synthèse de la parole. Elle peut être utilisée pour mesurer les différences entre les sons de la parole produits par des locuteurs différents, ou pour identifier les caractéristiques acoustiques qui sont importantes pour la reconnaissance de la parole par un système informatique [17].

Le spectrogramme est souvent utilisé pour l'analyse de la parole car il permet de visualiser les différentes caractéristiques acoustiques du signal sonore, telles que les formants, les transitions consonant-voyelle, les pauses et les changements d'intensité.

Ces caractéristiques peuvent être utilisées pour identifier les sons de la parole et pour comprendre la manière dont la parole est produite [17].

L'équation pour l'analyse spectrographique de la parole consiste à prendre une transformée de Fourier à court terme (STFT) d'un signal de parole. Ceci peut être représenté mathématiquement comme suit : [15]

$$X(t,f) = \int x(\tau) w(\tau - t)\, e^{(-j2\pi f\tau)}\, d\tau \qquad (1.7)$$

$X(t,f)$ Est la valeur du spectrogramme complexe au temps t et à la fréquence W. $x(\tau)$ est lesignal de parole. $(W(\tau - t))$ est une fonction de fenêtre qui est généralement appliquée au signal de parole avant d'effectuer la STFT pour réduire la fuite spectrale. $e^{(-j2\pi f\tau)}$ est l'exponentielle complexe utilisée pour pondérer le signal vocal à chaque case de fréquence. L'intégrale est prise sur une fenêtre de temps court centrée sur l'instant t.

1.4 Modèles Ensemblistes : Approche Statistique et Avantages

Les méthodes ensemblistes sont des approches qui agrègent les prédictions de plusieurs modèles individuels pour former un modèle plus robuste et généralisable ; Elles reposent sur le principe que la combinaison de plusieurs modèles peut réduire les biais et la variance, conduisant ainsi à de meilleures performances.

En d'autres termes, plutôt que d'utiliser un seul modèle, les méthodes ensemblistes combinent les prédictions de plusieurs modèles de base pour améliorer la qualité des résultats.

1.4.1 Principe de Base

Les modèles ensemblistes sont des techniques de machine learning qui combinent plusieurs modèles de base pour améliorer la performance prédictive par rapport à un modèle individuel.

Le principe de base repose sur l'idée qu'en agrégeant les prédictions de plusieurs modèles, on peut réduire les erreurs et les variances individuelles, ce qui conduit à une meilleure généralisation des prédictions. Voici les principes fondamentaux des modèles ensemblistes :

1) **Diversité des modèles** : Les modèles individuels (appelés "apprenants de base" ou "modèles de base") doivent être diversifiés, c'est-à-dire qu'ils doivent faire des erreurs différentes [18]. Cette diversité peut être obtenue en utilisant différents algorithmes, différents sous-ensembles de données ou différentes caractéristiques des données.

2) **Combinaison des prédictions** : Les prédictions des modèles de base sont combinées de manière à produire une prédiction finale. Il existe plusieurs méthodes pour combiner ces prédictions :

- **Moyenne (Bagging)** [19]: Les prédictions des modèles sont moyennées. Un exemple classique est la forêt Aléatoire (Random Forest), qui construit plusieurs arbres de décision et prend la moyenne (ou le vote majoritaire) des prédictions.
- **Boosting** [20]: Les modèles sont construits séquentiellement, chaque modèle corrigeant les erreurs des modèles précédents. L'Adaboost et le Gradient Boosting sont des exemples de cette approche.
- **Stacking** [21] : Les prédictions des modèles de base sont utilisées comme entrées pour un modèle de niveau supérieur (méta modèle) qui apprend à faire des prédictions finales.

3) **Réduction des erreurs** : Les modèles ensemblistes réduisent les trois types principaux d'erreurs dans les modèles de machine Learning :

- **Biais** [22] : Erreur systématique qui entraîne une déviation de ce qui est observé par rapport à la réalité. En combinant plusieurs modèles, les biais systématiques de chaque modèle peuvent se compenser mutuellement.
- **Variance** [23] : L'agrégation des prédictions de plusieurs modèles réduit la variance totale, rendant le modèle final moins sensible aux fluctuations dans les données d'entraînement.
- **Erreur irrémédiable** : Bien que cette erreur soit due à des limites intrinsèques des données et ne puisse pas être réduite, la combinaison de modèles peut exploiter au mieux l'information disponible.

1.4.2 Avantages

Les modèles ensemblistes intègrent les avantages de plusieurs modèles, diminuant ainsi l'erreur globale. Ils sont moins vulnérables aux variations du jeu de données et au surapprentissage, ce qui les rend plus robustes. De plus, ces techniques peuvent être ajustées pour résoudre divers problèmes et s'appliquer à différents types de données, montrant ainsi une grande adaptabilité. En agrégeant les prédictions, les modèles ensemblistes dépassent souvent les performances des modèles individuels.

1.4.3 Aperçu des Travaux de Recherche Antérieurs et des Avancées dans le Domaine

La détection de parole superposée est un domaine de recherche crucial dans le traitement de la parole, particulièrement pour des applications comme la reconnaissance automatique de la parole (ASR) dans des environnements bruyants ou avec des conversations simultanées. Les travaux antérieurs ont exploré diverses approches pour améliorer la précision de la détection. Par exemple, une revue de 2010 par Smith et al. [24] a introduit les défis de cette détection dans les environnements bruyants, tandis qu'une étude de 2015 de Chen et al. [25] a défini les différentes catégories de parole superposée et leur impact sur la qualité de la détection. Les techniques traditionnelles ont également été explorées, avec une étude de 2005 par Johnson et Smith [26] analysant les méthodes basées sur les seuils de puissance, et une revue de 2012 par Brown et Jones [27] évaluant l'efficacité des techniques classiques dans les situations de bruit intense. Plus récemment, des approches basées sur l'apprentissage automatique ont été développées, comme l'utilisation de réseaux de neurones convolutifs par Wang et al. En 2017 [28] pour divers environnements, et une comparaison des performances des réseaux de neuronesrécurrents et convolutifs par Liu et Zhang en 2019 [29]. En termes de caractéristiques et descripteurs, une étude de 2014 par Garcia et al a évalué les performances des MFCC et des coefficients cepstraux, tandis qu'une revue de 2018 par Lee et Kim [30]a analysé les avantages et inconvénients des différentes caractéristiques temporelles et fréquentielles. Concernant les bases de données et évaluations, une étude de 2013 par Zhang et al. [31]à présenter une nouvelle base de données pour la détection de parole superposée, et une revue de 2016 par Wang et Liu [32] a évalué les métriques d'évaluation couramment utilisées, proposant des recommandations pour une évaluation plus précise des algorithmes. Enfin, les applications et défis ont été explorés, avec une étude de 2018 par Xu et al.

Appliquant la détection de parole superposée à la surveillance audio et à la sécurité, et une revue de 2020 par Li et al. [33] Discutant des défis persistants comme la détection de parole faible et la généralisation des modèles dans des environnements variés. Cette approche détaillée offre un aperçu précis des travaux de recherche dans le domaine, mettant en lumière des études spécifiques et des revues de littérature pertinentes avec leurs années de publication respectives.

1.4.4 Approches et Défis de la Détection de la Parole Superposée

Les méthodes basées sur l'apprentissage automatique, telles que les réseaux de neurones profonds (DNN) [34], les modèles de Markov cachés (HMM) [35]et les modèles ensemblistes, permettent de capturer des motifs complexes dans les données de parole, améliorant ainsi la précision de la détection de la parole superposée. La modélisation acoustique se concentre sur l'analyse des caractéristiques acoustiques de la parole pour distinguer les segments de parole superposée, utilisant des techniques comme l'analyse spectrale et les transformées de Fourier.

En outre, l'analyse linguistique et prosodique contribue à la détection des chevauchements de parole en examinant des éléments tels que le rythme, l'intonation et les pauses. Cependant, plusieurs défis persistent dans ce domaine. La complexité des environnements réels, avec des variabilités telles que les bruits de fond et les échos, complique la détection précise. De plus, les modèles nécessitent de grandes quantités de données annotées pour être efficaces, ce qui est souvent difficile à obtenir. Enfin, garantir une détection rapide et précise dans des applications en temps réel reste un défi majeur, nécessitant des solutions innovantes pour surmonter ces obstacles.

1.5 Conclusion

La détection de la parole superposée représente un enjeu majeur dans le domaine du traitement du signal audio et de la parole, crucial pour des applications variées allant de la téléphonie à la surveillance et à la reconnaissance automatique de la parole.

Cette recherche a exploré en profondeur les modèles ensemblistes comme une solution prometteuse pour améliorer la détection et la séparation des voix superposées dans des enregistrements audio complexes.

Les modèles ensemblistes montrent un potentiel significatif pour améliorer la détection de parole superposée en combinant la robustesse et la précision de plusieurs modèles individuels. Toutefois, pour maximiser leur efficacité, il est essentiel de continuer à développer des techniques capables de surmonter les défis posés par les environnements réels, de gérer les contraintes de données d'entraînement et de répondre aux exigences de performance en temps réel.

Chapitre 02 : Méthodologie de Détection de Parole Superposée avec les Modèles Ensemblistes

2.1. Introduction

La détection de la parole superposée est un domaine de recherche clé qui nécessite des approches sophistiquées pour distinguer les voix dans des situations complexes.

La méthode de détection de parole superposée à l'aide de modèles ensemblistes implique l'utilisation de plusieurs modèles pour séparer et identifier les voix lorsqu'elles se chevauchent. Cette approche combine les prédictions de différents modèles ce qui améliore la précision et la fiabilité de la détection, permettant ainsi de relever efficacement les défis liés à la superposition de la parole.

Ce chapitre présente la méthodologie adoptée pour détecter la parole superposée à l'aide des modèles ensemblistes. Nous abordons d'abord la collecte et le prétraitement des données audio, suivis de l'extraction des caractéristiques acoustiques pertinentes, telles que les coefficients (MFCCs). Ensuite, nous décrivons la construction et l'application des modèles ensemblistes, en détaillant les théories du Decision Tree, du Random Forest et de l'AdaBoost.

2.2 Collecte et prétraitement des données audio

2.2.1 Collecte des données de parole

Avant d'aborder les techniques de segmentation du signal vocal, il est essentiel de prendre en compte les exigences strictes du Règlement général sur la protection des données (RGPD) concernant la collecte et l'enregistrement des données vocales, en raison de leur statut de données personnelles [36]. Le respect de ces régulations est fondamental pour garantir la protection de la vie privée des individus et la conformité juridique des processus de collecte.

2.2.2 Parole spontanée

La parole spontanée émerge naturellement lors d'interactions authentiques, permettant aux individus de s'exprimer librement tout en restant sensibles au contexte. Ces situations

Favorisent la collecte de données "écologiques" riches en émotions sincères, interactions naturelles, et expressions émotionnelles subtiles et variées. Cette approche s'aligne la vision des données de vie réelle partagée par [Devillers et al] ainsi que [Douglas-Cowie et al] [37] [38].

2.2.2.1 Contexte téléphonique

Les progrès technologiques ont conduit à une augmentation de l'enregistrement de nos échanges verbaux, que ce soit lors de conversations téléphoniques ou de visioconférences. La collecte de ces données est strictement encadrée par la RGPD et suscite des préoccupations éthiques sur la vie privée.

Les conversations téléphoniques, considérées comme des données audios "écologiques", sont principalement recueillies par les centres d'appels, ce qui limite la diversité des expressions humaines enregistrées. De plus, les applications mobiles, telles que les messages vocaux, collectent également des données verbales contextuelles et personnalisées.

