REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE YAHIA FARES DE MEDEA



Faculté de **T**echnologie

Département du **G**énie **E**lectrique

Mémoire de fin d'études de Master

Filière **:** *Télécommunications*

Spécialité : *Système de Télécommunications*

Thème :

Reconnaissance robuste de la parole synthétisée par GMM

Présentés par :

Zemirline amira

Benaicha rania

Proposé et Dirigé par :

Dr. RAGGAI Djahid

Promotion 2022-2023

Dédicace

Je dédie ces mots exprimant mon amour et ma reconnaissance à mes chers parents, qui m'ont transmis leur amour, leur sagesse et leur force, votre confiance en moi a été le moteur de ma réussite, Je suis si chanceuse de vous avoir comme parents, merci pour tous les sacrifices qui vous avez fait pour moi.

Je dédie ce fruit de travail à mes frères RIADH et ABDERRAHMANE, je souhaite exprimer ma gratitude pour votre présence constante, vos mots gentils, vos sourires ont été un rayon de soleil dans ma vie et surtout votre soutien qui a été le catalyseur qui m'a permis de surmonter les obstacles et de réaliser mes rêves.

Ce mémoire est également dédié à mes chères copines MANEL, ZAHRA, DOUNIA, ZOHRA, HANAA et MAHDIA vous m'avez laissées des belles choses qui ne finissent jamais.

À mon âme qui se tenait debout même quand elle était surmenée. Vos contributions ont été inestimables et je vous en suis profondément reconnaissante.

Sans oublier ma binôme RANIA, bien que nous ayons fait face à des difficultés mais cela n'a pas empêché notre expérience d'être amusante et enrichissante

Merci infiniment."

Amira

Avec l'expression de ma reconnaissance, je dédie ce modeste travail à ce qui, quelques soient les termes embrassés, je n'arriverais jamais à leurs exprimer mon amour sincère.

A la plus belle créature que dieu a crée sur terre, source de tendresse de patience et générosité ma lune qui allumer mon chemin ma moitie ma mère Fouzia.

Celui qui m'a fait une femme, mon support qui était toujours a mes cotes pour me soutenir et m'encourage à mon père Sofien.

A mes chères sœurs Manel et Amina mes amours mes confidentes que dieu les protèges et leurs offre la chance.

A mon adorable petite sœur Ines qui sait toujours comment procurer la joie et le bonheur pour toute la famille.

A la meilleure personne ma source de joie ABDELRAOUF WAIL qui m'as toujours donné le courage.

A mes chères sœurs que la vie m'offre Meriem Amira et Wissam qui n'ont pas cessées de me conseiller, encourager et soutenir tout au long de mes études.

A mes grandes mères Yamina et Massouda ; A mes tantes Ratiba Lamia Soulef et Zohra que dieu leurs donnes une longue et joyeuse vie.

A mes chères cousines Sabrina Lina Malek Yasmine Safia Nesrine.

A mes cousins Sidali Issam Mouatez Madjid.

A mes très chères amies Sarah et Manel.

Je remercie Monsieur Bencheref Aid, mon prof de math pour l'accompagnement, la disponibilité et les conseils. A toute la famille Benaicha , Sahari pour leurs amour et encouragements.

Sans oublier mon binôme Zemirline Amira pour son soutien moral, sa patience et sa compréhension tout au long de ce projet.

Rania

Remerciement

Nous tenons à remercier ALLAH le tout puissant qui nous a donné le courage, la connaissance, la patience et le pouvoir afin d'élaborer ce travail

En guise de reconnaissance, nous tenons à remercier notre encadrant

Dr. RAGGAI DJAHID pour ses conseils éclairés, son expertise pointue tout au long de notre projet, ainsi que pour sa disponibilité et sa bienveillance qui ont été des atouts majeurs pour la réussite de notre travail.

Nous sommes honorées d'avoir pu bénéficier de son savoir-faire pour mener à bien cette recherche." de plus, nous tenons à exprimer tout notre gratitude envers tous les enseignants, qui ont été des guides précieux tout au long de notre parcours académique, en particulier, les enseignants du département de télécommunications, merci amplement pour tout.

Nos sincères remerciements aux membres du jury qui nous ont fait l'honneur d'avoir accepté d'évaluer ce travail

Enfin, nous présentons également notre grand remerciement pour nos parents, nos familles et nos amis pour leur amour et leur soutien indéfectible.

Merci à tous

ملخص :

المعطيات ASVspoof 2017-V2. عينات من الكلام الطبيعي و المصطنع من نفس قاعدة المعطيات تم استخدامها لاختبار قدرات النظام. مدخل النظام يستخدم معاملات CQC. خلال هذه الدراسة حاولنا اظهار تاثير عدد من العوامل على كفاءة النظام. تم تغيير عدد و طبيعة المعاملات CQC بالاضافة الى عدد مركبات GMM. اظهرت النتائج ان استخدام 90 معاملCQC من نوع SDA بالاضافة الى 512 مركبة GMM تعطي نتائج مرضية معبر بنسبة خطا متساوية تقدر ب 10% .

الكلما ت المفتاحية: GMMC, CQCC , نسبة الخطأ المتساوية، ASVspoof 2017-V2

Abstract:

We present in this work the use of GMMs for the detection of spoofed speech. The GMMs were trained using the ASVspoof 2017-V2 database. Authentic and spoofed speech from this database were also used to test the performance of the system. The front end of the system uses CQC coefficients. Throughout this study we tried to determine the effect of certain parameters on the performances of the system. We changed the number and the nature of CQC coefficients used in the front-end stage, the number of gaussian mixtures was also changed. Results show that using 90 SDA coefficients combined with 512 mixtures for GMMs gave acceptable EER of around 10%.

Key words: GMM, CQCC, EER, ASVspoof 2017-V2 .

Résumé :

Nous présentons dans ce travail l'utilisation des GMMs pour la détection de parole usurpée. Les GMMs utilisés ont été entrainés en utilisant la base de données ASVspoof 2017-V2. Des échantillons de parole authentique et usurpée dans cette base de données ont été aussi utilisés por tester les performances du système. L'étage d'entrée du système utilise les coefficients CQC. Tout au long de cette étude il a été question d'étudier l'effet d'un certain nombre de paramètres sur les performances du système. Nous avons changé le nombre et la nature des coefficients CQC utilisés dans l'étage d'entrée, le nombre de gaussiennes des GMMs a été aussi changé. Les résultats obtenus montrent que l'utilisation de 90 coefficients SDA combiné avec un nombre de gaussiennes de 512 conduisent à une performance acceptable d'environ 10 % en termes de taux d'erreur égale TEE.

Mots clés : GMM, CQCC, TEE, ASVspoof 2017-V2 .

Table des matières

Dédicace ..... **I**

Remerciement ..... **III**

I.ملخص ..... **IV**

II. Liste des figures ..... **IX**

III. Liste des tableaux ..... **X**

IV. Liste des abréviations ..... **XI**

Introduction Générale **1**

Chapitre I Généralités sur le traitement de la parole 4

I.2 Introduction 5

I.3 Définition de la parole 5

I.4 La production de la parole 5

I.4.1 Larynx 6

I.4.2 Pharynx 6

I.5 La perception de la parole 7

I.5.1 L'appareil auditif humain 7

I.6 Le système nerveux et le cerveau 8

I.7 Les Caractéristiques phonétiques de la parole 9

I.7.1 Les voyelles 9

I.7.2 Les consonnes 9

I.8 Le signal de la parole 9

I.9 Caractéristiques fréquentielles du signal de parole 10

I.9.1 La fréquence fondamentale F0 (Le pitch) 10

I.9.2 Les formants 11

I.10 La Numérisation du signal de la parole 11

I.10.1 Echantillonnage 12

I.10.1.1 Théorème d'échantillonnage 12

I.10.2 La Quantification 13

I.10.3 Le codage 13

I.11 Domaines d'application 13

I.11.1 La reconnaissance automatique de langage 14
I.11.2 Reconnaissance de la parole 15
I.11.3 Reconnaissance et vérification du locuteur 15
I.11.4 La synthèse de la parole 16
I.12 Extraction de paramètres 16
I.12.1 Extraction des paramètres MFCC 16
I.12.2 Extraction des paramètres CQCC 18
I.13 Conclusion 19
Chapitre II Reconnaissance et vérification du locuteur 20
II.1 Introduction 21
II.2 La reconnaissance automatique du locuteur (RAL) 21
II.2.1 Les branches de RAL 22
II.2.1.1 Identification automatique du locuteur (IAL) 22
II.2.1.2 Vérification automatique du locuteur (VAL) 23
II.2.2 Classification automatique du locuteur 24
II.2.3 Segmentation automatique du locuteur 24
II.2.3.1 Suivi de locuteur 25
II.2.4 Modalités de la RAL 25
II.2.4.1 La méthode indépendante du texte 25
II.2.4.2 La méthode dépendante du texte 25
II.2.5 La variabilité de la voix 26
II.2.5.1 La variabilité interlocuteur 26
II.2.5.2 La variabilité intra-locuteur 26
II.2.6 Modélisation des locuteurs 26
II.2.6.1 Définition 26
II.2.6.1 Le modèle de mélange gaussien (GMM) 27
II.2.7 Mesures de performance 28
II.2.7.1 Le taux de faux rejets (FRR) 28
II.2.7.1 Le taux de fausses acceptations (FAR) 29
II.2.7.2 Taux d'erreur égal (EER) 29

II.2.8 Les Courbes de performance 30
II.2.8.1 La Courbe ROC (Receiver Operating Characteristic).....30
II.2.8.2 La courbe DET (Detection Error Tradeoff) 31
II.3 La Détection d'usurpation d'identité 31
II.4 Vulnérabilité de la vérification des conférenciers à la mystification 32
II.4.1 Points d'attaque possibles 32
II.4.1 Les types d'attaques 34
II.4.2 Mécanismes anti-usurpation d'identité 34
II.5 Les attaques par usurpation d'identité (Spoofing attacks) 34
II.5.1 Discours synthétique (Synthetic speech) 34
II.5.2 Conversion vocale (Voice conversion) 35
II.5.3 Les attaques rejeu (replay) 35
II.5.4 L'usurpation d'identité (Impersonation) 35
II.6 Conclusion 35
Chapitre III Résultats et interprétations 36
III.1 Introduction 37
III.2 Présentation de la base de données utilisée 37
III.1.1 Structure de la base de données ASV spoof 2017 37
III.1.2 Procédés d'obtention des données de la base 38
III.1.2.1 Simulation du rejeu (replay) 38
III.1.2.2 Environnements acoustiques 38
III.2.1 Dispositifs de lecture 39
III.1.2.3 Dispositifs d'enregistrements 40
III.2 Stratégie de l'étude **.....41**
III.3 Résultats obtenus **.....42**
III.3.1 Utilisation des coefficients CQCC statiques 42
III.3.2 Utilisation des coefficients dynamiques 42
III.3.3 Utilisation coefficients MFCC de type SDA (statiques + delta + accélération) 43
III.3.4 Effet du nombre de gaussiennes 44

III.4 Interprétations **44**

III.5 Conclusions **45**

Conclusion générale **46**

Liste des figures

Figure I.1: Modèle simplifié de l'appareil phonatoire 5
Figure I.2: Le larynx 6
Figure I.3 : L'oreille externe et le milieu d'oreille. 7
Figure I.4 : L'oreille interne.....8
Figure I.5: Domaines de la production vocale dans le cerveau. 8
Figure I.6 : Représentation des voyelles dans le plan F1 F2.....9
Figure I.7: Fréquence fondamentale.[6].....10
Figure I.8: la production d'une voyelle.....11
Figure I.9: Les étapes de numérisation d'un signal analogique. 12
Figure I.10: signal échantillonné.....12
Figure I.11 : Signal quantifié 13
Figure I.12: Structure d'un système d'identification automatique des langues. 14
Figure I.13 : Calcul des coefficients MFCC [12].....17
Figure I.14: Diagramme schématique pour l'extraction des coefficients (CQCC).....18
Figure II.1 : schéma simplifié le système d'avoir reconnaissance du locuteur.....21
Figure II.2: Architecture d'un système d'identification du locuteur. 23
Figure II.3 : Structure d'un système de vérification du locuteur 23
Figure II.4 : Principe de la segmentation et classification en locuteur 24
Figure II.5: modélisation gaussienne d'une distribution en 2 D 27
Figure II.6: Illustration du point d'intersection des courbes FRR et FAR.....30
Figure II.7: La Courbe ROC 30
Figure II.8:DET représentant la performance de quatre techniques de vérification automatique du locuteur.....31