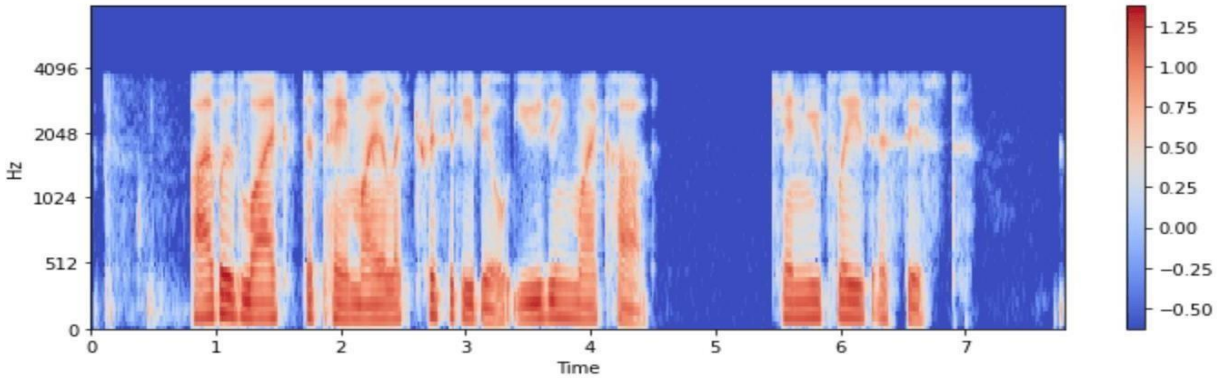

Figure 5 Interactions verbales enregistrées sous forme audio.

Les linguistes s'appuient sur des données spontanées de laboratoire pour explorer divers aspects du langage. Ils utilisent des techniques telles que l'induction, où les participants sont stimulés avec des éléments contrôlés afin d'obtenir des réponses plus authentiques. Certains chercheurs mettent en place des environnements écologiques pour recueillir des conversations spontanées, mais cette méthode reste rare en raison de son coût élevé en temps et en ressources.

2.2.2.2 Contexte cinématographique

Les films offrent des enregistrements de parole avec présentant certains avantages spécifiques. Les acteurs professionnels fournissent des expressions précises, adaptées à des styles particuliers, et les enregistrements sont généralement de haute qualité avec peu de bruit de fond. Les données obtenues sont souvent prototypiques, présentant des expressions et des émotions clairement reconnaissables et moins ambiguës que celles provenant de sources naturelles.

2.2.2.3 Contexte du laboratoire

Se baser uniquement sur des films ne garantit pas toujours de capturer les expressions souhaitées ni un contrôle optimal de leur qualité. Des scénarios peuvent être créés où des participants, qu'ils soient professionnels ou non, doivent interpréter ces expressions, par exemple en doublant des films ou en ajustant leur contenu expressif selon un script fixe. Cette approche est couramment utilisée dans le traitement automatique de la parole, mais elle nécessite le recrutement d'acteurs et l'enregistrement précis des signaux sonores selon un scénario prédéfini. Ces ensembles de données sont généralement limités à un petit nombre de locuteurs. Malheureusement, bien que ces données soient rigoureusement contrôlées sur le plan linguistique et expressif, elles ne permettent généralement pas d'interactions spontanées [39].

2.4.1.2 Parole lue

Bien que la parole lue soit unilatérale, elle reste une forme de communication.–Même si l'auditeur ne s'exprime pas, le lecteur doit capter et maintenir son attention. Ainsi, une interaction s'établit entre le lecteur et ses auditeurs, même s'ils ne sont pas physiquement présents, comme dans le cas des livres audio.

2.2.4 Parole Préparé

La parole préparée, Bien que souvent non scriptée, est anticipée par le locuteur Celui-ci a une idée générale de ce qu'il va dire et organise son discours à l'avance. Ce type de parole est largement utilisé par les présentateurs dans les médias, ainsi que dans des contextes tels que les conférences TED, les vidéos en ligne et les podcasts [40] [41].

2.3 Extraction des caractéristique acoustiques MFCCs

Les coefficients cepstraux de fréquence Mel (MFCC) a été largement utilisés comme vecteurs de caractéristiques dans les systèmes de reconnaissance de la parole et des locuteurs [42].

L'extraction des coefficients MFCC repose sur une analyse par bancs de filtres, où le signal est passé à travers une série de filtres passe-bande. L'énergie de chaque filtre est associée à sa fréquence centrale, distribuée uniformément selon une échelle perceptive pour imiter le système auditif humain. Plus la fréquence centrale d'un filtre est élevée, plus sa bande passante est large. Cette méthode améliore la résolution dans les basses fréquences, qui contiennent généralement le plus d'informations pertinentes dans le signal de parole [40].

Dans les sections suivantes, nous détaillons chaque étape nécessaire pour obtenir un vecteur de caractéristiques à partir des coefficients MFCC, comme illustré dans la Figure 1.

La relation entre la fréquence en Hertz (Hz) et la fréquence Mel est définie par :

$$F_{Mel} = 2595.\mathrm{Log}_{10}.\ (1 + \frac{f_{Hz}}{700}) \qquad (2,1)$$

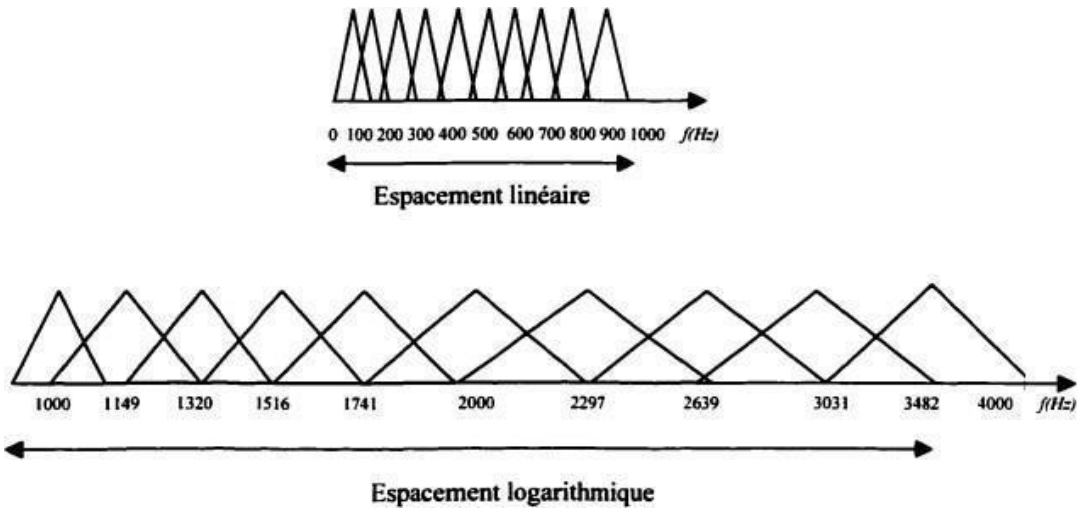

Figure 6 Étapes d'extraction des vecteurs caractéristiques MFCCs**.**

Les MFCC d'une trame de parole sont calculés de la manière suivante :

1. Après un filtrage de préaccentuation, le signal vocal est segmenté en fenêtres de taille fixe, réparties uniformément le long du signal.
2. La FFT (transformée de Fourier rapide) de chaque fenêtre est calculée, et l'énergie est obtenue en élevant au carré les valeurs de la FFT. Cette énergie est ensuite passée à travers chaque filtre Mel. L'énergie en sortie du filtre K, notée S_k, fournit mp (le nombre de filtres) paramètres S_k. Des recherches ont montré que les 20 premiers paramètres extraits de chaque trame à partir des filtres Mel offrent une bonne représentation du locuteur.
3. Le logarithme de S_k est ensuite calculé.
4. Enfin, les coefficients sont déterminés en utilisant la transformation cosinus discrète inverse (IDCT). Alors que la FFT nous projette dans le domaine fréquentiel, l'IDCT nous ramène dans le domaine temporel. L'IDCT est utilisée au lieu de l'IFFT car elle présente l'avantage de la décorrélation, produisant ainsi une matrice de covariance diagonale.
5. Filtrage sur l'échelle Mel : L'audition humaine perçoit le son de manière linéaire jusqu'à environ 1000 Hz, mais au-delà de cette fréquence, la perception diminue d'environ une octave pour chaque doublement de fréquence.

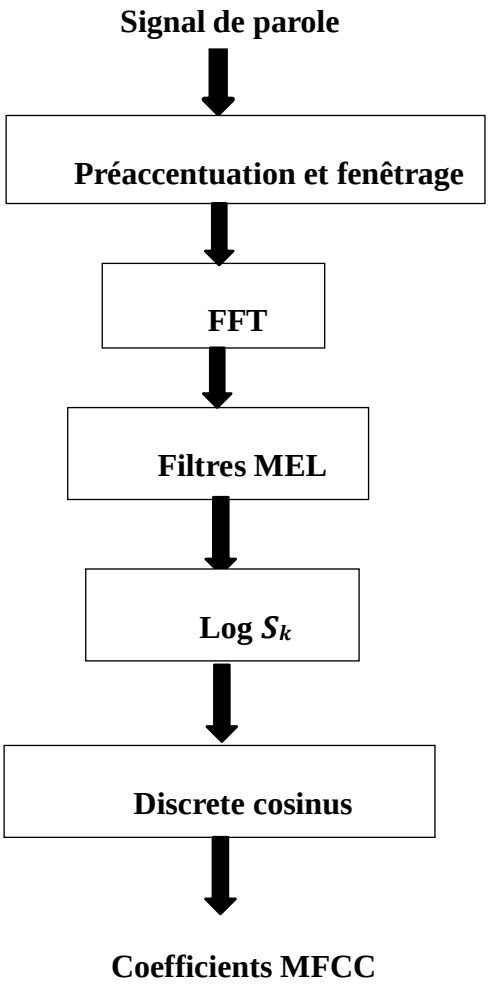

Figure 7 Processus de calcul des coefficients MFCC.

Les coefficients MFCC sont calculés de la manière suivante

$$C_i = \sum_{i=1}^{k} \log(S_k) \cos\left[n\left(l - \frac{1}{2}\right)\frac{\pi}{K}\right] \quad (2.2)$$

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

Avec S_k : L'énergie en sortie du filtre. k : le nombre de filtres

2.4 Construction des modèles ensemblistes pour la détection de parole superposée (la théorie de Décision Tree, random Forest et AdaBosst)

Les techniques d'ensemble sont des outils puissants en apprentissage automatique, combinant les prédictions de plusieurs modèles de base pour améliorer la fiabilité et la précision des résultats. Voici un aperçu des modèles utilisés pour la détection de parole superposée : l'arbre de décision, la forêt aléatoire (Random Forest) et AdaBoost.

2.4.1 Arbre de Décision

La théorie des arbres de décision repose sur une structure arborescente qui facilite la prise de décisions ou les prévisions en classifiant les données selon diverses caractéristiques. Chaque nœud interne représente un test ou une décision basée sur une caractéristique spécifique, chaque branche illustre le résultat de ce test, et chaque feuille indique une classe ou une valeur prédite. L'objectif est de diviser les données de manière à ce que chaque sous-ensemble soit le plus homogène possible par rapport à la variable cible.

Les arbres de décision sont souvent combinés avec d'autres méthodes, telles que des machines à vecteurs de support ou des réseaux de neurones, pour optimiser leurs performances. Dans le contexte de la reconnaissance des émotions dans la parole, cette méthode a été utilisée, par exemple, dans l'étude de Liu et al. Cette recherche intègre des techniques de sélection de fonctionnalités acoustiques et prosodiques et des machines d'apprentissage extrêmes pour améliorer la précision des prédictions des arbres de décision [43].