Figure III.1 : illustration du scénario de rejeu…….……38
Figure III.2 : Courbe ROC obtenue dans le cas de CQCC statiques……..…42
Figure III.3 : Courbe ROC obtenue dans le cas de CQCC dynamiques……..…...44

Liste des tableaux

Tableau III. 1: Statistiques actualisées du corpus ASVspoof 2017 2.0 38

Tableau III. 2: Environnements acoustiques utilisés dans la collecte de données 39

Tableau III. 3 : dispositifs de lecture utilisés dans la collecte des données 40

Tableau III. 4: dispositifs d'enregistrement utilisés dans la collecte des données 41

Tableau III. 5 : EER en fonction du nombre de coefficients CQCC statiques 42

Tableau III. 6: EER en fonction du nombre de coefficients CQCC dynamiques 43

Tableau III. 7: Résultats pour coefficients MFCC de type SDA 44

Tableau III.8 : EER en fonction du nombre de gaussiennes..... 44

Liste des abréviations

ASV	***A**utomatic **S**peaker **V**erification*
CQCC	***C**onstant-**Q** **Ce**pstral **C**oefficient*
CQT	***C**onstant-**Q** **T**ransform*
DCT	***D**iscrete **C**osine **T**ransform*
DET	***D**etection **E**rror **T**radeoff*
DFT	***D**iscrete **F**ourier **T**ransform*
EER	***E**qual **E**rror **R**ate*
EM	***E**xpectation-**M**aximization*
FAR	***F**alse **A**cceptance **R**ate*
FFT	***F**ast **F**ourier **T**ransform*
FRR	***F**alse **R**ejection **R**ate*
GMM	***G**aussien **M**ixture **M**odel*
GMM	***G**aussian **M**ixture **M**odel*
IAL	***I**dentification **A**utomatique du **L**ocuteur*
LA	***L**ogical **A**ccess*
LA	***L**ogical **A**ccess*
MAP	***M**aximum **A** **P**osteriori*
MFCC	***M**el-**F**requency **C**epstral **C**oefficient*
PA	***P**hysical **A**ccess*
RAL	***R**econnaissance **A**utomatique du **L**ocuteur*
ROC	***R**eceiver **O**perating **C**haracteristic*
SV	***S**peaker **V**erification*
TVP	***T**aux de **V**rais **P**ositifs*

UBM ***U**niversal **B**ackground **M**odel*

VAL ***V**érification **A**utomatique du **l**ocuteur*

VC ***V**oice **C**onversion*

TTS ***T**ext- **T**o – **S**peech*

Introduction Générale:

Ces dernières années, la vérification automatique du locuteur (ASV) a été largement utilisée pour la biométrie vocale. Cela a conduit à se concentrer de plus en plus sur la sécurisation de ces systèmes biométriques vocaux pour des applications pratiques. Il existe quatre principaux moyens de générer des attaques : l'usurpation d'identité, la conversion vocale (VC), la synthèse vocale (TTS) et la relecture. Les attaques par usurpation d'identité ont reçu le moins d'attention ; les attaques réussies nécessitent souvent une expertise particulière, comme celle d'un imitateur professionnel.

La nécessité d'un ensemble commun de données, de protocoles et de mesures était la principale conclusion d'une session spéciale sur l'usurpation d'identité et les contre-mesures ASV lors de la conférence INTERSPEECH 2013 à Lyon, France. Pour améliorer les performances des systèmes de vérification du locuteur ASV, un ensemble de défis a été proposé visant à encourager le développement de contre-mesures (CM) pour empêcher les attaques d'usurpation de manipuler le système ASV. ASVspoof collecte et distribue de grandes bases de données réelles et falsifiées générées à l'aide de diverses techniques de pointe VC (Voice Conversion) et TTS (Text To Speech),de dispositifs matériels d'usurpation, d'enregistrement et de relecture.

La série de défisASVspoof (usurpation de vérification automatique du locuteur) a été créée pour faciliter ce développement. Il existe jusqu'à présent plusieurs défis de ce type :

Le premier défi ASVspoof organisé en 2015 présente un certain nombre de lacunes, dont certaines ont déjà été discutées lors de la session spéciale d'INTERSPEECH 2015.Premièrement, en termes d'attaques, ASVspoof 2015 se concentre sur TTS et VC, à l'exclusion des attaques par relecture (Replay Attaque). Contrairement aux TTS et VC à la pointe de la technologie, qui nécessitent des investissements importants, une attaque par relecture pourrait en principe être effectuée par toute personne utilisant un appareil grand public ordinaire pour diffuser du contenu et de l'audio vers un microphone du système ASV. Deuxièmement, en termes de qualité des données, le corpus SAS consiste en une parole propre à large bande passante sans variation extrinsèque des conditions. Troisièmement, ASVspoof 2015 est conçu sur la base du système ASV indépendant du texte, bien que les systèmes dépendants du texte soient les systèmes les plus couramment utilisés dans les scénarios d'authentification. Enfin, la métrique d'évaluation de l'ASVspoof 2015, le taux d'erreur égal moyen, ne tient pas compte du calibrage des contre-mesures dans différentes conditions ou attaques, puisqu'elle utilise un seuil (implicite) en fonction de l'attaque, qui n'est pas connu dans la réalité.

Le deuxième défiASVspoof2017 se concentre sur la détection d'usurpation autonome sans "intégration" avec un système ASV. Des protocoles standards comprenant des ensembles de données de développement et d'évaluation seront publiés. Le protocole prescrit simplement une liste de segments d'essai, chacun correspondant à un fichier audio nommé de manière aléatoire d'audio réel ou usurpé. Les participants ont attribué à chaque essai un score fini qui reflétait la valeur réelle reflétant la force relative de deux hypothèses concurrentes, c'est-à-dire si l'essai était un discours authentique ou usurpé. Après la publication de la base de données ASVspoof 2017 Les organisateurs ont remarqué de nombreuses anomalies dans les données. Ceux-ci ont été corrigés dans la deuxième version de la base de données publiée en février 2018(ASVspoof2017 V2)

L'édition 2019 a exploré les attaques par relecture à l'aide de dispositifs d'évaluation plus contrôlés et de conditions acoustiques soigneusement contrôlées sous la forme d'attaques par relecture simulées. Cette base de données est beaucoup plus grande que la base de données ASVspoof précédente et beaucoup plus diversifiée en termes d'algorithmes d'attaque. Elle est passée à la mesure du coût de la fonction de détection en tandem (t-DCF) alors que cette dernière reflète l'impact de l'usurpation et des contre-mesures sur les performances de l'ASV, Une bonne performance dépend de la généralisation, à savoir des contre-mesures qui donnent de bons résultats face à des attaques d'usurpation d'identité qui ne figurent pas dans les données de formation ou de développement.

Notre projet de fin d'étude s'inscrit dans le cadre de la reconnaissance robuste de la parole synthétisé. Grâce à cette technique, il est possible d'utiliser la voix des personnes pour vérifier leur identité et contrôler l'accès à des services tels que les achats par téléphone, les messageries, le démarrage automatique d'une voiture sécurisé par l'empreinte vocale. Nous avons utilisé le Modèle de mélange gaussien (GMM) et l'extraction de paramètres CQCC & MFCC pour développer le système.

Ce mémoire s'articule autour de trois Chapitres :

Dans le premier chapitre nous allons décrire d'une manière générale les mécanismes de production de la parole, les principes de l'audition et les propriétés générales des signaux vocaux. Dans ce contexte nous présenterons la numérisation d'un signal vocal, l'extraction des paramètres, les différents axes relevant du domaine de traitement de la parole avec un accent particulier sur la reconnaissance de la parole et du locuteur.

Le deuxième chapitre sera dédié à la reconnaissance et la vérification du locuteur(ASR et ASV) ainsi que les techniques utilisées dans ce domaine. La technique GMM (mélange de Gaussiennes) sera la plus étalée étant donné qu'elle constitue la méthode utilisée dans ce travail. Il sera question également de présenter les différentes métriques utilisées pour mesurer la performance de détection d'usurpation d'identité telles que le taux de fausse acceptation (FAR), de faux rejet (FRR) et le taux d'erreur égal (EER).

Le troisième chapitre concerne exclusivement l'aspect expérimental traitant l'utilisation de la technique GMM pour concevoir un système de détection d'usurpation d'identité en se basant sur le corpus ASV 2017 V2.

Une conclusion générale terminera naturellement se travail.

Chapitre I

Généralités Sur le traitement de la parole.

I.1 Introduction :

Ce chapitre aborde les aspects généraux en relation avec la parole. Il sera donc question d'exposer les notions de production et de perception de la parole ainsi que ses caractéristiques phonétiques et acoustiques. L'autre aspect important abordé sera celui du traitement du signal de parole à savoir la numérisation et l'extraction de paramètres ; à cet effet les deux techniques de base d'extraction des paramètres seront abordées : les coefficients MFCC et les coefficients CQCC. Enfin, les principales applications relevant du domaine de la parole seront abordées.

I.2 Définition de la parole :

La parole est la réalisation phonétique du langage résultant d'un acte psychophysiologique et volontaire de la part d'un individu. Aucun mode de communication animal ne peut correspondre à la complexité du langage parlé. Ceci est rendu possible par une anatomie particulière et par l'existence de régions spécialisées dans le cerveau [1].

I.3 La production de la parole :

Produire de la parole, c'est tout d'abord expulser de l'air des poumons ; ce flux d'air va être mis en forme par la multitude de cavités, constrictions et orifices qui marqueront son chemin depuis les poumons jusqu'aux lèvres et au nez [2].

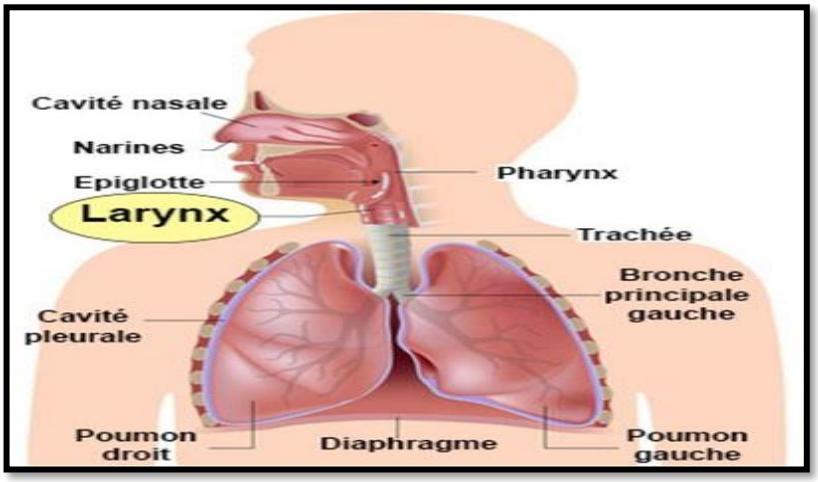


Figure I.1 : Modèle simplifié de l'appareil phonatoire

La Fig. I.1 représente schéma générale de l'appareil phonatoire humaine, On y voit, pour l'essentiel :

- Les cordes vocales : membranes musculeuses qui coiffent l'orifice de la trachée (larynx).Leur ouverture (glotte) présente une surface variable.

- Le conduit vocal : cavité de forme variable, allant des cordes vocales aux lèvres.sa longueur est d'environ 17.5 cm chez l'adulte. On y distingue souvent une cavité buccale et une cavité pharyngale.
- La cavité nasale : qui peut être mise en dérivation sur le conduit vocal, par l'abaissement du velum (comme c'est le cas sur la figure I.1).
- Les articulateurs : (langue, lèvres, mâchoires, …) qui permettent de modifier la forme du conduit vocal

I.3.1 Larynx :

Le larynx est un organe situé au fond de la gorge entre le pharynx et la trachée. Son rôle principal est d'assurer le passage de l'air vers les poumons lors de l'inspiration ou vers l'extérieur lors de l'expiration. Il contient également une membrane qui se ferme automatiquement lors de la déglutition lorsque les aliments pénètrent dans l'œsophage par la gorge, empêchant les aliments de pénétrer dans les voies respiratoires (comme la trachée).