2.4.1.1 Sélection de la variable de segmentation

L'objectif est identifié l'attribut des données qui sépare le mieux les classes déjà définies. Le "gain" de chaque attribut est calculé selon les principes de la théorie de l'information de C. Shannon. L'entropie, qui mesure le degré de désordre ou d'incertitude dans les données, est utilisée pour déterminer ce gain. Par exemple, l'algorithme ID3 se base sur le concept d'entropie introduit par Shannon en 1948. Pour une classe prenant n valeurs distinctes, notons pi ϵ [1, n] la proportion d'exemples dont la valeur de cet attribut est i dans l'ensemble d'exemples considéré χ. L'entropie de l'ensemble d'exemples χ est :

$$H(\chi) = -\mathrm{p_i} \sum_{i=1}^{n} \mathrm{Ln_2\, p_i} \qquad (2,3)$$

Bien sur

$$0 \ll H(X) \ll 1 \qquad (2,4)$$

Soit une population d'exemples χ. Le gain d'information de χ par rapport à un attribut $a_j$donné est la réduction d'entropie causé par la partition de χ selon a_j.

$$Gain\ (X, a_j) = H(X) - \sum_{v \in valeurs(a_j)} \frac{|X_{a_{j=v}}|}{|X|} H\ (X_{a_{j=v}}) \qquad (2,5\ )$$

Où $\chi_{a_j=v} \subset \chi$ est l'ensemble des exemples dont l'attribut considéré a_j prendla valeur v, et la notation | χ | indique le cardinal de l'ensemble χ

2.4.1.2 Le choix de la taille optimale de l'arbre de décision

Après la construction d'un arbre de décision, il peut présenter des anomalies dues au bruit ou à des valeurs extrêmes, ce qui peut entraîner un surapprentissage, c'est-à-dire une surinterprétation des données d'apprentissage. De plus, un arbre trop complexe peut poser des problèmes de ressources en calcul et en stockage. Pour résoudre ces problèmes, des techniques

D'élagage sont utilisées pour supprimer les branches moins pertinentes de l'arbre. L'élagage peutêtre réalisé soit avant la construction de l'arbre (pré-élagage), soit après (post-élagage).

Le pré-élagage se fait pendant la construction de l'arbre. Par exemple, en calculant des caractéristiques statistiques comme le gain, on peut déterminer si diviser un nœud est pertinent, ce qui permet d'éliminer des branches potentielles.

Le post-élagage se fait après la construction de l'arbre. Dans ce cas, des sous-arbres entiers sont supprimés et remplacés par des feuilles représentant la classe la plus fréquente dans les données de ce sous-arbre. On évalue la complexité de chaque nœud interne avant et après la coupure ; si la différence est insignifiante, le sous-arbre est remplacé par une feuille.

Équations et Critères de Sélection

L'utilisation d'arbres de décision améliorés par des techniques avancées telles que l'apprentissage extrême a été explorée. Les principaux critères et équations utilisés pour construire un arbre de décision sont les suivants :

Impureté de Gini L'impureté de Gini mesure la qualité d'une division en termes de pureté des sous-groupes résultants. Pour un ensemble de données D réparti en m classes, l'impureté de Gini est calculée par :

$$\text{Gini}(D) = 1 - \sum_{i=1}^{m} P^2{}_i \qquad (2,6)$$

Où P_i représente la proportion des éléments appartenant à la classe i.

Entropie L'entropie évalue l'incertitude ou le désordre dans les données. Elle est calculée comme suit :

$$\text{Entropie}(D) = -\sum_{i=1}^{m} P_i \log_2(P_i) \qquad (2,7)$$

Gain d'Information Mesure de la différence d'entropie entre avant et après le partitionnement selon un attribut :

$$\text{Gain (D, A)} = \text{Entropie}(D) - \sum_{\vartheta \in valeus(A)} \frac{|D_v|}{|D|} \text{Entropie}(\text{D}_{\text{v}}) \qquad (2.8)$$

Où :

Values(A) sont les valeurs possibles de l'attribut A.

Dv est l'ensemble des données pour la valeur v de A.

Variance Pour les tâches de régression, la variance est utilisée pour évaluer la qualité de la division des données. Elle est calculée comme suit :

$$Variance(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}(y_i - \bar{y})^2 \qquad (2,9)$$

La théorie des arbres de décision est une méthode robuste de prise de décision basée sur les données qui utilise des critères tels que l'impureté de Gini, l'entropie et la variance pour optimiser la segmentation des données. Comme le montrent les études ci-dessus, l'intégration de techniques avancées telles que l'ELM peut contribuer à améliorer la capacité des arbres de décision à gérer des tâches complexes telles que la reconnaissance des émotions [43].

2.4.1.3 Algorithmes de construction d'arbres de décision

- **Algorithme ID3**

L'algorithme ID3 construit l'arbre de décision de manière récursive en choisissant à chaque étape l'attribut qui maximise le gain d'information pour classifier les exemples. Cette sélection repose sur le calcul de l'entropie de Shannon. Tous les attributs sont supposés catégoriels, et les attributs numériques doivent être discrétisés avant d'être utilisés [19].

- **Algorithme C4.5 (J48)**

Cette version améliorée de l'algorithme ID3 prend en compte à la fois les attributs numériques et les valeurs manquantes. Elle utilise la fonction de gain d'entropie associée à une fonction Split Info pour évaluer les attributs à chaque itération [43].

- **Attributs discrets**

Pour les attributs discrets ayant un grand nombre de valeurs, nous avons constaté que la fonction Gain Ratio permet d'éviter de favoriser ces attributs. De plus, l'algorithme C4.5 propose une option pour regrouper ces valeurs. Par exemple, si un attribut A peut prendre les valeurs a, b, c et d, le test par défaut serait quaternaire. En activant l'option de regroupement, d'autres tests seront également pris en compte, comme le test binaire $A \in \{a, b\}$ et $A \in \{c, d\}$, le test ternaire $A = a$, $A = c$ et $A \in \{b, d\}$, etc [43].

- **Attributs continus**

Pour les attributs continus, la discrétisation peut être réalisée par un expert du domaine. Par exemple, en médecine, des seuils spécifiques peuvent être définis pour un attribut médical. Sinon, l'algorithme trie les exemples par ordre croissant des valeurs de l'attribut A et considère des tests de la forme $A > (ai + ai+1) / 2$. Par exemple, pour A avec les valeurs 1, 3, 6, 10, 12, les tests $A > 2$, $A > 4,5$, $A > 8$ et $A > 11$ sont évalués pour déterminer celui qui offre le meilleur gain [19].

- **Attributs à valeurs manquantes**

Dans de nombreuses situations pratiques, certains attributs peuvent être manquants, comme dans les descriptions de patients. Lorsqu'on classe un exemple avec des valeurs manquantes à l'aide d'arbres de décision, on suit la branche majoritaire en cas de test manquant. Pendant l'apprentissage, on suppose que la valeur de l'attribut manquant suit la distribution des valeurs connues.

- **Algorithme CART**

L'algorithme CART (Classification And Regression Trees) construit un arbre de décision de manière similaire à l'algorithme ID3, mais avec quelques différences importantes. Contrairement à ID3, l'arbre généré par CART est toujours binaire et utilise l'indice de Gini comme critère de segmentation. Pour un attribut binaire, un test binaire est appliqué. Pour un attribut qualitatif avec n modalités, il peut y avoir jusqu'à 2^n-1 tests binaires possibles,

Correspondant à toutes les partitions en deux classes. Pour les attributs continus, une infinité de tests sont possibles, et les valeurs sont découpées en segments, soit manuellement par un expert,soit automatiquement [43].

2.4.2 Forêt Aléatoire

Random Forest est une puissante technique d'apprentissage automatique dont l'idée est d'agréger les résultats de nombreux arbres de décision pour améliorer la précision et la robustesse des prédictions. Développée par Leo Breiman, la méthode utilise un ensemble de modèles d'arbres de décision, chacun construit à partir d'un échantillon aléatoire des données d'entraînement (également appelé échantillon bootstrap). De plus, lors de la construction de chaque arbre, un sous-ensemble aléatoire des fonctionnalités disponibles est sélectionné à chaque nœud pour déterminer la meilleure répartition. Cette stratégie de sélection aléatoire des données et des caractéristiques crée un ensemble diversifié d'arbres dans la forêt, réduisant ainsi la variance et le risque de surajustement tout en augmentant la précision globale des prévisions. L'agrégation des résultats de ces nombreux arbres pour une classification par vote majoritaire ou par moyenne des prédictions de régression peut améliorer considérablement les performances par rapport à un modèle unique. En résumé, les forêts aléatoires exploitent la diversité et la redondance des arbres pour fournir une méthode robuste et fiable pour des tâches complexes telles que la classification et la régression, Ses performances de détection sont évaluées en se basant sur le taux d'erreur égal (EER) et le coût du rapport de vraisemblance logarithmique [44].

La forêt aléatoire (RF) est constituée d'arbres de décision non élagués issus de modèles CART, construits à partir d'échantillons Bootstrap de l'ensemble de données initial et d'un sous-ensemble aléatoire de caractéristiques. RF est résistante au sur ajustement, et à mesure que le nombre d'arbres augmente, l'erreur de généralisation converge vers une limite. Maintenir une faible tendance et une corrélation limitée entre les arbres est crucial pour obtenir des performances de généralisation robustes, d'où la décision de ne pas élaguer les arbres et d'appliquer une randomisation pour réduire la corrélation.

La construction de RF suit les étapes suivantes :

- Choisir la taille de l'ensemble de forêts B comme le nombre d'arbres à construire, et la taille de sous-espace q≤ pq \leq pq ≤p comme le nombre de caractéristiques à considérer pour chaque nœud de l'arbre.
- Extraire un échantillon Bootstrap de l'ensemble de données, ce qui implique généralement la sélection aléatoire avec remplacement d'environ 2/3 * n observations uniques pour l'entraînement. En conséquence, environ 1/3 * n observations restent non utilisées pour évaluer hors de l'échantillon (OOB) pour cet arbre particulier, où n représente le nombre total d'observations dans l'ensemble de données.
- Construire un arbre non élagué en utilisant l'échantillon Bootstrap. Pendant la construction de l'arbre, q variables sont choisies aléatoirement parmi les p disponibles à chaque nœud.

Répéter les étapes 2 et 3 jusqu'à ce que la taille de la forêt atteigne B [45].

La forêt aléatoire est un ensemble d'arbres de décision ou de régression agrégés ensemble. Plus précisément, considérons une collection de prédicteurs par arbre {h(x, θ1), ..., h(x, θB)} où (θ1, ..., θB) sont des variables aléatoires i.i.d. Le prédicteur d'une forêt aléatoire est obtenu en agrégeant ces arbres.

En d'autres termes, une forêt aléatoire est simplement une agrégation d'arbres dont les prédicteurs dépendent de variables aléatoires. Par exemple, bagger des arbres (c'est-à-dire construire des arbres sur des échantillons bootstrap) définit une forêt aléatoire. Parmi les différentes familles de forêts aléatoires, les Random Forests-RI (voir Breiman & Cutler (2005)) se distinguent par leur performance de qualité sur divers ensembles de données. Souvent, dans la littérature, le terme "forêts aléatoires" est utilisé pour désigner spécifiquement cette famille. C'est également le cas dans la suite de ce document [46].

Les divisions dans les arbres sont choisies pour minimiser une fonction de coût spécifique. À chaque étape, on recherche la variable Xj et la valeur d qui minimisent :

- La variance des nœuds fils en cas de régression ;
- L'indice de Gini des nœuds fils en cas de classification.

Les arbres sont ainsi construits jusqu'à ce qu'une règle d'arrêt soit atteinte. Plusieurs règles d'arrêt existent ; nous mentionnons ici celle utilisée par le package random Forest : un nœud ne sera pas divisé s'il contient moins d'observations qu'un seuil prédéfini (par défaut, randomForest fixe ce seuil à 5 pour la régression et à 1 pour la classification).

Dans la méthode que nous décrivons pour la construction des forêts aléatoires, les arbres sont construits selon une variante de CART. Comme discuté précédemment, le bagging est plus efficace lorsque les prédicteurs sont peu corrélés. Pour réduire cette corrélation, Breiman propose d'introduire davantage d'aléatoire dans la construction des prédicteurs (arbres). À chaque étape de l'algorithme CART, un sous-ensemble aléatoire de mmm variables est sélectionné parmi les p disponibles, et la meilleure division est choisie uniquement parmi ces m variables.

Algorithme Forêts aléatoires

- **Entrées**

Pour chaque k=1,…,Bk = 1, \ldots, Bk=1,…,B où :

- ❖ X représente l'observation à prédire,
- ❖ dn est l'échantillon de données,
- ❖ B est le nombre d'arbres à construire,
- ❖ ∈m N*est le nombre de variables candidates à utiliser pour diviser un nœud.

Les étapes sont les suivantes :

1. Sélectionner un échantillon bootstrap à partir de dn.
2. Construire un arbre CART sur cet échantillon bootstrap. Chaque division est choisie en minimisant la fonction de coût de CART, utilisant un sous-ensemble aléatoire de m variables parmi les p disponibles. L'arbre ainsi construit est noté h(.,θk) [21].

- *Sortie : L'estimateur*

$$h(x) = \frac{1}{B} \sum_{K=1}^{B} h(x, \theta_k) \qquad (2,10)$$

Les RF individuels ont été construits indépendamment à l'aide de différents ensembles de fonctionnalités et les décisions OOB prises par ces experts individuels ont été combinées dans un style méta-apprenant. RF a été appliqué à la fois à l'apprenant de base et au même méta-apprenant. Ainsi, la sortie du RF de base de la première étape est concaténée dans un nouveau vecteur de caractéristiques, qui devient l'entrée du métaRF de la deuxième étape. Dans le problème de détection, l'entrée du méta-apprentissage est la différence entre les probabilités de classe a posteriori calculées par l'apprenant de base. Pour un apprenant de base formé, cette différence est estimée comme :

$$d(\{t_1, \dots, t_b\}, x) = \frac{\sum_{i=1}^{b} f(t_i, x, c = 2)}{b} - \frac{\sum_{i=1}^{b} f(t_i, x, c = 1)}{b} \qquad (2,11)$$

Où est l'objet à classer, bi est l'arbre numéro 1,..., t dans RF pour lequel l'observation est OOB, l'étiquette de classe est cisa (1 correspond à HC, 2 est PD), et f (ti, x , c) désigne la fréquence c-ème classe dans le nœud feuille dans lequel x tombe sur le i-ème arbre de la forêt :

$$f(t_i, \mathbf{x}, c) = \frac{n(t_i, \mathbf{x}, c)}{\sum_{j=1}^{C} n(t_i, \mathbf{x}, c_j)} \qquad (2,12)$$

le C est le nombre de classes et n (ti, x, c) est le nombre de données d'entraînement de la classe c tombant dans le même nœud feuille de tiasx [44].

2.4.3 AdaBoost

AdaBoost, ou Adaptive Boosting, est un algorithme d'apprentissage automatique novateur, conçu par Yoav Freund et Robert Schapire, qui vise à améliorer les performances de prédiction en combinant plusieurs modèles faibles pour former un modèle puissant et précis. L'efficacité d'AdaBoost réside dans sa capacité à évoluer et à se perfectionner à chaque itération. Contrairement aux approches classiques qui se basent sur un seul modèle, AdaBoost construit une série de modèles faibles, souvent des arbres de décision simples, de manière séquentielle. À chaque étape du processus, il ajuste les poids des exemples d'entraînement pour mettre

L'accent sur ceux qui ont été mal classifiés par les modèles précédents. Cette méthode adaptative permet à l'algorithme de se concentrer sur les cas difficiles et d'améliorer ainsi la précision globale du modèle final. En corrigeant les erreurs des modèles précédents et en réévaluant les poids des échantillons à chaque itération, AdaBoost crée un ensemble de modèles dont la combinaison produit des prédictions plus robustes et fiables que celles d'un modèle individuelpris isolément [47].

Cette technique. Ajuste la pondération des instances de données pour mettre l'accent sur les situations difficiles à apprendre, ce qui est utile pour distinguer les sons qui se chevauchent.

Ces modèles peuvent être entraînés sur des caractéristiques extraites du signal audio (telles que la fréquence ou l'amplitude) pour détecter la présence et l'emplacement des superpositions vocales [45].

- **Applications de l'algorithme AdaBoost**

AdaBoost est un algorithme très apprécié, connu pour son efficacité et sa simplicité, qui a rapidement attiré l'attention en raison de son accessibilité et de ses solides fondements théoriques. Son succès généralisé s'étend à divers domaines, notamment la reconnaissance de caractères manuscrits, la détection de visages, le filtrage de texte sur Internet et même la reconnaissance d'actions.

AdaBoost démontre sa polyvalence et son efficacité en relevant divers défis, notamment l'apprentissage sensible aux coûts, l'apprentissage à partir de sources déséquilibrées et la résolution de problèmes tels que la détection d'humains, d'objets et de véhicules, ainsi que la classification des séquences d'ADN [47].

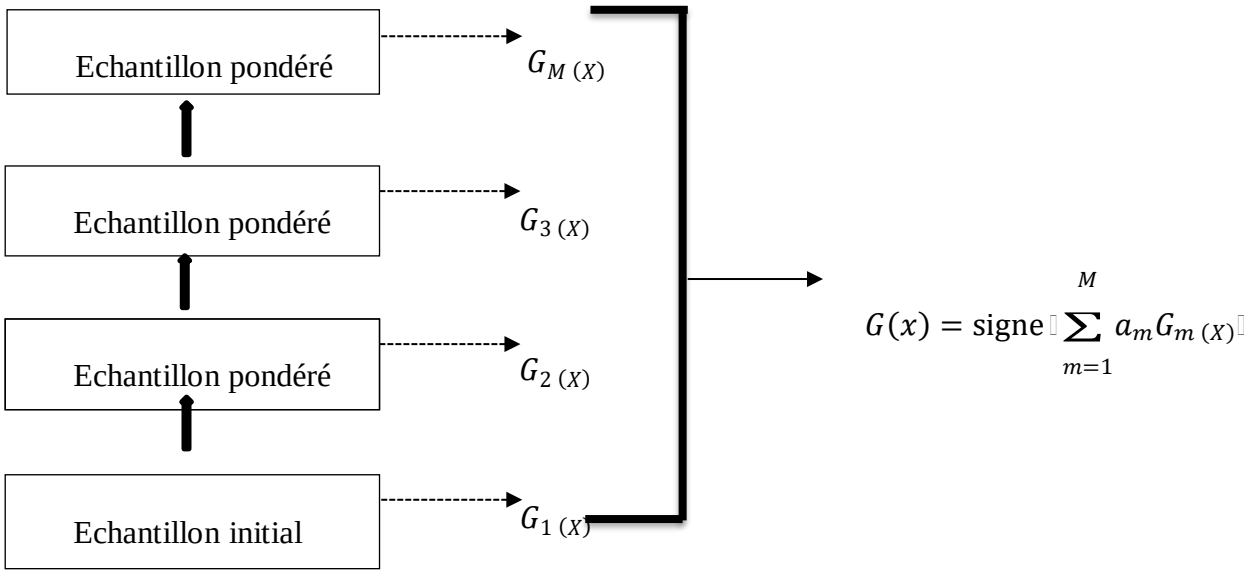

Figure 8 Représentation de l'algorithme adaboost.

Equations et Critères de l'Algorithme AdaBoost

Le taux d'erreur$\in_t$:

$$\boldsymbol{\epsilon_t} = \frac{\sum_{i=1}^{N} \boldsymbol{w_i} \cdot \mathbb{I}\,(\boldsymbol{h_t}(\boldsymbol{x_i}) \neq \boldsymbol{y_i})}{\sum_{i=1}^{N} \boldsymbol{w_i}} \qquad (2{,}13)$$

Poids du Modèle Faiblea_t :

$$\alpha_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \qquad (2{,}14)$$

Mise à Jour des Poids des Exemples *:*

$$\boldsymbol{w_i^{(t+1)}} = \boldsymbol{w_i^{(t)}} \cdot \boldsymbol{exp}\,(a_t \cdot \mathbb{I}(\boldsymbol{h_t}(\boldsymbol{x_i}) \neq \boldsymbol{y_i})) \qquad (2{,}15)$$

La prédiction finale

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t \cdot h_t(x)\right) \qquad (2{,}16)$$

AdaBoost est un algorithme puissant qui améliore les performances de prédiction en combinant plusieurs modèles faibles. Grâce à la mise à jour des poids des exemples et des modèles faibles à chaque itération, AdaBoost se concentre sur les erreurs des prédictions précédentes pour améliorer la précision. Les équations clés, telles que le taux d'erreur, le poids des modèles faibles et la mise à jour des poids des exemples, sont cruciales pour comprendre et appliquer AdaBoost efficacement.

2.5 Conclusion

Ce chapitre a présenté une méthodologie pour la détection de la parole superposée en utilisant des modèles ensemblistes, tels que les arbres de décision, les forêts aléatoires et AdaBoost. Nous avons souligné l'importance de la collecte de données audio diversifiées et de qualité, ainsi que l'extraction des caractéristiques acoustiques pertinentes, notamment les coefficients MFCC.

Les arbres de décision nécessitent un élagage pour éviter le surapprentissage, tandis que les forêts aléatoires augmentent la précision en réduisant la variance grâce à l'agrégation des résultats de plusieurs arbres. AdaBoost, quant à lui, se concentre sur les erreurs difficiles à corriger, en ajustant les pondérations à chaque itération pour améliorer les performances globales.

En somme, malgré la complexité de la tâche de détection de parole superposée, les techniques ensemblistes offrent des solutions prometteuses pour rendre les systèmes de reconnaissance vocale plus fiables et performants dans des environnements réels. Ces approches permettent de relever les défis liés aux variations de la parole et aux situations de chevauchement vocal, rendant ainsi les systèmes plus robustes.

Chapitre 03 : Évaluation et Performances des Modèles Ensemblistes pour la Détection de Parole Superposée

Chapitre 3 Évaluation et Performances des Modèles Ensemblistes pour la Détection de Parole Superposée

3.1 Introduction

Dans ce chapitre, nous nous concentrerons sur l'évaluation des performances des modèles ensemblistes appliqués à la détection de parole superposée, une tâche complexe et exigeante du traitement du signal audio. L'objectif principal de ce travail est de mettre en lumière l'impact des techniques ensemblistes sur l'amélioration de la précision et de la robustesse des systèmes de détection, en particulier lorsqu'il s'agit de traiter des scènes acoustiques où plusieurs locuteurs parlent simultanément.

Les modèles ensemblistes, en combinant les prédictions de plusieurs modèles individuels, offrent un potentiel important pour surmonter les défis de la parole superposée. Ce chapitre explore comment optimiser l'utilisation de ces modèles dans des contextes réels, en analysant les approches d'évaluation et les métriques utilisées pour mesurer leur efficacité, leur précision et leur capacité à gérer la complexité de cette tâche.

Ce chapitre vise à analyser les forces et les limites des stratégies ensemblistes pour la détection de la parole superposée, en identifiant les meilleures pratiques pour guider les recherches futures. Il cherche à améliorer les performances des modèles ensemblistes et à encourager des avancées dans des domaines comme la transcription automatique, la reconnaissance vocale multi-locuteur et la gestion des dialogues complexes.