Le larynx joue également un rôle important dans la production de la parole. Il contient des cordes vocales et des muscles qui modifient l'ouverture de la lumière laryngée pour moduler le son produit. La Figure I.2 donne un aperçu de cet organe

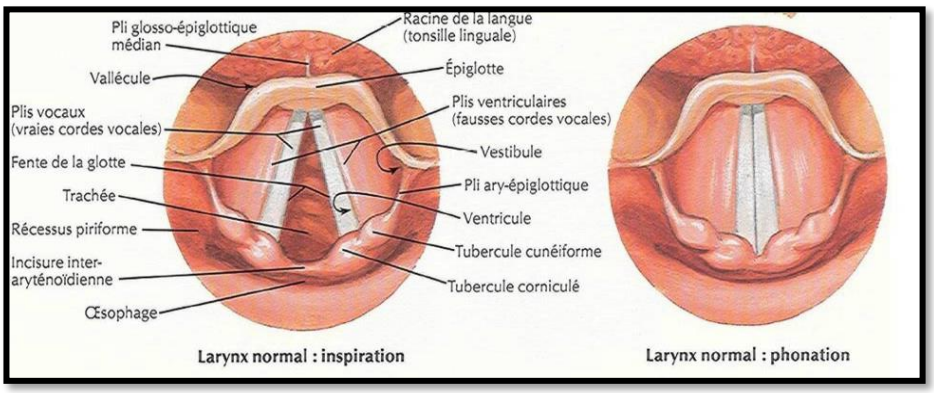


Figure I.2: Le larynx

I.3.2 Pharynx :

Remontant, après que l'air ait traversé les plis vocaux et presque a immédiatement quitté le larynx, le pharynx commence qui a une forme irrégulière avec de nombreuses sections. C'est essentiellement le début de ce qu'on appelle le tractus vocal puisque tout contrôle d'articulation commence ici. Selon la quantité d'air passant par les plis vocaux et leur état de

tension, l'air laminaire vibrant ou turbulent passera par les différentes sections du pharynx pour produire différents types de sons [3].

I.4 La perception de la parole :

L'appareil phonatoire, émetteur d'informations, ne serait d'aucune utilité si l'information générée ne pouvait être captée et analysée par un récepteur. Parmi tous les récepteurs existants, l'homme a acquis la capacité de découvrir le sens caché sous les sons produits par son interlocuteur. Nous allons présenter l'anatomie de l'oreille, organe récepteur de l'information sonore, et les capacités de perception qui caractérisent cet organe lorsqu'il est en parfait état et n'a subi aucune atteinte venue amoindrir ses capacités intrinsèques [4].

I.4.1 L'appareil auditif humain :

La partie mécanique de l'audition (l'oreille) est composée de trois sections. La première section, oreille externe, est la combinaison des cartilages dans l'oreille et le meatus acoustique externe (le canal auditif).

Deuxièmement, l'oreille moyenne comprend la membrane tympanique (tympan) et tout ce qui se trouve à sa droite, comme le montre la figure I.3. C'est principalement la membrane elle-même et les trois os spéciaux qui transfèrent le mouvement de la membrane tympanique induit par les ondes sonores qui traversent et qui sont amplifiées par l'oreille externe.

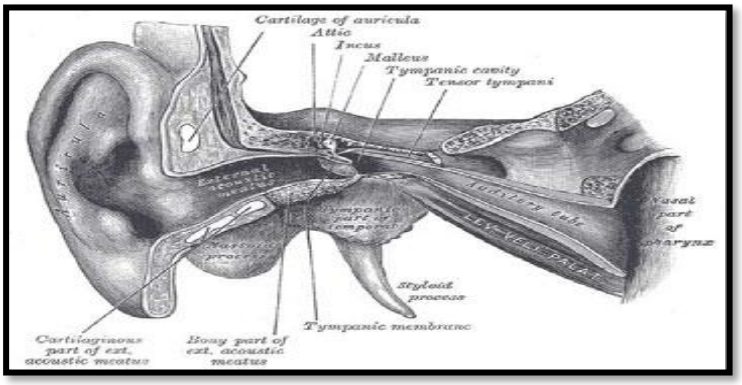


Figure I.3 : L'oreille externe et le milieu d'oreille.

Enfin, l'oreille interne (figure 1.4) est constituée de la cochlée, une cavité ressemblant à un escargot, et trois canaux semi-circulaires appelés l'ampoule supérieure, l'ampoule

antérieure, et l'ampoule postérieure. La cochlée et les trois canaux semi-circulaires sont remplis d'un fluide incompressible [3].

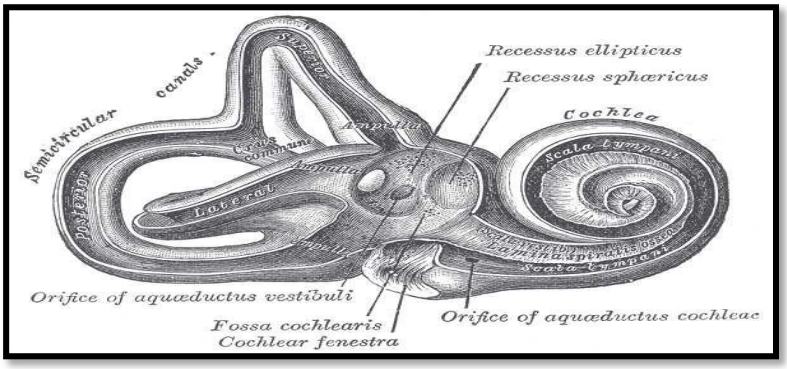


Figure I.4: L'oreille interne

I.5 Le système nerveux et le cerveau :

Illustré sur la figure I.5, Le système nerveux central se compose de deux parties principaux, la moelle épinière et le cerveau. Le cerveau est divisé en trois divisions principaux : le cerveau antérieur (cerveau humain), le mésencéphale (le mésencéphale) et le cerveau postérieur (céphale rhombique).Le cerveau contient de nombreuses parties, dont la plus importante est le cortex cérébral, qui est la couverture grise des tissus à la surface du cerveau. Ila des fonctions telles que la pensée, la production et la compréhension du langage, la vision et l'ouïe.

Le cortex cérébral est divisé en quatre lobes séparés par de grandes fissures. Le lobe occipital est situé derrière le cortex cérébral et est principalement responsable des fonctions visuelles. Le lobe temporal inférieur effectue mieux le traitement de l'information visuelle, le traitement de la mémoire, le traitement de la parole et leurs relations.

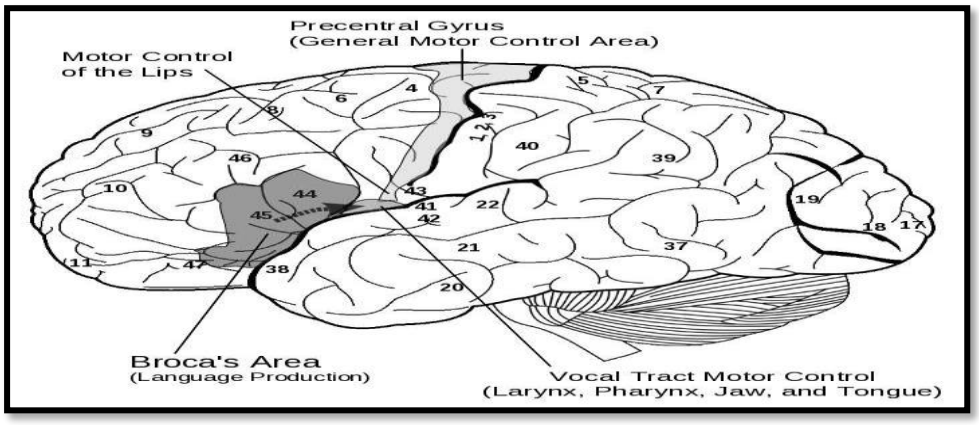


Figure I.5: Domaines de la production vocale dans le cerveau.

I.6 Les Caractéristiques phonétiques de la parole :

I.6.1 Les voyelles :

Si le conduit vocal est suffisamment ouvert pour que l'air expulsé par les poumons le traverse sans obstacle, il y a production d'une voyelle. Le rôle de la bouche se réduit alors à une modification du timbre vocalique. En effet, la voyelle est produite avec un canal aérifère ouvert, sans constriction majeure et en l'absence de génération de bruit de friction. Le paramètre caractéristique des voyelles est la position des formants (maximas de spectre), la figure I.6 montre la distribution des voyelles dans le plan F1,F2 (premier et deuxième formant).

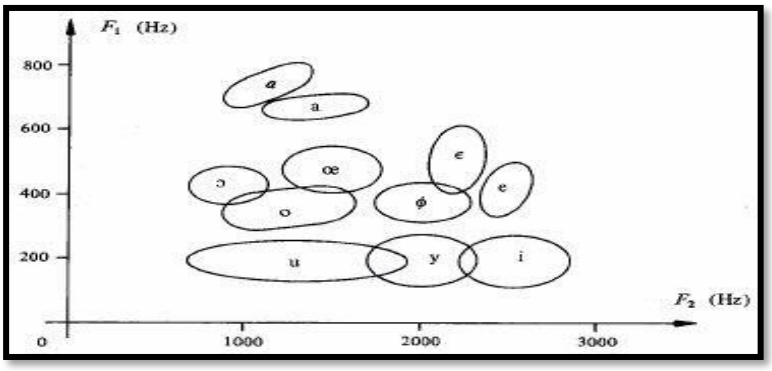


Figure I.6 : Représentation des voyelles dans le plan F1 F2

I.6.2 Les consonnes :

Lorsque le conduit vocal au passage de l'air se rétrécit par endroits, ou même s'il se ferme temporairement, le passage forcé de l'air donne naissance à un bruit : une consonne est produite. On classe principalement les consonnes en fonction de leur mode d'articulation, de leur lieu d'articulation, de leur voisement et leur nasalisation Comme pour les voyelles, d'autres critères de différenciation peuvent être nécessaires dans un contexte plus général : l'organe articulatoire, la source sonore, l'intensité, l'aspiration, la palatalisation, et la direction du mouvement de l'air.

I.7 Le signal de la parole :

La parole est considérée, depuis la nuit des temps, comme étant le moyen le plus simple, le plus naturel et le plus pratique pour communiquer. La parole est avant tout un signal acoustique plongé dans un espace de communication Linguistique. La communication parlée n'est donc possible qu'entre deux locuteurs ayant un même espace de communication

linguistique, formés à coder et à décoder un sens au moyen de la même clé. En effet, un message émis d'une source, n'est compris que si le récepteur est en possession de la clé de l'espace de communication.[5]

I.8 Caractéristiques fréquentielles du signal de parole :

Les caractéristiques les plus importantes sont la fréquence fondamentale et les formants

I.8.1 La fréquence fondamentale F0 (Le pitch) :

Le pitch est une quantité perçue qui est liée à la fréquence fondamentale de vibration des cordes vocales sur une certaine durée. C'est une fonction de leur forme et de leur tension, des caractéristiques du flux d'air et de la configuration des voies vocales au moment de l'expression.[3]

Les différentes variations de la fréquence fondamentale F0, de l'intensité et de la durée, peut faire ressortir bien des caractéristiques du locuteur, ou bien même pour caractériser les différences lexicales entre les mots F0 moyen-homme 100 à 150 Hz, F0 moyen-femme 200 Hz à 300 Hz et F0 moyen enfant 350 à 400 Hz

L'extraction du pitch est une tâche particulièrement difficile pour trois raisons :

- La vibration des cordes vocales n'a pas nécessairement une périodicité complète.
- Il est difficile de séparer le pitch des effets du trait vocal.
- La plage de dynamique de la fréquence fondamentale est très grande.[6]

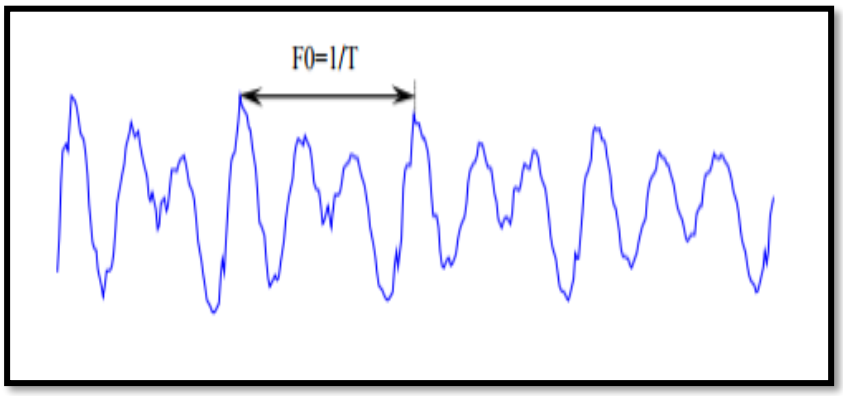


Figure I.7: Fréquence fondamentale.[6]

I.8.2 Les formants :

Un formant est un renforcement spectral créé par une cavité du conduit vocal (telle que la bouche, le pharynx…) agissant comme un résonateur. Les formants décrivent les structures spectrales des sons voisés qui nous permettent d'identifier le type du son prononcé en particulier pour les voyelles.

En outre, il est bien établi que les formants sont liés à la forme géométrique du conduit vocal. Afin d'améliorer les modèles de production et d'étudier l'identification des sons, il est important d'avoir une estimation précise des formants et de leurs transitions

Donc le conduit vocal possède une fonction de transfert (filtre) appliquée à une source produit un phonème, on appelle formant les maxima locaux (pics) de cette fonction de transfert notés F1, F2, F3… et correspondent aux zones de renforcement maximal. Néanmoins, du point de vue perceptif, seuls quelques formants jouent un rôle central au niveau de la parole. En théorie on décrit souvent les voyelles grâce à leurs 2 premiers formant F1, F2 à travers un triangle acoustique

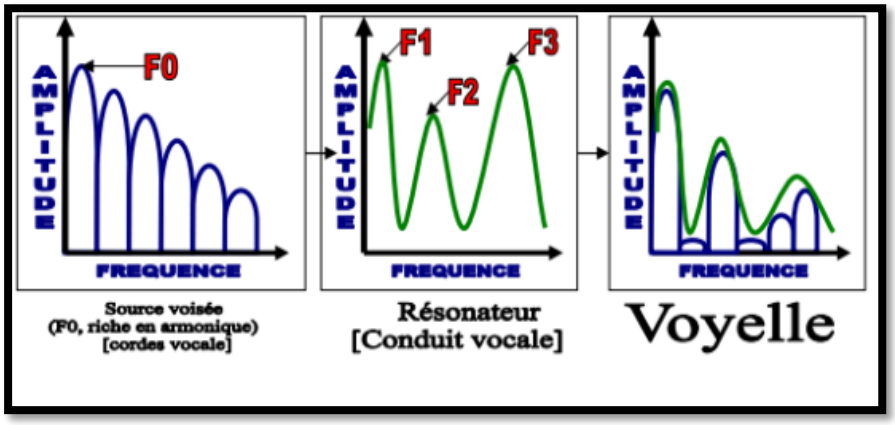


Figure I.8: la production d'une voyelle

I.9 La Numérisation du signal de la parole :

La numérisation consiste à transformer un signal analogique qui contient une quantité infinie d'amplitudes en un signal numérique contenant, lui, une quantité finie de valeurs.

Le passage de l'analogique au numérique repose sur trois étapes successives : l'échantillonnage, la quantification, et le codage.

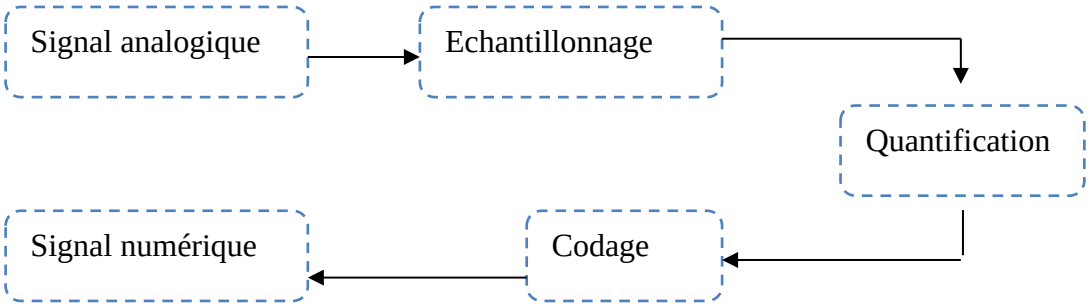


Figure I.9: Les étapes de numérisation d'un signal analogique.

I.9.1 Echantillonnage :

Le point de départ naturel est d'échantillonner le signal audio analogique à traiter plus tard. Dans le traitement de la parole, nous utilisons généralement un échantillonnage périodique dans lequel la fréquence d'échantillonnage (taux d'échantillonnage) est fixe. il y a une question fondamentale : quelle période d'échantillonnage (ou fréquence d'échantillonnage) devrions-nous utiliser? Pour répondre à cette question très importante, examinons un théorème fondamental dans la théorie de l'information et le traitement du signal appelé, le théorème d'échantillonnage.[3]

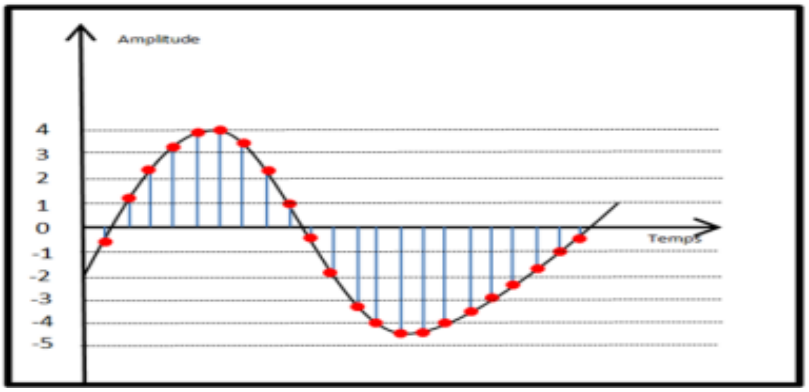


Figure I.10: signal échantillonné

I.9.1.1 Théorème d'échantillonnage :

« La fréquence d'échantillonnage assurant un non repliement du spectre doit être supérieure à 2 fois la fréquence haute du spectre du signal analogique fe >= 2 fmax »

Pour la téléphonie, on estime que le signal garde une qualité suffisante lorsque son spectre est limitée à 3400 Hz et l'on choisit fe = 8000 Hz.

Pour les techniques d'analyse, de synthèse ou de reconnaissance de parole, la fréquence peut varier de 6000 à 16000 Hz. Puisque fm = 5 KHz alors fe = 10 KHz et donc il va falloir mesurer le signal tous le $\mathbf{1}$ / $\mathbf{10000}$ $\boldsymbol{seconde}$ = $\mathbf{0}$,1 $\boldsymbol{ms}$. [7]

I.9.2 La Quantification :

Cette étape consiste à approximer les valeurs réelles des échantillons selon une échelle à n niveaux appelée échelle de quantification. Par conséquent, l'échantillon quantifié a 2^n valeurs possibles entre $-2^n - 1$ et $2^n - 1$. L'erreur systématique causée par l'assimilation de la valeur réelle de l'écart au niveau de quantification le plus proche est appelée bruit de quantification. Ce procédé est illustré dans la figure I.11.

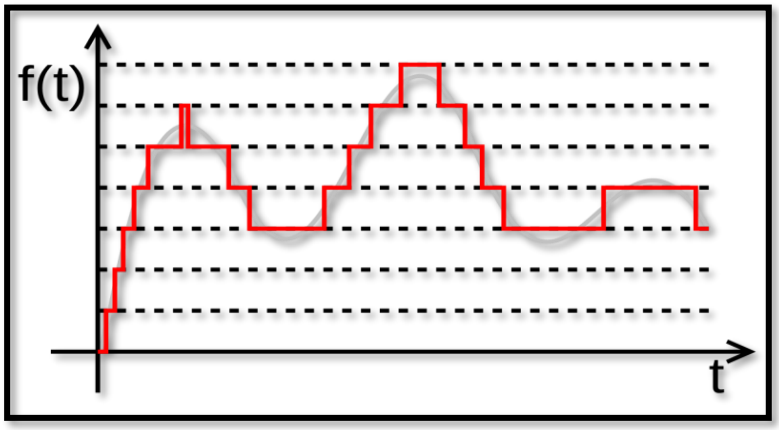


Figure I.11: Signal quantifié

Les deux étapes : échantillonnage, quantification représentent en fait une discrétisation dans le temps et dans l'amplitude respectivement.

I.9.3 Le codage :

Les valeurs quantifiées du signal sont transformées (codées) en valeurs binaires (usuellement 16 bits). Une représentation linéaire des valeurs quantifiés permet le traitement du signal sur machine.[7]

I.10 Domaines d'application :

Les domaines d'application sont vastes et variés, on peut en dresser une liste non exhaustive :

- Reconnaissance automatique du langage.
- Reconnaissance automatique de la parole.
- Reconnaissance et vérification du locuteur.
- Protection contre l'usurpation d'identité.

- Reconnaissance d'émotion.
- Synthèse de la parole.

I.10.1 La reconnaissance automatique de langage :

Une langue est définie comme un regroupement de dialectes partageant un vocabulaire similaire et ayant des systèmes phonologiques et grammaticaux similaires. L'identification automatique de langage a pour rôle de reconnaitre la langue parlée par un locuteur inconnu, parmi un ensemble fini de langues. Ainsi cette tache joue un rôle très important dans la consultation de bases de données documentaires multilingues (notamment sur internet), l'enseignement ou encore la traduction automatique et autant d'autres applications issus de l'ouverture mondiale des télécommunications.[8]

Quatre domaines linguistiques représentent la source d'informations discriminantes pour la reconnaissance des langues :

- La prosodie : Le rythme, l'intonation et l'accentuation varient suivant les langues.
- La phonologie : science qui étudie les phonèmes non en eux-mêmes, mais quant à leur fonction dans la langue [9]
- La morphologie : c'est l'étude des formes des mots (flexion et dérivation) et chaque langue a son propre vocabulaire et sa propre manière de former.
- La syntaxe : c'est l'étude de la façon dont les mots d'une langue donnée peuvent être associés pour former des phrases ou des énoncés Le fonctionnement chaque système de reconnaissance, le système d'identification de la langue se décompose en deux phases (figure I.12).

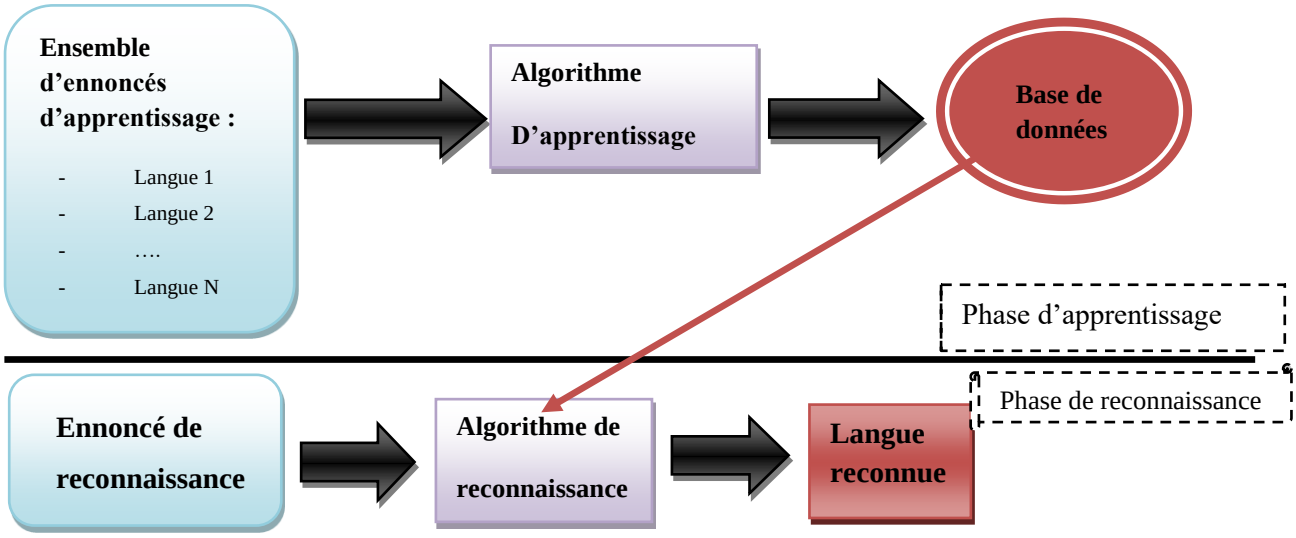


Figure I.12: Structure d'un système d'identification automatique des langues.

La conception du système passe par deux phases : apprentissage, reconnaissance

Pendant l'apprentissage, des paramètres sont extraits pour les signaux de parole de chaque langue. Pour chaque source d'information prise en compte, un modèle spécifique à chaque langue est appris à partir de ces paramètres.

Pendant la reconnaissance, des paramètres sont extraits pour un signal de parole d'une langue inconnue. La langue la plus vraisemblable est déterminée en fonction des modèles issus de la phase d'apprentissage.[10]

I.10.2 Reconnaissance de la parole :

La reconnaissance automatique de la parole permet à la machine de comprendre et de traiter des informations fournies oralement par un utilisateur humain. Elle consiste à employer des techniques d'appariement afin de comparer une onde sonore à un ensemble d'échantillons, composés généralement de mots mais aussi, plus récemment, de phonèmes (unité sonore minimale).