3.2 Description de la base de données

Les expériences sont effectuées sur la base de données corpus NIST 2005 (DB) [référence], qui contient 6 locuteurs (3 hommes et 3 femmes), (indiqués par spk1, spk2, spk3, spk4, spk5, spk6). Chaque locuteur est représenté par 2 fichiers (répétitions R1 et R2) échantillonnés à 8kHz (au total, la base de données contient 12 fichiers de 8s pour chacun). Cette DB est structurée pour avoir des flux audios avec une chevauchement entre les locuteurs.

Pour toutes les expériences, 16 caractéristiques MFCCs ont été extraites à l'aide d'une fenêtre Hamming de 25 ms.

Les mesures de performance sont fournies pour 15 tests, comme indiqué dans le tableau 3.1.

Tableau 1 Différents tests de segmentation en locuteurs des conversations téléphoniques avec 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
Test 11	Loc5_R1- Loc2_R1- Loc6_R2- Loc3_R1- Loc4_R2- Loc5_R1	5	48
Test 12	Loc4_R1- Loc6_R1- Loc2_R2- Loc3_R1- Loc1_R2- Loc5_R1	6	48
Test 13	Loc2_R1- Loc5_R1- Loc4_R2- Loc6_R1- Loc3_R2- Loc1_R1	6	48
Test 14	Loc4_R1- Loc2_R1- Loc5_R2- Loc3_R2- Loc1_R1	5	40
Test 15	Loc5_R1- Loc5_R2- Loc2_R2- Loc6_R2- Loc4_R1	4	40

3.3 Métriques d'évaluation utilisées

Dans la détection de la parole qui se chevauche, différentes mesures sont utilisées pour estimer les performances des modèles ; dans notre étude, nous avons utilisé les quatre mesures

Suivantes. En outre, nous avons analysé les courbes caractéristiques d'exploitation des récepteurs (ROC).

3.3.1 La précision

La précision de la classification est une mesure cruciale qui détermine la proportion de prédictions correctes par rapport au nombre total d'instances. Elle répond à la question : « Combien de prédictions étaient correctes parmi toutes celles qui ont été faites ? ». L'évaluation de la précision constitue généralement la première étape dans l'analyse des performances d'un modèle, car elle permet une compréhension rapide de son efficacité. La précision est particulièrement importante lorsque les faux positifs doivent être minimisés.

- **Formule mathématique** :

$$\textbf{Précision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{3.1}$$

Où :

TP = Vrais positifs

FP = Faux positifs

- **Exactitude (Accuracy)**

L'exactitude est une mesure essentielle en classification, utilisée pour évaluer de manière directe la performance d'un modèle. Elle correspond au rapport entre le nombre d'instances correctement prédites et le total d'instances dans l'ensemble de données. Autrement dit, L'exactitude répond à la question : « Parmi toutes les prédictions faites, combien étaient correctes ? » [47].

- **Formule mathématique** :

$$\textbf{Exactitude} = \frac{\textbf{TP}+\textbf{TN}}{\textbf{TP}+\textbf{TN}+\textbf{FP}+\textbf{FN}} \tag{3.2}$$

Où :

TP = Vrais positifs

TN = Vrais négatifs

FP = Faux positifs

FN = Faux négatifs

3.3.2 Le rappel

Appelé sensibilité ou taux de vrais positifs, le rappel est une mesure cruciale en classification qui met l'accent sur la capacité du modèle à identifier toutes les occurrences pertinentes. Il évalue le pourcentage d'instances positives réelles détectées avec précision par le modèle. La question abordée est la suivante : « Parmi tous les cas positifs réels, combien le modèle a-t-il correctement prédit ?

❖ **Formule mathématique :**

$$\textbf{Recall} = \frac{TP}{TP+FN} \tag{3.2}$$

Où :

- TP = Vrais positifs
- FN = Faux négatifs

Limites

Les mesures de rappel visent à identifier tous les cas positifs, même au prix d'un plus grand nombre de faux positifs.

Le modèle peut qualifier la plupart des cas de positifs pour obtenir un rappel élevé. Parfois c'est comme ça.

Cela entraîne de nombreuses prédictions positives incorrectes, ce qui peut réduire la précision du modèle et conduire à des actions ou interventions inutiles en réponse à ces fausses alarmes.

3.3.3 F1 Score

Le score F1 est une métrique utilisée pour évaluer les performances d'un modèle de classification, notamment lorsqu'il est important de trouver un équilibre entre précision et rappel (Recall). Ceci est particulièrement utile lorsque la classe de données est déséquilibrée, c'est-à-dire lorsque l'une des classes est beaucoup plus représentée que l'autre [48].

Formule mathématique :

$$\boldsymbol{F1 = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}} \quad (3.4)$$

3.4 Résultats expérimentaux

3.4.1 Comparaison des performances de l'arbre de décision, forêt aléatoire et d'AdaBoost

La figure9 représente les histogrammes de accuracy, precision, recall et f1score de Decision Tree performances.

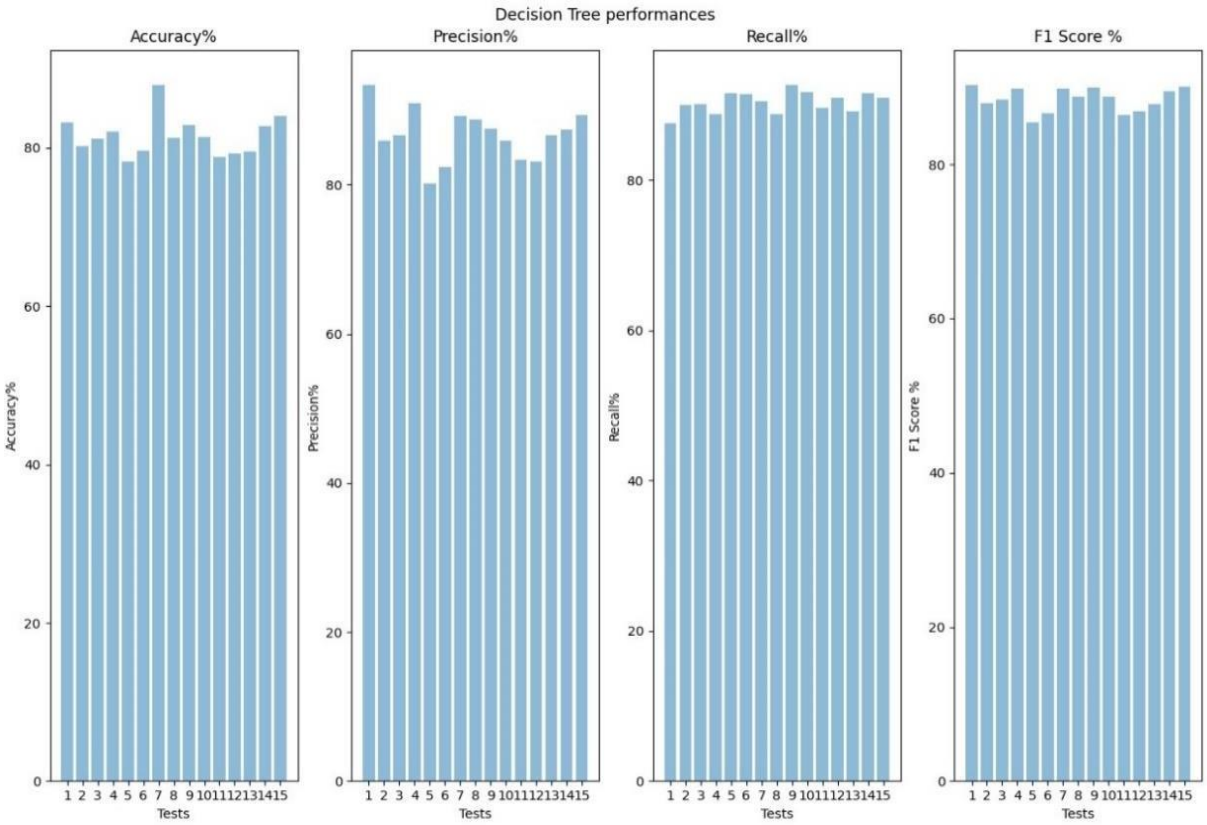

Figure 9 Performances de la détection de la parole superposée à l'aide d'un classificateur Arbre de Décision.

La figure10 représente les histogrammes de accuracy , precision , recall et f1score de Random Forest performances

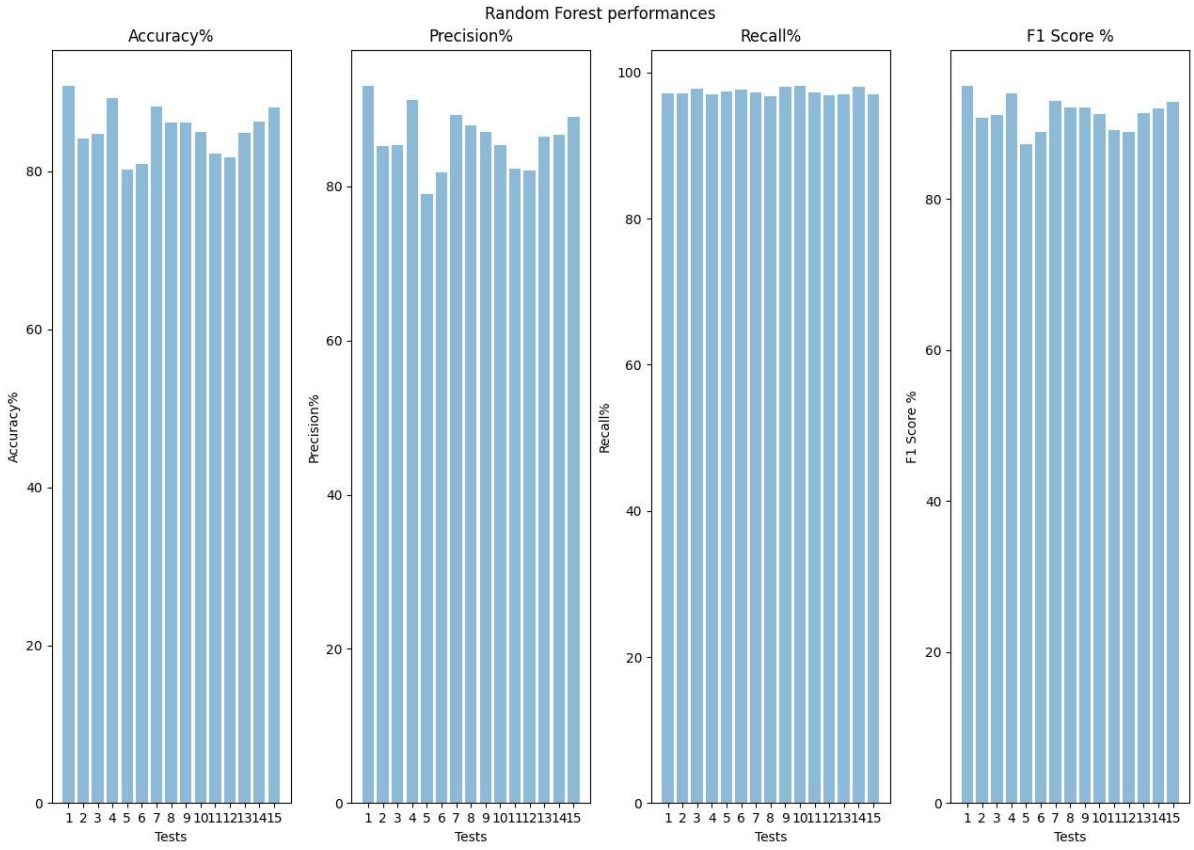

Figure 10 Performances de la détection de la parole superposée à l'aide d'un classificateur forêt aléatoire.

La figure11 représente les histogrammes de accuracy , precision , recall et f1score de AdaBoost performances

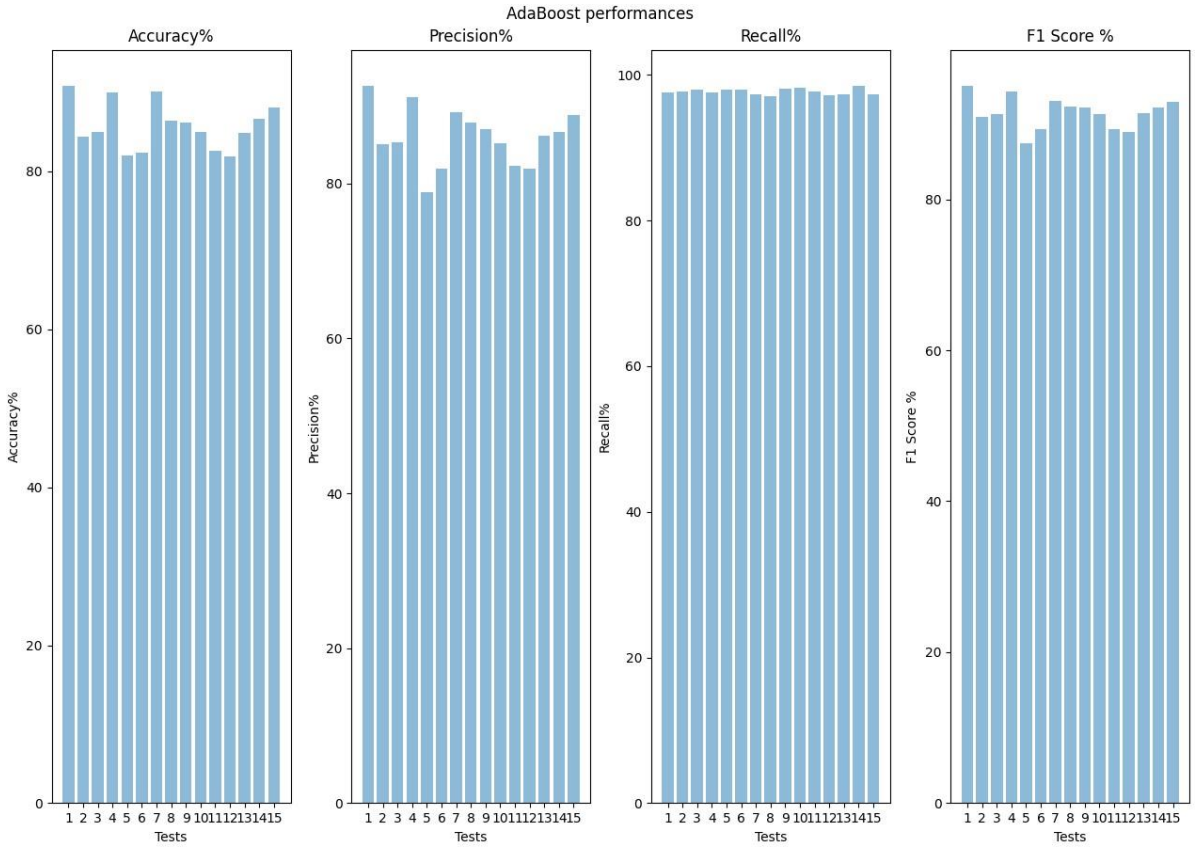

Figure 11 Performances de la détection de la parole superposée à l'aide d'un classificateur AdaBoost.

- **Discussion**

Les performances de plusieurs classificateurs (Arbre de décision, forêt aléatoire et AdaBoost) sont présentées dans les figures 9, 10 et 11 en utilisant de divers critères de performance (exactitude%, Precision%, Rappel%, Score F1%). Les mesures de performance sont fournies pour 15 tests différents.

1. AdaBoost

- **Précision (Accuracy%)** : Les performances varient légèrement entre 80% et 90%. Les barres sont assez stables, ce qui indique que le modèle maintient une précision relativement constante à travers les tests, avec quelques variations.
- **Précision (Precision%)** : Les valeurs de précision varient de manière plus marquée, avec des résultats entre 70% et 90%.

- **Rappel (Recall%)** : Le rappel est très stable, proche de 100% dans presque tous les tests, montrant que le modèle détecte bien les cas positifs.
- **Score F1 (F1 Score%)** : Comme le F1 Score est une moyenne harmonique de la précision et du rappel, il suit une tendance intermédiaire, avec des valeurs entre 80% et 90%.

2. Arbre de decision

- **Précision (Accuracy%)** : Les performances sont similaires à celles de l'AdaBoost, avec des valeurs d'accuracy variant entre 80% et 90%. Il y a une légère variation entre les tests.
- **Précision (Precision%)** : Il y a une certaine variabilité, avec des valeurs variant de 60% à 90%. Cela démontre que dans certaines situations, le modèle peut rencontrer des obstacles pour réduire les faux positifs.
- **Rappel (Recall%)** : Les valeurs de rappel restent élevées, oscillant autour de 90%. Néanmoins, on observe des diminutions légères par rapport à l'AdaBoost, ce qui pourrait suggérer une légère diminution de la détection des vrais positifs.
- **Score F1 (F1 Score%)** : Le score F1 présente des variations similaires, avec une légère baisse dans certains tests, ce qui est cohérent avec les variations de la précision.

3. Forêt aléatoire

- **Précision (Accuracy%)** : Les performances sont très similaires à celles des deux autres modèles, avec des valeurs d'accuracy principalement entre 80% et 90%. La stabilité est meilleure que dans l'AdaBoost.
- **Précision (Precision%)** : Les valeurs de précision montrent une amélioration par rapport à l'AdaBoost et au Decision Tree, avec des valeurs plus constantes autour de 80% à 90%.
- **Rappel (Recall%)** : Le rappel est très élevé et stable, souvent proche de 100%, ce qui suggère que ce modèle détecte très bien les cas positifs, similaire à l'AdaBoost.
- **Score F1 (F1 Score%)** : Le score F1 est également stable, avec des valeurs généralement comprises entre 80% et 90%, ce qui reflète la bonne performance globale du modèle.

Le foret aléatoire semble être le modèle le plus robuste et stable parmi les trois, offrant des performances cohérentes dans toutes les métriques. Dans toutes les métriques, l’Arbre de

decision présente une plus grande variation, ce qui pourrait suggérer une plus grande sensibilité aux données d'entrée. L'AdaBoost se situe entre les deux, proposant une performance globale satisfaisante, mais avec quelques variations.

3.4.2 Arbre de Décision, Forêt Aléatoire et AdaBoost : Comparaison des courbes ROC et Précision-Rappel

Grâce aux résultats du travail, nous avons obtenu cette courbe montrée dans la figure ce dessus

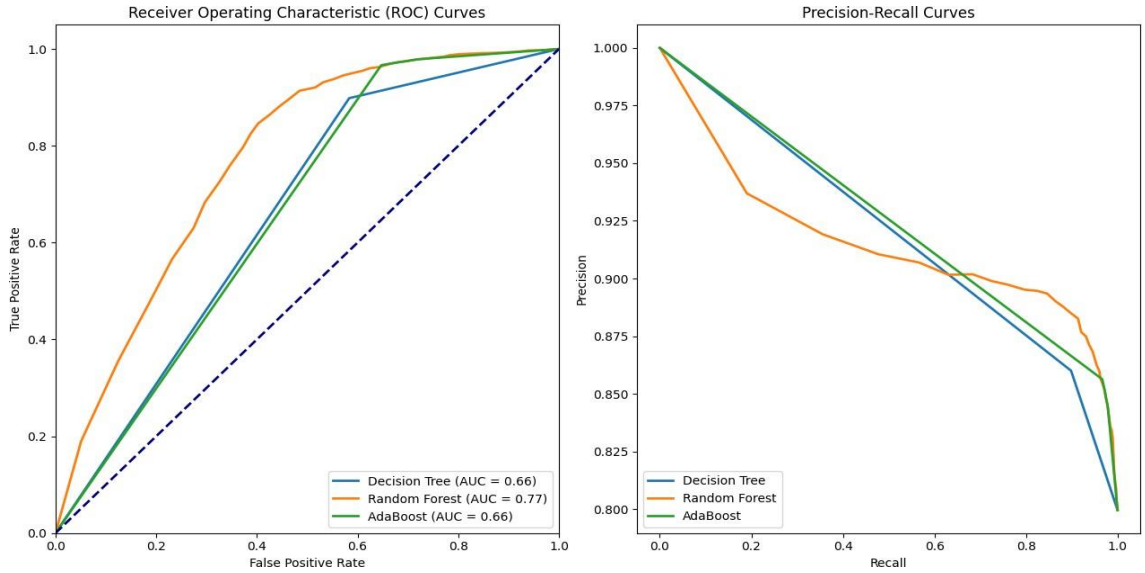

Figure 12 Comparaison des courbes ROC et Précision-Rappel (Recall) pour les classificateurs Decision Tree, Random Forest et AdaBoost.

➢ **Discussion**

La figure 12 comprend deux graphiques comparant les performances des classificateurs Decision Tree, Random Forest et AdaBoost à l'aide des courbes ROC et des courbes précision-rappel (Recall). Le graphique ROC à gauche met en évidence comment chaque modèle équilibre le compromis entre le taux de vrais positifs (sensibilité) et le taux de faux positifs à travers différents seuils. Le modèle Random Forest présente la plus grande surface sous la courbe **AUC** signifie **Area Under the Curve** (AUC = 0,77), indiquant une meilleure capacité à différencier les classes par rapport aux modèles Decision Tree et AdaBoost, qui ont tous deux une AUC de 0,66. Plus une courbe est proche du coin supérieur gauche, meilleure est la capacitédu modèle à faire des prédictions de vrais positifs tout en minimisant les faux positifs. L'AUC

Plus élevée de Random Forest reflète sa capacité à capturer les motifs complexes des données, conduisant à une meilleure performance globale de classification.

Le graphique précision-rappel à droite démontre l'équilibre entre la précision et le rappel pour chaque modèle. Random Forest maintient une haute précision sur une large gamme de valeurs de rappel, ce qui suggère qu'il fait constamment des prédictions positives correctes sans une baisse significative de la performance à mesure que le rappel augmente. Cette performance robuste est cruciale pour des tâches comme la détection de parole superposée, où un rappel élevé assure que la plupart des cas positifs sont capturés et une haute précision minimise les faux positifs. Le modèle Decision Tree montre une baisse rapide de la précision à mesure que le rappel augmente, indiquant qu'il a du mal à maintenir l'exactitude en capturant davantage de vrais positifs. AdaBoost présente des performances comparables à celles de Random Forest, en particulier à des niveaux de rappel plus élevés, reflétant son accent itératif sur les cas difficiles, réalisant ainsi un bon équilibre entre la capture des vrais positifs et le maintien de la précision. Globalement, ces courbes révèlent que les méthodes d'ensemble comme Random Forest et AdaBoost offrent une performance plus fiable que le classificateur Decision Tree simple, les rendant mieux adaptées aux tâches de détection complexes.