En reconnaissance, il existe un problème supplémentaire qui est lié au fait qu'il n'existe pas un appareil phonatoire humain unique et universel, mais qu'au contraire chaque homme possède des cordes vocales et un conduit vocal unique qui peut s'avérer très différents de ceux de son voisin. Il est ainsi probable que le signal de parole tel qu'il sera capté par un microphone renfermera une grande variabilité suivant les personnes. On s'aperçoit en fait qu'il existe plusieurs niveaux de variabilité :

- La variabilité intra-locuteur : qui représente la variabilité de la parole d'un même locuteur au cours du temps. Cette variabilité dépend d'un grand nombre de paramètres tels que la force de la voix, l'état physique (voix enrouée) et de l'état émotionnel (fatigue, colère, excitation).
- La variabilité interlocuteur : qui représente la variabilité entre les différents locuteurs dues aux différences physiologiques, de style d'élocution, d'accents régionaux, etc….

La conception d'un tel système passe par les mèmes étapes d'apprentissage et de reconnaissance expliquées plus haut.

I.10.3 Reconnaissance et vérification du locuteur :

La structure d'un tel système est similaire à celle du système de reconnaissance de la parole, sauf que l'objectif ici est de déterminer ou de vérifier l'identité du locuteur dans le but de permettre à un utilisateur particulier d'accéder à un service particulier offert par le système.

Une attention particulière a été attachée à ce domaine ces dernières années donnant naissance à un créneau de recherche ayant pour but de protéger les systèmes vocaux des attaques par usurpation d'identité. Le système à concevoir contient deux barrières : une première pour la vérification de l'identité du locuteur et une deuxième pour la vérification de l'authenticité du signal traité par le système (signal naturel ou non). C'est d'ailleurs dans ce cadre que s'inscrit notre travail.

I.10.4 La synthèse de la parole :

Il s'agit de générer un signal de parole par une machine (parole artificielle). Les applications sont nombreuses, on peut citer par exemple les horloges parlantes ou la conversion d'un texte vers de la parole (TTS : Text To Speech).

I.11 Extraction de paramètres :

Dans la grande majorité des cas, les systèmes exposés plus haut dans ce chapitre n'utilisent pas le signal de parole sous sa forme temporelle brute. En effet une phase d'extraction de paramètres est nécessaire pour extraire les paramètres informatifs pertinents permettant au système de bien fonctionner.

Deux techniques vont être exposées ici à savoir : l'extraction des coefficients MFCC et l'extraction des coefficients CQCC.

I.11.1 Extraction des paramètres MFCC :

La parole peut être représentée sous la forme d'un modèle source-filtre. Cette représentation permet ainsi de représenter le signal de parole s(t) sous la forme de la convolution du signal source g (t) par la réponse impulsionnelle du filtre h(t) représentant le conduit vocal [11] :

$$\boldsymbol{s(t) = g(t) * h(t)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{I.1})$$

L'étude de ce signal à l'aide de la FFT présente un défaut particulier liée à cette convolution qui rend difficile l'observation de la seule contribution du conduit vocal.

Le cepstre (parfois appelé lissage cepstral) permet de séparer les contributions respectives de la source et du conduit vocal. En effet, l'équation (1) se récrit dans le domaine spectral sous la forme [11] :

$$\boldsymbol{S(\omega) = G(\omega)H(\omega)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{I.2})$$

Où S(ω), G(ω) et H(ω) représentent respectivement les transformées de Fourier de s(t), g(t) et h(t). Le cepstre qui est défini par la transformée de Fourier inverse de la logarithme du module de S(ω) s'écrit donc sous la forme :

$$\boldsymbol{c(\tau) = FFT^{-1} log|S(\omega)| = FFT^{-1} log|G(\omega)| + FFT^{-1} log|H(\omega)|} \ldots\ldots\ldots\ldots\ldots (\mathbf{I.3})$$

On peut alors noter que le spectre s'exprime comme la somme de deux termes. Le premier terme FFT−1 log |G(ω)| est caractéristique de la source et représente ainsi la structure fine, tandis que le second terme est caractéristique de l'enveloppe spectrale et représente la contribution du conduit vocal. Le paramètre τ homogène à un temps est appelé quéfrence. A l'aide de cette représentation, il est possible d'isoler soit le pic (qui correspond au pitch) qui se trouve dans la région des hautes quéfrences (on a ici une méthode d'estimation de la fréquence fondamentale) soit d'isoler la partie correspondant aux basses quéfrences qui représente une version lissée de l'enveloppe spectrale. [11]

Ce procédé de séparation des éléments cepstraux est appelé un liftrage (par dérivation de l'appellation filtrage).Lorsque le cepstre est obtenu en calculant la transformée de Fourier discrète, on obtient la forme suivante :

$$\boldsymbol{c_n = \frac{1}{N}\sum_{k=0}^{N-1} log|X(k)|e^{2j(\pi)kn/N} \quad , 0 \le n \le N-1} \ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{I.4})$$

Le procédé est illustré sur le schéma de la figure I.13

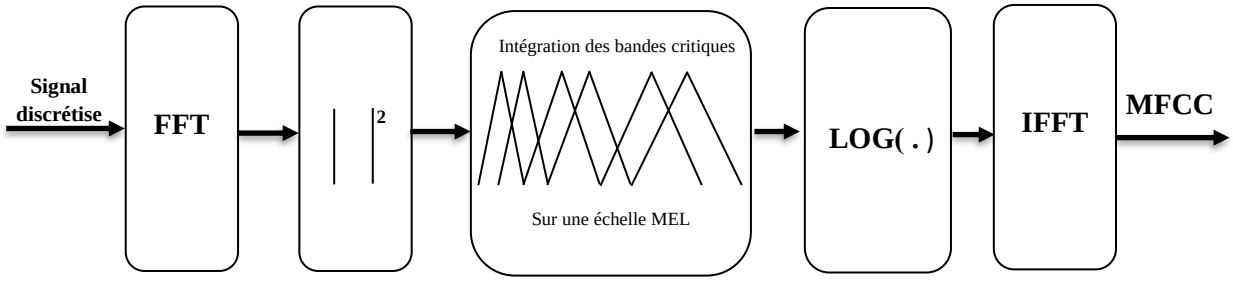


Figure I.13 : Calcul des coefficients MFCC [12]

Une représentation améliorée peut être obtenue en incluant des informations sur l'évolution temporelle des coefficients cepstraux. Ceci peut être déterminé, par exemple, en utilisant les dérivées première et seconde des coefficients du cepstre. Soit cm(t) le coefficient de cepstre pris au temps t (ou plus précisément une fenêtre d'indice t). Étant donné que cette séquence est obtenue à des instants discrets.

Il est bien connu qu'une simple moyenne de différence ne donne pas une estimation sans bruit. Les dérivations sont donc souvent obtenues en faisant la moyenne sur une plus longue période de temps sous la forme de [11]:

$$\Delta c_m(t) \approx \mu \sum_{k=-K}^{K} k c_m(t+k) \quad \text{(I. 5)}$$

Usuellement on prend les dérivées d'ordre 1 et 2 des coefficients statiques obtenus, ainsi si le nombre de coefficients statiques est 13 (valeur couramment adoptée) alors on aura 13 autres coefficients représentant la dérivée d'ordre 1 et 13 autre représentant la dérivée d'ordre 2. On aura donc au total 39 coefficients pour chaque trame du signal.

I.11.2 Extraction des paramètres CQCC :

CQCC utilise un rééchantillonnage uniforme pour convertir le spectre de puissance d'octave en un spectre de puissance linéaire, puis effectue une transformée en cosinus discrète (DCT) sur le spectre de puissance linéaire pour obtenir CQCC. La raison derrière cela est que la DCT ne peut pas être directement appliquée aux spectres de puissance d'octave car chaque bande de fréquence a une bande passante différente. Cependant, nous pensons que la DCT peut être appliquée à l'ensemble du spectre de puissance d'octave pour décorréler les caractéristiques. Nous notons que les spectres de puissance d'octave ne fournissent pas le même niveau de détail que les spectres de puissance linéaire, mais les spectres de puissance d'octave et les spectres de puissance linéaire peuvent fournir des informations complémentaires. De plus, le spectre de puissance d'octave peut refléter certaines caractéristiques du système auditif humain, Par exemple, les basses fréquences donnent une meilleure résolution en fréquence et les hautes fréquences donnent une meilleure résolution temporelle, contrairement aux spectres linéaires. Nous supposons que l'information spectrale d'octave peut être combinée avec l'information obtenue par échantillonnage uniforme [13]. Le schéma de la figure I.14 illustre le procédé d'extraction des coefficients CQCC.

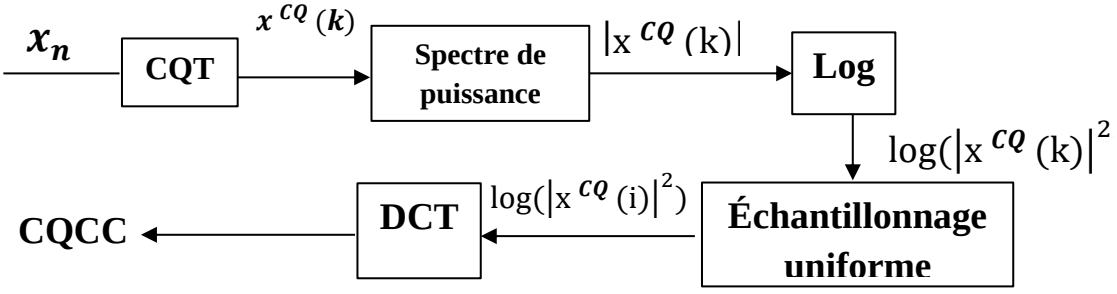


Figure I.14: Diagramme schématique pour l'extraction des coefficients (CQCC).

La CQT est différente de la DFT car le rapport entre la fréquence centrale et la bande passante est constant dans CQT. En conséquence, CQT a une résolution de fréquence plus élevée en basse fréquence et une résolution temporelle plus élevée pour une fréquence plus élevée[11]. Pour un signal x(n), son CQT Y(k, n) est défini comme :

$$\boldsymbol{Y(k,n) = \sum_{j=n-\frac{N_k}{2}}^{n+\frac{N_k}{2}} x(j) a_k^* \left(j - n - \frac{N_k}{2}\right)} \ldots\ldots\ldots\ldots (\mathbf{I.6})$$

Où k = 1, 2, ..., K est l'indice bin de fréquence, Nk sont les longueurs de fenêtre variable, $a_k^*(n)$ désigne le complexe conjugué $a_k(n)$ qui est définis par :

$$\boldsymbol{a_k(n) = \frac{1}{C} v\left(\frac{n}{N_k}\right) e^{i(2\pi n \frac{f_k}{f_s} + \emptyset_k)}} \ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{I.7})$$

Oùf_k est la fréquence centrale de bin f_k , f_s est la fréquence d'échantillonnage, et v(t) est une fonction de fenêtre (par ex : fenêtre Hanning)et $\emptyset_k$ est un décalage de phase. Le facteur d'échelle C est calculé comme suit :

$$\boldsymbol{C = \sum_{m=-\frac{N_k}{2}}^{\frac{N_k}{2}} v\left(\frac{m + \frac{N_k}{2}}{N_k}\right)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{I.8})$$

Comme pour les coefficients MFCC, les dérivées d'ordre 1 et 2 des coefficients statiques sont souvent ajoutés.

I.12 Conclusion :

Ce chapitre ne se veut en aucun cas avoir la prétention de couvrir tous les aspects en relation avec le domaine de traitement de la parole. Nous avons ainsi passé en revue dans ce chapitre quelques notions fondamentales sur le signal de la parole à commencer par son origine de production (appareil phonatoire), sa perception (système auditif), les différents domaines d'application et enfin les procédés usuels d'extraction de paramètres. Parmi les applications exposées dans ce chapitre, la protection contre l'usurpation d'identité constitue l'objet de ce travail. Le deuxième chapitre sera donc entièrement dédié à cet aspect.

Chapitre II

Reconnaissance et vérification du locuteur

II.1 Introduction :

La reconnaissance du locuteur fait généralement référence à la fois à l'identification du locuteur et la vérification du locuteur. Un système d'identification du locuteur identifie qui est le locuteur, tandis qu'un système de vérification automatique du locuteur (ASV) décide si une déclaration d'identité est vraie ou fausse. Le premier est un problème de classification multi-classes, tandis que le second est un test d'hypothèse. Un système ASV général est robuste aux imposteurs sans effort, ils sont vulnérables aux attaques plus sophistiquées.