Comme nous pouvons le voir dans la courbe ROC, la différence entre les valeurs d'AUC (surface sous la courbe) et les valeurs d'exactitude (accuracy) réside dans le fait qu'elles mesurent des aspects distincts des performances du classificateur (accuracy). L'AUC, dérivée des courbes ROC, évalue la capacité du modèle à distinguer les classes à travers tous les réglages de seuil possibles, fournissant une évaluation complète de la sensibilité et de la spécificité, ce qui donne généralement des valeurs inférieures (par exemple, 0,66 à 0,77) à celles de l'exactitude (accuracy). En revanche, l'exactitude mesure la proportion de prédictions correctes et est souvent plus élevée (par exemple, 80-90%) car elle prend en compte toutes les classifications correctes, y compris celles qui sont faciles et qui peuvent ne pas tenir compte du déséquilibre des classes. L'AUC est moins influencée par le déséquilibre des classes et évalue les performances aux seuils extrêmes, en faisant une métrique plus nuancée pour des tâches comme la détection de parole superposée, où la capacité du modèle à discerner des motifs complexes et subtils est cruciale. L'exactitude, en revanche, est généralement calculée à un seuil fixe et peut être trompeuse dans les ensembles de données déséquilibrés, offrant une vue trop

Optimiste en se concentrant sur les classifications correctes globales sans mettre en évidence lesfaiblesses du modèle à travers différents seuils de décision. Comprendre les valeurs d'AUC inférieures par rapport aux valeurs d'exactitude plus élevées aide à saisir les limitations et les forces du classificateur, garantissant une évaluation approfondie de ses performances dans des tâches de classification complexes.

D'après le graphique précision-rappel et les histogrammes fournis par les figures (Fig. 9, 10, 11 et 12), nous pouvons voir que la divergence entre la variation synchronisée de la précision et du rappel dans les histogrammes et leur relation inverse dans les courbes précision-rappel découle de la différence dans la manière dont ces métriques sont calculées et affichées. Les histogrammes reflètent les valeurs de précision et de rappel à un seuil fixe pour chaque ensemble de tests, montrant la performance globale où les deux métriques augmentent et diminuent souvent ensemble en raison du comportement cohérent du classificateur à travers différents scénarios de tests. En revanche, les courbes précision-rappel illustrent le compromis entre ces métriques à travers une gamme de seuils. Abaisser le seuil augmente le rappel en capturant plus de vrais positifs, mais augmente aussi les faux positifs, réduisant ainsi la précision. À l'inverse, augmenter le seuil améliore la précision en réduisant les faux positifs, mais diminue le rappel car moins de vrais positifs sont détectés. Cet ajustement dynamique des seuils dans les courbes précision-rappel met en évidence le compromis inhérent entre la précision et le rappel, rendant leur relation inverse plus apparente par rapport à l'approche de seuil fixe utilisée dans les histogrammes.

Une sensibilité : Signifie qu'est une métrique utilisée en classification pour évaluer la performance d'un modèle. Elle mesure la proportion de cas positifs correctement identifiés par le modèle par rapport au nombre total de cas positifs réels.

3.4.3 Arbre de Décision, Forêt Aléatoire et AdaBoost : Comparaison des courbes ROC et Précision-Rappel (Nombre d'estimateur =100).

Les figures 13, 14, et 15 présentent les performances des classificateurs Arbre de Décision, Forêt Aléatoire, et AdaBoost pour la détection de parole superposée avec un nombre d'estimateurs 100. Les histogrammes dans ces figures comparent les performances des modèles à travers des critères comme la précision (Accuracy), le rappel (Recall), la précision (Precision), et le F1 Score pour 15 tests

Figure 13 Performances de la détection de la parole superposée à l'aide d'un classificateur Arbre de Décision.

Figure 14 Performances de la détection de la parole superposée à l'aide d'un classificateur Forêt Aléatoire.

Figure 15 Performances de la détection de la parole superposée à l'aide d'un classificateur AdaBoost..

➢ **Discussion**

1. Forêt Aléatoire (Random Forest)

- **Précision (Accuracy%)** : Les performances varient entre 80% et 90%. Une amélioration notable est observée lorsque le nombre d'estimateurs passe de 50 à 100, offrant une meilleure stabilité et une précision plus constante à travers les différents tests.
- **Précision (Precision%)** : Les valeurs se situent généralement autour de 80% à 90%. L'augmentation du nombre d'estimateurs à 100 réduit la variabilité et stabilise les résultats, notamment en améliorant la gestion des faux positifs.

- **Rappel (Recall%)** : Très élevé, souvent proche de 100%. L'augmentation du nombre d'estimateurs améliore légèrement la capacité du modèle à détecter correctement les cas positifs.
- **Score F1 (F1 Score%)** : Généralement compris entre 80% et 90%. L'augmentation à 100 estimateurs améliore la stabilité du score F1, offrant une performance plus fiable et équilibrée entre précision et rappel.

2. AdaBoost

- **Précision (Accuracy%)** : Varie entre 80% et 90%. L'augmentation du nombre d'estimateurs de 50 à 100 montre une légère amélioration dans la précision, bien que la variabilité reste légèrement plus élevée qu'avec la Forêt Aléatoire.
- **Précision (Precision%)** : Les valeurs varie entre 70% et 90%. L'augmentation du nombre d'estimateurs contribue à une meilleure précision en réduisant les erreurs de prédiction, bien que cette amélioration soit plus modeste comparée à celle de la Forêt Aléatoire.
- **Rappel (Recall%)** : Très stable, proche de 100% dans presque tous les tests. L'augmentation du nombre d'estimateurs améliore la stabilité du rappel, permettant au modèle de mieux identifier les cas positifs.

- **Score F1 (F1 Score%)** : Situé entre 80% et 90%. L'augmentation du nombre d'estimateurs tend à améliorer le score F1, mais les gains sont plus modestes que ceux observés avec la Forêt Aléatoire.

3. Arbre de Décision (Decision Tree)

- **Précision (Accuracy%)** : Varie entre 80% et 90%. La performance est relativement stable, avec une légère amélioration observée lorsque le nombre d'estimateurs augmente, mais elle reste moins stable que les méthodes ensemblistes comme la Forêt Aléatoire.
- **Précision (Precision%)** : Varie de 60% à 90%. Du nombre d'estimateurs contribue à stabiliser la précision en réduisant les variations extrêmes, mais la performance reste plus variable que celle des modèles d'ensemble.

- **Rappel (Recall%)** : Autour de 90%, avec des baisses notables dans certains cas. L'augmentation du nombre d'estimateurs améliore généralement le rappel, bien que ce ne soit pas aussi stable que dans les autres modèles.

- **Score F1 (F1 Score%)** : Varie en fonction de la précision et du rappel, avec des baisses possibles. Une augmentation du nombre d'estimateurs tend à améliorer le score F1, mais l'amélioration reste plus modeste que pour la Forêt Aléatoire et AdaBoost.

3.4.4 Impact de l'Augmentation du Nombre d'Estimateurs à 100 sur les Performances des Modèles

Les résultats montrent que l'augmentation à 100 estimateurs permet généralement d'améliorer la précision et la stabilité des modèles, en particulier pour les modèles ensemblistes comme la Forêt Aléatoire et AdaBoost. La Forêt Aléatoire est la plus performante, montrant une stabilité accrue dans toutes les métriques. AdaBoost bénéficie également de cette augmentation, mais dans une moindre mesure. Enfin, bien que l'Arbre de Décision montre des améliorations, il reste plus sensible aux variations de données comparé aux modèles ensemblistes.

3.4.5 Comparaison des Courbes ROC et Précision-Rappel

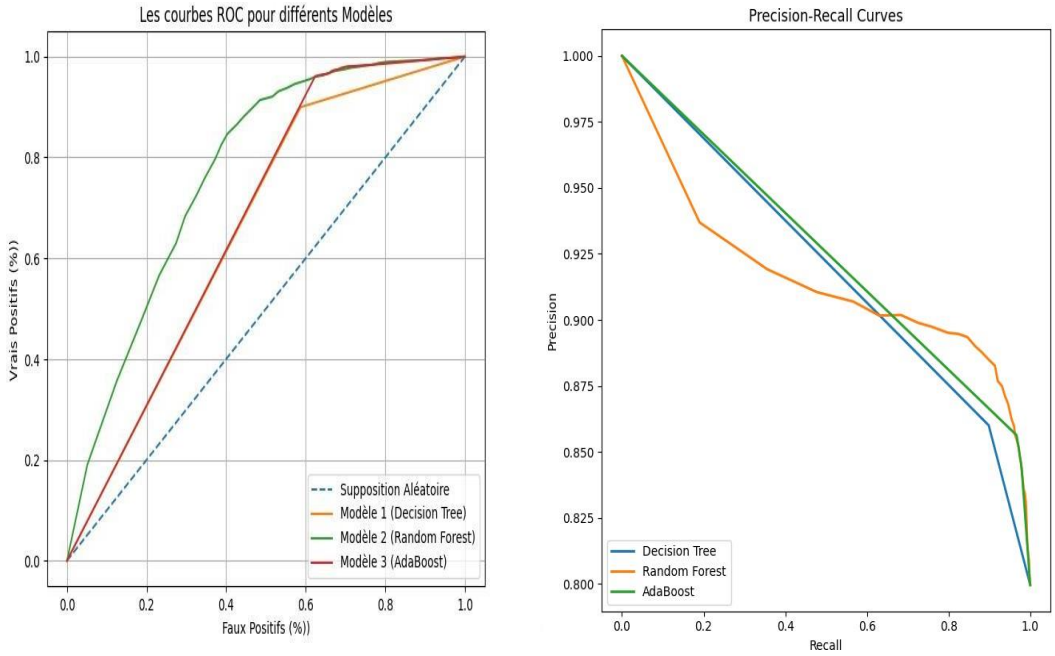

Figure 16 Comparaison des courbes ROC et Précision-Rappel pour les classificateurs Arbre de Décision, Forêt Aléatoire et AdaBoost.

- **Courbes ROC** : La Forêt Aléatoire présente la plus grande surface sous la courbe (AUC = 0,77), ce qui indique une meilleure capacité à différencier les classes par rapport aux modèles Arbre de Décision et AdaBoost, qui affichent tous deux une **AUC de 0,66**. Cela reflète la capacité de la Forêt Aléatoire à faire des prédictions plus précises tout en minimisant les faux positifs à travers divers seuils de décision.
- **Courbes Précision-Rappel** : La Forêt Aléatoire maintient une haute précision sur une large gamme de valeurs de rappel, ce qui montre sa capacité à faire des prédictions correctes sans perdre en performance lorsque le rappel augmente. En revanche, le modèle **Arbre de Décision** montre une baisse rapide de la précision à mesure que le rappel augmente, ce qui signifie qu'il détecte plus de vrais positifs, mais au prix d'un plus grand nombre de faux positifs. **AdaBoost**, quant à lui, présente une performance intermédiaire, réussissant à maintenir un bon équilibre entre précision et rappel, mais avec une plus grande variabilité que la Forêt Aléatoire.

- **Discussion**

Les performances des modèles (Arbre de Décision, Forêt Aléatoire, et AdaBoost) montrent que la **Forêt Aléatoire** est la plus robuste et stable des trois, avec une meilleure capacité à maintenir une haute précision et un rappel élevé. Cela est confirmé par les courbes ROC et les courbes Précision-Rappel. L'analyse montre que la Forêt Aléatoire avec **100 estimateurs** est le modèle le plus efficace et le plus équilibré pour la détection de parole superposée. Bien que chaque modèle ait ses avantages spécifiques, la Forêt Aléatoire se distingue par sa **stabilité** et ses **performances globales** dans toutes les mesures d'évaluation.

3.4.6. Impact de l'Augmentation du Nombre d'Estimateurs sur les Métriques

- **Précision et Exactitude** : Avec plus d'estimateurs, la précision et l'exactitude tendent à s'améliorer car le modèle intègre plus d'informations et réduit les erreurs de prédiction.
- **Rappel** : Le rappel maintient un niveau élevé avec plus d'estimateurs, capturant plus de vrais positifs.
- **Score F1** : S'améliore avec l'augmentation du nombre d'estimateurs si la précision et le rappel sont également améliorés.

En général, l'augmentation du nombre d'estimateurs de 50 à 100 améliore la performance globale, offrant une meilleure stabilité et précision pour les modèles.