Une telle vulnérabilité représente l'un des problèmes de sécurité des systèmes ASV. L'usurpation d'identité implique un adversaire (attaquant) qui se fait passer pour le locuteur cible pour accéder à un système.

Dans ce chapitre, nous présenterons la RAL (identification et vérification) et ses branches, le mélange de Gaussiennes (GMM) ainsi que la détection d'usurpation.

II.2 La reconnaissance automatique du locuteur (RAL) :

La reconnaissance automatique du Locuteur RA s'inscrit dans le domaine du traitement de la parole, c'est une technique informatique qui permet d'analyser un mot ou une phrase captée au moyen d'un microphone pour la transcrire sous la forme d'un texte exploitable par une machine.

Grâce à cette technique, il est possible d'utiliser la voix des personnes pour vérifier leur identité et contrôler l'accès à des services tels que les achats par téléphone, les messageries, le démarrage automatique d'une voiture sécurisée par l'empreinte vocale du propriétaire [14] ou encore, pour résoudre des investigations criminelles en trouvant lequel des suspects correspond aux enregistrements trouvés sur la scène du crime par exemple [15].

Les trois modules principaux qui constituent le fondement de base d'un système de reconnaissance du locuteur [16], sont résumés sur la (figure II.1) :



Figure II.1 : schéma simplifié le système d'avoir reconnaissance du locuteur.

- Prétraitement et paramétrisation :

Le prétraitement permet quelque conditionnement du signal tel que la normalisation de l'amplitude, la suppression des silences.

Ceci rendra le signal propice à la phase de paramétrisation qui consiste à extraire les paramètres acoustiques de chaque locuteur [17].

- Modélisation :

Ce module va créer un modèle de chaque personne à partir des vecteurs de caractéristiques obtenue précédemment.[17]

- Reconnaissance :

Ce module, pendant la phase de test du système, va tester les nouveaux sujets sur les modèles existants et stockés dans une base de données lors de la phase d'apprentissage du système afin de vérifier ou d'identifier le locuteur qui a parlé.[17]

II.2.1 Les branches de RAL :

II.2.1.1 Identification automatique du locuteur (IAL):

L'identification automatique du locuteur (IAL) a été une des premières utilisations delà RAL. En IAL, la liste des locuteurs à identifier est connue du système. Le système doit pouvoir décider, à partir d'un échantillon de voix, à quelle identité connue du système correspond l'échantillon [18]. La figure II.2 illustre l'architecture d'un tel système.

L'identification automatique du locuteur se découpe en deux étapes :

Une étape d'apprentissage et une étape de test : À partir d'un ensemble d'enregistrements de voix de chaque locuteur, le système apprend un modèle pour chaque locuteur lors de l'étape d'apprentissage. Lors de l'étape de test, le système confrontera l'échantillon de voix qu'il recevra aux différents modèles qu'il aura déjà appris afin de déterminer si l'identité du locuteur est déjà connue

Deux modes sont proposés en IAL :

- L'identification en ensemble fermé :

Dont on suppose que la séquence de parole est effectivement prononcée par un locuteur.

- L'identification en ensemble ouvert :

Le système doit décider de la fiabilité de son jugement en acceptant ou rejetant l'identité qu'il a trouvée [19].

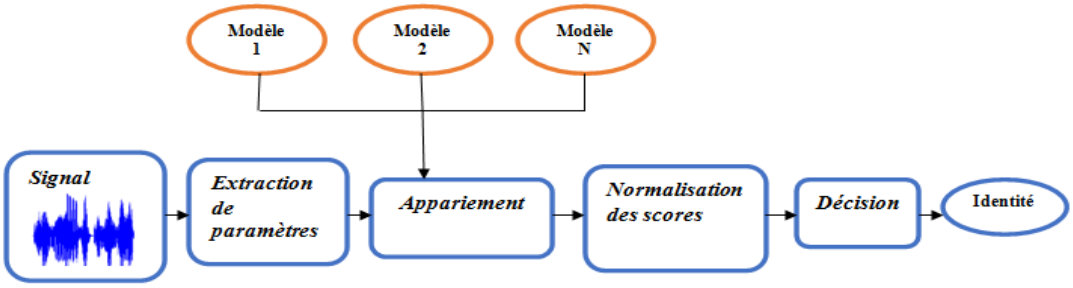


Figure II.2: Architecture d'un système d'identification du locuteur.

II.2.1.2 Vérification automatique du locuteur (VAL) :

La Vérification Automatique du Locuteur (VAL) est le processus décisionnel permettant de déterminer, au moyen d'un message vocal, la véracité de l'identité revendiquée par un individu dont la figure II.3 représente le principe de VAL [20, 21]. L'identité ainsi que le message vocal constituent les deux entrées du système de VAL. L'identité, nécessairement connue du système, désigne automatiquement la référence caractéristique d'un locuteur. Une mesure de similarité basée sur le rapport de vraisemblance est calculée entre cette référence et le message vocal puis comparée à un seuil de décision.

Dans le cas où la mesure de similarité est supérieure au seuil, l'individu est accepté. Dans le cas contraire, l'individu est considéré comme un imposteur et sera rejeté.

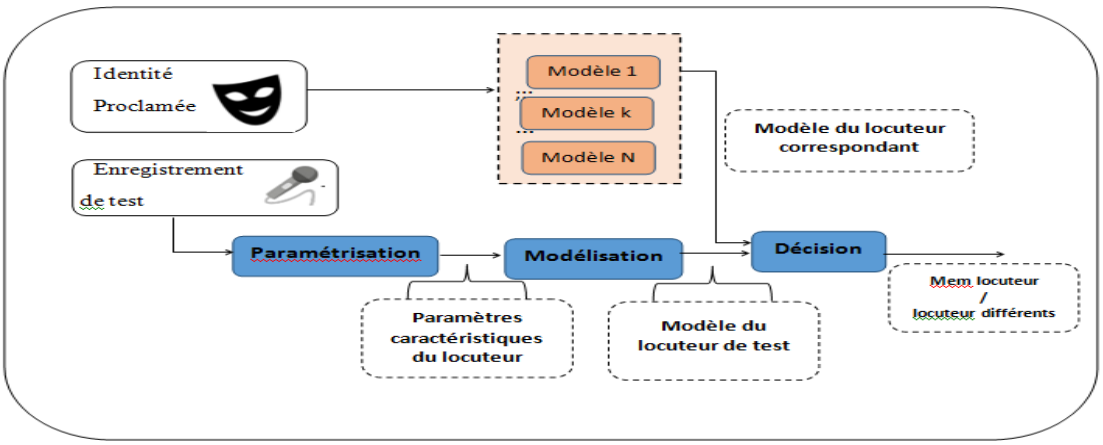


Figure II.3 : Structure d'un système de vérification du locuteur

II.2.2 Classification automatique du locuteur :

L'objectif de la classification est un peu plus vague. C'est l'étiquette générale pour toute technique qui regroupe des signaux audios similaires dans des bacs individuels. La classification selon le sexe, la classification selon l'âge et la classification des événements sont quelques exemples des nombreux scénarios de classification. La classification par sexe, comme son nom l'indique, tente de séparer les locuteurs masculins des locuteurs féminins. Les versions plus avancées distinguent également les enfants et les placent dans une corbeille séparée ; classer un homme et une femme n'est pas si simple chez les enfants puisque leurs caractéristiques vocales sont assez semblables avant le début de la puberté. [3]

II.2.3 Segmentation automatique du locuteur :

La segmentation automatique d'un flux audio en parties contenant la parole de locuteurs distincts, la musique, le bruit et différentes conditions de fond a de nombreuses applications. Ce type de segmentation est élémentaire aux considérations pratiques de la reconnaissance des locuteurs ainsi que de la parole et d'autres systèmes de reconnaissance audio. Différents enregistreurs spécialisés peuvent être utilisés pour la reconnaissance de catégories distinctes d'audio dans un flux. Une émission de nouvelle radio typique contient des signaux vocaux et non vocaux provenant d'une grande variété de sources comme la parole propre, la parole à bande limitée comme les sources téléphoniques, les segments musicaux, la parole sur la musique, la parole sur le bruit ambiant, parole sur parole, etc. Le défi de segmentation est de pouvoir séparer le discours produit par différents intervenants les uns des autres. [3]

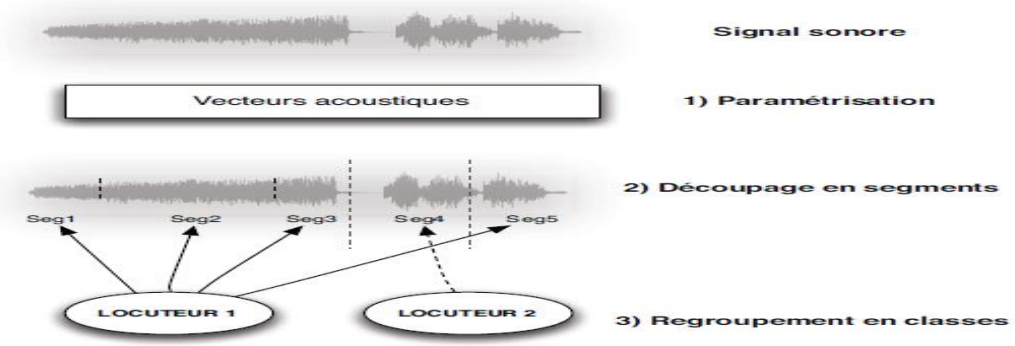


Figure II.4 : Principe de la segmentation et classification en locuteur

II.2.3.1 Suivi de locuteur :

Le suivi des locuteurs est quelque peu semblable à la détection des locuteurs, avec la différence subtile qu'un ou plusieurs des locuteurs sont suivis dans le flux.

Dans ce cas, on peut imaginer des conditions où aucune donnée d'inscription n'est disponible, mais non seulement l'audio est segmenté en segments de locuteur unique, mais les segments sont également étiquetés avec des étiquettes signifiant les locuteurs individuels dans le flux. Si les données d'inscription sont disponibles, les étiquettes de locuteur pour les segments peuvent être ajustées pour refléter les vrais locuteurs de ces segments dans la base de données d'inscription. Dans la plupart des cas, tant que les différents locuteurs du flux sont identifiés par des étiquettes générales telles que des étiquettes alphanumériques, l'objectif est atteint. L'application la plus importante du suivi est l'étiquetage des conférenciers dans une conversation comme une conférence téléphonique. [3]

II.2.4 Modalités de la RAL :

Selon leur application, les méthodes de reconnaissance du locuteur se divisent en deux :

II.2.4.1 La méthode indépendante du texte :

La reconnaissance des locuteurs indépendants du texte est la plus polyvalente des modalités. C'est aussi la seule modalité viable qui peut être utilisée dans toutes les branches de la reconnaissance des orateurs. Il y a différents degrés d'indépendance du texte. Certains enregistreurs sont complètement indépendants du texte et de la langue. Il y a des moteurs qui dépendent un peu de la langue.

Certains sont complètement dépendants de la langue, mais le texte est indépendant dans une certaine mesure. Elle est aussi indépendante du texte souffre également du problème d'évaluation de la vivacité. Dans ce cas-ci, il n'est même pas nécessaire que la personne prononce des mots précis pour pouvoir fausser le système. Tout type d'enregistrement de haute qualité de la voix de l'individu sera suffisant. [3]

II.2.4.2 La méthode dépendante du texte :

La modalité dépendante du texte s'applique uniquement à la branche de vérification des locuteurs. La plupart des autres branches ne peuvent pas être utilisées avec des phrases spécifiques puisqu'elles se produisent d'une manière plus passive où un système de reconnaissance écoute une parole et prend une décision.[3]

II.2.5 La variabilité de la voix :

On distingue les variabilités inter et intra-locuteur :

II.2.5.1 La variabilité interlocuteur :

Ces variabilités proviennent des différences physiologiques (différences dimensionnelles du conduit vocal, fréquences d'oscillations des cordes vocales) et de différences de style de prononciation (p.ex. accent, niveau social). Certaines de ces différences qui influencent la représentation de chaque locuteur, nous permettent de les séparer.

II.2.5.2 La variabilité intra-locuteur :

Variabilités intra-locuteur font que la voix dépend de l'état physique et émotionnel d'un individu. La voix humaine varie avec le temps ou les conditions physiologiques et psychologiques du locuteur. Cependant, ces variations intra-locuteur ne sont pas identiques pour tous les humains. En effet, hormis les variations lentes de la voix dues au vieillissement, certains phénomènes extérieurs tels que l'état de santé d'une personne ont une influence variable sur sa voix.

II.2.6 Modélisation des locuteurs

Parmi les approches proposées, on a choisi d'utiliser dans notre travail celle du mélange de Gaussiennes (GMM) pour la voix, Le dernier permet une approximation d'une large gamme de distributions complexes dans l'espace de représentation avec beaucoup de rapidité, simplicité et facilité.

II.2.6.1 Définition :

GMM est l'acronyme de "Gaussian Mixture Model" (modèle de mélange gaussien en français). Il s'agit d'une technique de modélisation statistique utilisée dans le domaine de l'apprentissage automatique (machine learning) et de la reconnaissance des formes. C'est une approche statistique qui consiste à estimer une loi de probabilité inconnue à l'aide d'une combinaison de plusieurs Gaussiennes chacune identifiée par M {1, …, M}

Où M est le nombre de groupes (Cluster) de notre ensemble de données (figure II.5).

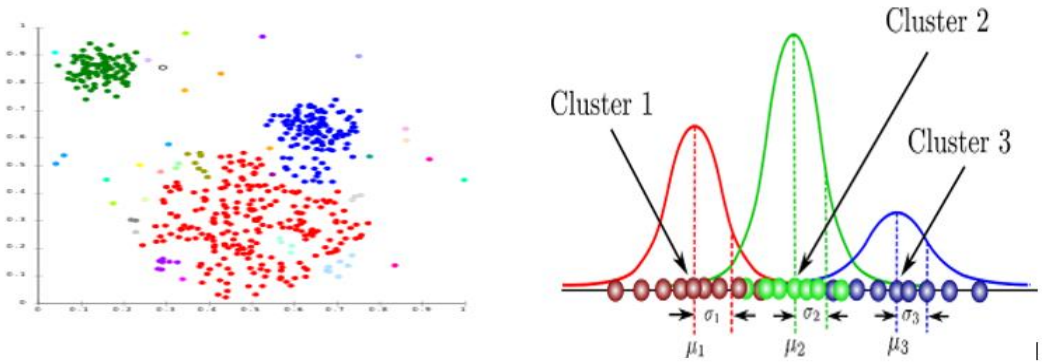


Figure II.5: modélisation gaussienne d'une distribution en 2 D

II.2.6.1 Le modèle de mélange gaussien (GMM) :

Une densité de mélange de gaussiennes est une somme pondérée de M densités gaussiennes. Elle est donnée par l'équation (II.1) :

$$\boldsymbol{P(X|\lambda) = \sum_{i=1}^{M} \pi_i b_i(X)} \ldots\ldots\ldots\ldots\ldots\ldots (\boldsymbol{II.1})$$

Avec :

- x : un vecteur aléatoire de dimension D.
- $b_i(X)$: densité de probabilité de la i-ème gaussienne.
- $\mu_m \; et \; \sum m$: vecteur moyen et matrice de covariance respectivement.
- π_m: poids des gaussiennes constituant le mélange.

La densité de chaque gaussienne est donnée par l'équation (II.2) :

$$\boldsymbol{b_i(x) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} exp\left[-\frac{1}{2}(x-\mu_i)'(\sum_i)^{-1}(x-\mu_i)\right]} \quad \ldots\ldots\ldots (\boldsymbol{II.2})$$

Un modèle GMM est donc représenté par les vecteurs moyens, les matrices de covariance et les poids des gaussiennes, ceci est définit par la notation :

$$\boldsymbol{\lambda = \{\pi_i, \mu_i, \sum_i\} \; avec \; i = 1 \ldots M} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (\boldsymbol{II.3})$$

L'algorithme EM (Expectation- Maximisation) est utilisé pour estimer les paramètres GMM. Chaque intervenant est représenté par son modèle. Pour une séquence de vecteurs d'entraînements on a $X = \{x_1, x_2, \dots, x_T\}$, la vraisemblance GMM peut être écrite comme :

$$\boldsymbol{P(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda)} \dots\dots\dots\dots\dots (\mathbf{II.4})$$

L'objectif de l'estimation du maximum de vraisemblance est de trouver les paramètres du modèle qui maximisent la probabilité de GMM compte tenu des données d'apprentissage. Le modèle d'arrière-plan universel (UBM) est formé à l'aide de l'algorithme EM qui utilise tous les locuteurs formant des énoncés vocaux. Le GMM spécifique au locuteur est créé à partir de l'UBM à l'aide de l'estimation du paramètre Maximum A Posteriori (MAP). Lors de l'identification du locuteur, la notation GMM est effectuée lorsque nous calculons le rapport de vraisemblance log du signal de test étant donné le modèle de locuteur au modèle de fond universel et celui avec le score maximum est identifié comme le locuteur.[22]

II.2.7 Mesures de performance :

Il est nécessaire de définir des mesures permettant d'évaluer les performances de reconnaissance pour une méthode donnée afin de pouvoir la comparer avec d'autres méthodes

II.2.7.1 Le taux de faux rejets (FRR) :

Le taux de faux rejet (FRR : False Rejection Rate), également connu sous le nom de taux de faux négatifs, représente la probabilité qu'un système biométrique rejette une personne autorisée. Le FRR est généralement exprimé en pourcentage et est calculé en divisant le nombre de fausses rejections par le nombre total d'essais :

$$\boldsymbol{FRR = \frac{le\ nombre\ des\ clients\ rejetés(FR)}{le\ nombre\ total\ d'accès\ clients}} \dots\dots\dots (\mathbf{II.5})$$

Le FRR est un paramètre important pour évaluer la convivialité et la praticité d'un système biométrique. Un taux de rejet trop élevé peut entraîner des inconvénients pour les utilisateurs, qui devront effectuer plusieurs tentatives pour être authentifiés.

Toutefois, une FRR trop faible peut entraîner une augmentation du FAR, ce qui compromet la sécurité du système.[23]

II.2.7.1 Le taux de fausses acceptations (FAR) :

Le taux de fausses acceptations (FAR), également connu sous le nom de taux de faux positifs, représente la probabilité qu'un système biométrique accepte une personne non autorisée. Le FAR est généralement exprimé en pourcentage et est calculé en divisant le nombre de fausses acceptations par le nombre total d'essais.

$$\boldsymbol{FAR = \frac{le\ nombre\ des\ imposteurs\ acceptés(FA)}{le\ nombre\ total\ d'accès\ imposteurs}} \quad \ldots\ldots\ldots (\mathbf{II.6})$$

Le FAR est un paramètre important pour évaluer la fiabilité et la sécurité d'un système biométrique. Plus le FAR est faible, plus le système est sûr et précis. Le FAR doit être maintenu à un niveau suffisamment bas pour éviter toute intrusion non autorisée, tout en permettant un accès facile et rapide aux personnes autorisées.[23]

II.2.7.2 Taux d'erreur égal (EER) :

Le taux d'erreur égal (EER), également connu sous le nom de point d'opération de seuil égal, est un indicateur des performances globales d'un système biométrique.

Le taux d'erreur égal est le point où le taux de fausses acceptations (FAR) est égal au taux de fausses rejections (FRR), exprimé en pourcentage. En d'autres termes, le taux d'erreur égal est le point où les taux de fausses acceptations et de fausses rejections sont égaux.

$$\boldsymbol{ERR = \frac{Nombredefaussesacceptations(FA) + Nombredefauxrejets(FR)}{Lenombretotald'accès}} \ldots (\mathbf{II.7})$$

L'EER est un indicateur important de la performance d'un système biométrique, car il permet de mesurer la capacité du système à maintenir un équilibre entre la sécurité et la convivialité. Un EER faible indique que le système a un niveau de sécurité élevé tout en permettant un accès facile et rapide pour les utilisateurs autorisés.[24]

La figure (II.6) illustre les mesures de performance expliquées précédemment

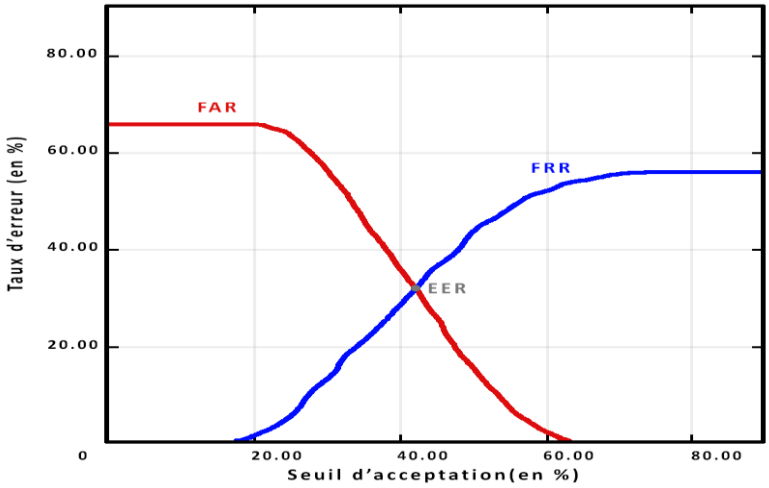


Figure II.6: Illustration du point d'intersection des courbes FRR et FAR

II.2.8 Les Courbes de performance :

Les courbes de performance sont des graphiques qui permettent de visualiser les performances d'un système biométrique en fonction de différents seuils de décision. Les deux courbes les plus couramment utilisées sont la courbe ROC (Receiver Operating Characteristic) et la courbe DET (DetectionErrorTradeoff).

II.2.8.1 La Courbe ROC (Receiver Operating Characteristic):

La courbe ROC représente la relation entre le taux de faux rejets (FRR) et le taux de fausses acceptations (FAR) pour différents seuils de décision. Elle permet d'évaluer la capacité d'un système biométrique à différencier les utilisateurs autorisés des imposteurs.[25]

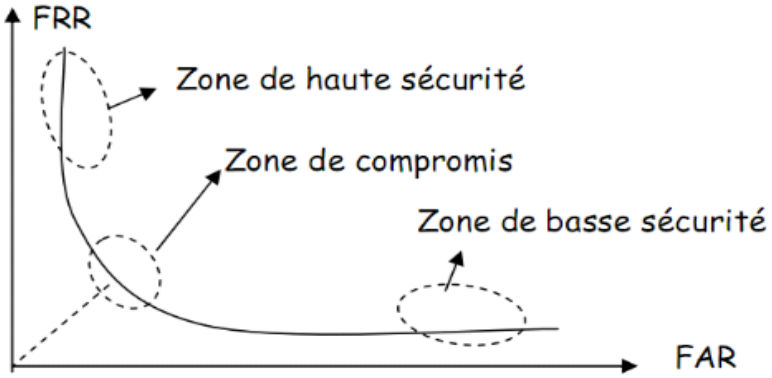


Figure II.7: La Courbe ROC

II.2.8.2 La courbe DET (Detection Error Tradeoff):

La courbe DET représente la relation entre le taux de fausses rejections (FRR) et le taux de fausses acceptations (FAR) pour différents seuils de décision. Elle permet d'évaluer la capacité d'un système biométrique à détecter les utilisateurs autorisés tout en minimisant les fausses alarmes.[23]

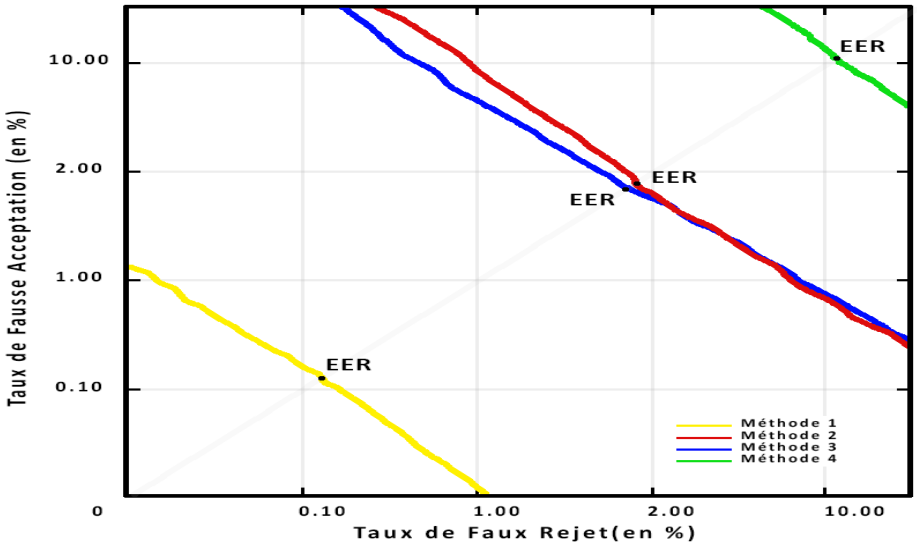


Figure II.8:DET représentant la performance de quatre techniques de vérification automatique du locuteur.