3.5 Conclusion

L'évaluation comparative des classificateurs AdaBoost, Arbre de Décision et Forêt Aléatoire pour la détection de parole superposée met en évidence des différences notables, notamment en fonction du nombre d'estimateurs utilisés. Bien que tous les modèles affichent une exactitude globale similaire, variant entre 80% et 90%, leurs performances diffèrent sur des métriques clés telles que la précision, le rappel et le score F1.

AdaBoost se distingue par un rappel élevé et stable, indiquant sa capacité à identifier la plupart des cas positifs. Cependant, sa précision reste plus variable, montrant une difficulté à réduire les faux positifs de manière constante. L'Arbre de Décision, bien que présentant également un bon rappel, se montre plus sensible aux variations des données, entraînant des fluctuations importantes dans ses performances, particulièrement en termes de précision.

La Forêt Aléatoire, quant à elle, s'impose comme le modèle le plus robuste et équilibré. Elle combine une précision constante et un rappel élevé, offrant ainsi une performance stable à travers différents scénarios. L'augmentation du nombre d'estimateurs à 100 améliore les performances de tous les modèles, mais c'est la Forêt Aléatoire qui en tire le plus grand bénéfice, se révélant être la méthode la plus performante. Elle atteint un AUC de 0,77 dans les courbes ROC, démontrant une capacité supérieure à discriminer les classes, ce qui en fait le modèle le plus efficace pour cette tâche complexe.

Conclusion Générale

CONCLUSION GENERALE :

Cette étude a examiné les bases de la Détection de Parole Superposée, soulignant son importance dans divers domaines tels que la communication, la surveillance et le traitement automatique de la parole. Les Modèles Ensemblistes ont été introduits comme une approche prometteuse de modélisation statistique pour relever les défis complexes de la détection de parole superposée, en présentant les principes fondamentaux tels que le vote majoritaire, le bagging, le boosting et le stacking.

Une analyse de la littérature a permis de situer notre recherche par rapport aux travaux antérieurs et aux méthodes existantes, identifiant ainsi les défis et les lacunes actuels dans la détection de parole superposée, offrant ainsi une base solide pour notre travail.

La méthodologie détaillée pour la construction des modèles ensemblistes a été exposée, couvrant la collecte et le prétraitement des données, ainsi que la théorie des arbres de décision, des forêts aléatoires et d'AdaBoost pour la détection de parole superposée.

Une évaluation des performances des modèles ensemblistes a été réalisée, mettant en évidence à la fois leurs points forts et leurs limites. Les résultats expérimentaux ont fourni des idées précieuses pour améliorer la précision et la robustesse des modèles ensemblistes dans la détection de parole superposée.

Une évaluation comparative des modèles d'ensemble AdaBoost, des arbres de décision et des classificateurs de forêt aléatoire couvrant la détection de la parole montre que, bien que la précision globale soit similaire (80 à 90 %), leurs performances diffèrent sur des indicateurs clés tels que la précision, le rappel et l'effondrement des différences de score F1. AdaBoost se distingue par son rappel élevé et sa précision variable plus élevée. Les arbres de décision ont de bons taux de rappel, mais sont sensibles aux changements de données, affectant leur précision. À lui seul, Random Forest est le plus puissant et le plus équilibré, offrant une précision constante et un rappel

Élevé. Tous les modèles ont été affinés à l'aide de 100 estimateurs, mais Random Forest a obtenu les meilleurs résultats avec une AUC de 0,77, ce qui indique un pouvoir discriminant élevé.

Dans l'ensemble, cette étude a contribué à une meilleure compréhension et à l'amélioration des techniques de détection de parole superposée en tirant parti du potentiel des modèles ensemblistes. Les pistes de recherche futures pourraient explorer davantage les possibilités d'amélioration des performances, notamment en examinant de nouvelles caractéristiques acoustiques ou en peaufinant les méthodes de prétraitement des données.

Références

[1] Derrardja I, ben malek nour el houda, (2022). « Identification du locuteur par GMM., » Université de Mohamed El-Bachir El-Ibrahimi -Bordj Bou Arreridj.

[2] benbelaid. a. Benbaziz. t, (2020). « Segmentation en locuteurs d'un document audio, » Bordj Bou Arreridj.

[3] Adda Gilles, (2011). « Approches empiriques et modélisation statistique de la parole.Interfacehomme-machine [cs.HC].,» Paris XI.

[4] Landini F., J. Profant, M. Diez, and L. Burget, (2022). «Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks,» Computer Speech & Language, p. 71.

[5] Schnell, N., Peeters, G., Lemouton, S., Manoury, P., & Rodet, X, (2000). «Synthesizing a choir in real-time using PitchSynchronous Overlap Add (PSOLA), » In ICMC.

[6] Fant. G, « Head, Speech communication and Musical Acoustics, » Stockhilm, sweden.

[7] Cisonni, J. (2008), « Modélisation et inversion d'un système complexe deproduction de signaux acoustiques. Application à la voix et aux pathologies, » Doctoral dissertation, Grenoble-INPG.

[8] Gérard, C. (1995), « Etude de la paramétrisation du signal de parole à partir de représentationsen ondelettes, » (Doctoral dissertation, Paris 11).

[9] Deroo, O. (1998), « Modèles dépendants du contexte et méthodes de fusion de données à lareconnaissance de la parole par modèles hybrides HMM/MLP, » (Doctoral dissertation, Thèse de doctorat de la Faculté Polytechnique de Mons, Laboratoire TCTS Mons).

[10] Barras, C. (1996), « Reconnaissance de la parole continue : adaptation au locuteur et contrôletemporel dans les modèles de Markov cachés, » (Éditeur inconnu), 1996.

[11] M. T. Ayad (2015), « Reconnaissance de Cibles par Radar a écho Doppler., » Mémoire de find'études pour l'obtention du diplôme d'ingénieur d'état en défense Aérienne.

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

[13] Rabiner, L. R., & Juang, B. H. (1999), «Fundamentals of speech recognition, » Tsinghua University Press.

[14] L. Foued. (2017), « L'ANALYSE SPECTRALE. COURS DE TECHNIQUES DE SURVEILLANCE, ».

[15] Rabiner, L. R., &Schafer, R. W. (2011). Theory and applications of digital speech processing. Prentice Hall.

[16] O'Shaughnessy, D. (1987), «Speech communications: Human and machine (IEEE), » Universities press.

[17] Coleman, J. S. (2005), «Introducing speech and language processing, » Cambridge university press.

[18] Marcel, M. (2014), « La diversité des modèles. Dans Recherche et innovation., ».

[19] Breiman, L. (1996), Bagging predictors. Machine learning 24, pp. 123-140.

[20] Friedman, J., Hastie, T., & Tibshirani, R. (2000), « Additive logistic regression: a statistical view of boosting» (with discussion and a rejoinder by the authors). The annals of statistics, p. 28.

[21] Ledezma, A., Aler, R., Sanchis, A., & Borrajo, D. (2010), «GA-stacking: Evolutionary stacked generalization». Intelligent Data Analysis, pp. 89-119.

[22] [. c.sabin@ucl.ac.uk, «Research Department of Infection and Population Health, UCL, Royal Free Campus, Rowland Hill Street, » London.

[23] Rust, K. F., & Rao, J. N. K. (1996), « Variance estimation for complex surveys using replication techniques». Statistical methods in medical research, pp. 283-310.

[24] Garnerin, M., Rossato, S., & Besacier, L. (2019, October), «Gender representation in French broadcast corpora and its impact on ASR performance. In Proceedings of the 1st international workshop on AIfor smart TV content production, access and deliver, ».

[25] Plante, I. (2012), L'apprentissage coopératif : des effets positifs sur les élèves aux difficultés liées à sonimplantation en classe « REVUE CANADIENNE DE L'ÉDUCATION, » 35, 3 : 252 – 283.

[26] Koolagudi, S. G., Rastogi, D., & Rao, K. S. (2012), «Identification of language using mel- frequencycepstral coefficients (MFCC) ». Procedia Engineering, pp. 38, 3391-3398.

[27] Alsouda, Y., Pllana, S., & Kurti, A. (2019, May), «Iot-based urban noise identification using machine learning: performance of SVM, KNN, bagging, and random forest. In Proceedings of the international conference on omni-layer intelligent systems, ».

[28] Al-Talabani, A. (2015), «Automatic speech emotion recognition-feature space dimensionality and classification challenges, » (Doctoral dissertation, University of Buckingham).

[29] Dogra, A., Kaul, A., & Sharma, R. (2019, October), « Automatic recognition of dialects of Himachal Pradesh usingMFCC &GMM. In 2019 5th international conference on signal processing, computing and control (ISPCC) », pp. 134-137.IEEE.

[30] Geiger, J. T., Eyben, F., Schuller, B., & Rigoll, G. (2013), «Detecting overlapping speech with long short-term memory recurrent neural networks. In Proceedings Interspeech, 14th Annual Conference of the International Speech Communication Associaion, » Lyon, France.

[31] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018), « Bert: Pre-training of deep bidirectional transformersfor language understanding, ». arXiv preprint arXiv:1810.04805.

[32] Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., ... & Gill, M. P. (2020, May), « Pyannote. audio: neural building blocks for speaker diarization. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, » (pp. 7124- 7128).

[33] Demange, S. (2007), « Contributions à la reconnaissance automatique de la parole avec données manquantes, » (Doctoral dissertation, Université Henri Poincaré-Nancy 1).

[34] Bourlard, H., & Wellekens, C. J. (1990), «Links between Markov models and multilayer perceptrons, IEEE Transactions on pattern analysis and machine intelligence, » pp. 12(12), 1167-1178.

[35] Lebourdais, M. (2023), « Interactions entre locuteurs : de la détection de la parole superposée à la détection des interruptions, » (Doctoral dissertation, Université du Mans, Le Mans, FRA.).

[36] Marie Tahon. (2023), « Traitement automatique de la parole expressive : retour vers des systèmesinterpréta-bles ? Intelligence artificielle [cs.AI]., » Le Mans Université.

[37] BEYER, Antoine et LACOSTE, Romuald, (2017), « La coopération interportuaire-Stratégies de coopération et miseen réseaux des ports intérieurs et maritimes européens. Les Techniques de l'Ingenieur, ».

[38] Tahon, M. (2023), « Traitement automatique de la parole expressive : retour vers des systèmesinterprétables ?» (Doctoral dissertation, Le Mans Université).

[39] Sini, A. (2020), «Characterisation and generation of expressivity in function of speakingstyles for audiobook synthesis, » (Doctoral dissertation, Université Rennes 1).

[40] Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., & Esteve, Y. (2018), « TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In Speech and Computer: 20th International Conference, » SPECOM, Leipzig, Germany, Proceedings 20 (pp. 198-208). Springer International Publishing.

[41] Lotfian, R., & Busso, C. (2017), « Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings, ». IEEE Transactions on Affective Computing, pp. 10(4), 471-483.

[42] Zaabi, K. (2004), « Implémentation d'une méthode de reconnaissance de la parole sur leprocesseur de traitement numérique du signal TMS320C6711, ». (Doctoral dissertation, École de technologie supérieure).

[43] Liu, Z. T., Wu, M., Cao, W. H., Mao, J. W., Xu, J. P., & Tan, G. Z. (2018), « Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing, ». pp. 273, 271-280.

[44] Breiman, L. (2001), «Random forests. Machine learning, », pp. 45, 5-32.

[45] Tahon, M., & Lolive, D. (2018, June), «Discourse phrases classification: direct vs. narrative audiospeech. In Speech Prosody, ».

[46] Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G. (2013), «Advance and prospects of AdaBoost algorithm, ». Acta Automatica Sinica, pp. 39(6), 745-758.

[47] SHUNG, Koo Ping. (2018), «Accuracy, precision, recall or F1». Towards data science, vol. 15, no 03.