II.3 La Détection d'usurpation d'identité :

La voix est une biométrie populaire pour les approches d'identification. Néanmoins, la biométrie vocale n'est pas exempte de menaces et des stratégies d'attaque ciblant les systèmes de reconnaissance du locuteur ont été décrites. En tant que tel, la conception de contre-mesures pour produire des applications plus robustes est devenue une direction de recherche populaire. Comme illustré dans la figureII.9, les attaques peuvent être conçues à différentes étapes d'un pipeline d'authentification basé sur la biométrie. Notamment, les attaques contradictoires récemment introduites ciblant les réseaux de neurones artificiels peuvent agir au niveau du modèle. Dans le cas particulier de la vérification du locuteur et des systèmes de biométrie vocale, d'autres stratégies d'attaque existent également. Celles-ci sont appelées attaques d'usurpation d'identité et représentent une personne ou un programme informatique qui tente de contourner un système d'authentification en falsifiant les données d'un utilisateur légitime.[26].

Dans ce travail, nous nous concentrons sur les attaques au niveau du microphone. Les attaques au niveau du microphone peuvent être réalisées avec différentes méthodes ; (i)l'usurpation d'identité d'un autre utilisateur, (ii) la parole synthétique et (iii) l'audio préenregistré(rejoué) d'un utilisateur donné.

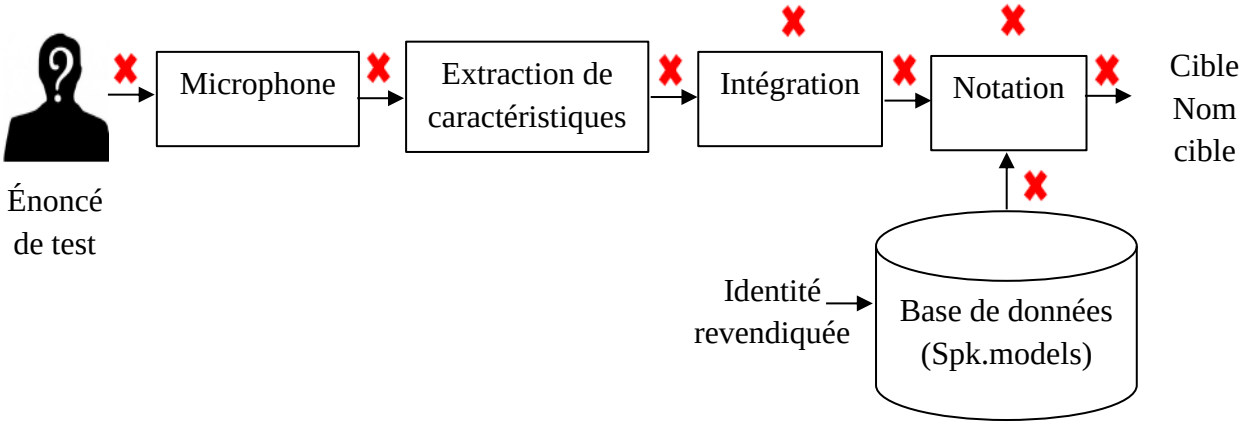


Figure II.9 : points d'attaque (croix rouges) d'un système ASV

II.4 Vulnérabilité de la vérification des conférenciers à la mystification :

II.4.1 Points d'attaque possibles :

Un système ASV typique implique deux processus : l'inscription hors ligne et la vérification du temps d'exécution. Lors de l'inscription hors ligne, un modèle de locuteur cible est formé en utilisant des caractéristiques extraites d'un exemple de discours. Le processus de vérification du temps d'exécution est illustré à la figureII.10, lorsqu'un orateur affirme d'abord une demande d'identité et fournit ensuite un échantillon de son discours. Les caractéristiques extraites de cet échantillon sont comparées au modèle afin de déterminer si le locuteur correspond ou non à l'identité revendiquée modèles, un correspondant à locuteur hypothétique et un second représentant l'hypothèse alternative. Le classificateur détermine une note de correspondance qui représente la similarité relative de l'échantillon avec chacun des deux modèles. Enfin, le module de logique de décision utilise le score relatif (habituellement un rapport log-vraisemblance) pour accepter ou rejeter la demande d'identité. Ces éléments et les liens entre eux représentent tous des points d'attaque possibles. Huit de ces points de

vulnérabilité pour un système générique de ASV sont également illustrés à la figure II.9. Ils peuvent être classés comme suit :

- Attaques directes :

Également appelés attaques d'usurpation, peuvent être appliquées au niveau du microphone ainsi qu'au niveau de la transmission – étiquetées comme points d'attaque 1 et 2 dans la figureII.10. Par exemple, l'usurpation d'identité d'une autre personne ou la présentation d'un signal vocal enregistré ou synthétisé au microphone.

- Attaques indirectes :

Les attaques indirectes nécessitent généralement un accès au niveau du système, par exemple les attaques qui interfèrent avec l'extraction des fonctionnalités (points 3 et 4), les modèles (points 5 et 6) ou calcul de score et logique de décision (points 7 et 8) [26].

Les attaques aux points 1-2 sont considérées comme des attaques directes alors que celles aux points 3-8 sont des attaques indirectes [26].

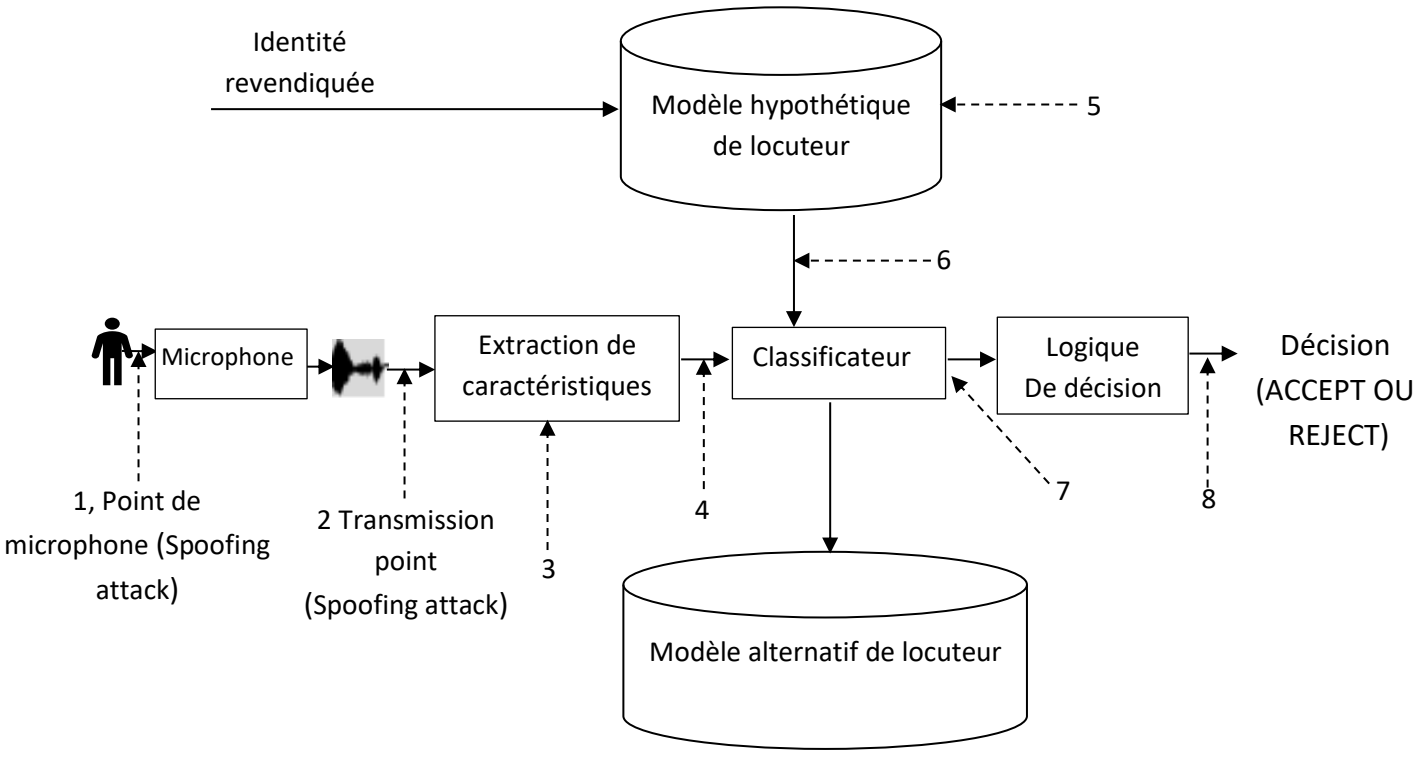


Figure II.10 : Illustration d'un système typique de vérification automatique des locuteurs (ASV)

II.4.1 Les types d'attaques :

Il en existe deux types :

- *Accès physique (PA) :*

Une méthode d'attaque directe, au cours de laquelle un attaquant interagit directement avec le capteur d'un système biométrique. (Microphone).

- *Accès logique (LA) :*

Méthode d'attaque indirecte. Elle vise les vulnérabilités dans le code ou le matériel d'un système biométrique.

II.4.2 Mécanismes anti-usurpation d'identité :

Les systèmes ASV de pointe ne disposent d'aucun moyen particulier de se protéger contre les attaques par usurpation d'identité. Ainsi, lorsqu'un système ASV est systématiquement confronté à la parole convertie, il a tendance à augmenter son FAR, dégradant ainsi ses performances [27]. L'augmentation du FAR est très gênante lorsque l'application doit être robuste. Afin de surmonter ce problème et diminuer le FAR, il est utile d'introduire de nouveaux modules dans les systèmes ASV qui détectent la parole [28]. Les techniques de pointe utilisées dans ces modules s'inspirent des limites des systèmes SV et VC de telle manière qu'ils exploitent les imperfections de discours converti. Ces modules de détecteur de parole convertis peuvent être inclus en tant que post-module de traitement, après le système ASV lui-même. Leur entrée ne devrait être que les énoncés de test qui ont été acceptées par le système. La sortie doit être une décision binaire entre parole convertie ou parole naturelle. C'est exactement dans ce créneau que s'inscrit notre travail.

II.5 Les attaques par usurpation d'identité (Spoofing attacks) :

Les attaques par usurpation d'identité sont largement classées en quatre types usurpation d'identité, SS, VC et rejouer.

II.5.1 Discours synthétique (Synthetic speech) :

SS est aussi connu sous le nom de Text-To-Speech (TTS), qui prend le texte comme entrée et génère la parole comme sortie. Il émule un système humain de production vocale et représente une véritable menace. L'attaque par SS est maintenant en mesure de générer une

voix de haute qualité en raison du progrès dans la sélection des unités paramétriques, statistiques, hybrides, et les méthodes TTS basées sur DNN. Récemment, des techniques fondées sur l'apprentissage profond, sont capables de produire un discours sonore très naturel à la fois dans le timbre et la prosodie [29]

II.5.2 Conversion vocale (Voice conversion) :

VC est le processus de conversion de la voix du haut-parleur source en un son similaire à la voix du haut-parleur cible. VC traite l'information qui se rapporte aux caractéristiques segmentaires et supra segmentaires et garde le contenu du langage similaire [29]

II.5.3 Les attaques rejeu (replay) :

Une attaque par rejeu consiste à rejouer un discours préenregistré à un système [30]

II.5.4 L'usurpation d'identité (Impersonation) :

L'usurpation d'identité est définie comme le processus de production de modèle de voix et comportement de parole similaires de la voix du locuteur cible.

II.6 Conclusion :

Dans ce chapitre nous avons exposé le concept de reconnaissance et de vérification du locuteur. Le diagramme fonctionnel de tels systèmes a été détaillé, de l'étage de paramétrisation jusqu'à l'étage de décision.

La protection de ces systèmes contre les attaques d'utilisateurs mal intentionnés a été aussi abordé avec détail, ainsi les différents types d'attaques et les étages concernés du système ont été expliqués. Un point important soulevé dans ce chapitre est que les systèmes étudiés possèdent deus sous-systèmes : un système de vérification du locuteur et un système de classification binaire de la parole (naturelle - non naturelle) qui protège le système global des attaques.

C'est la réalisation du système binaire de classification qui nous intéresse dans ce travail et fera l'objet des expérimentations exposées dans le chapitre 3.

CHAPITRE III

Résultats et interprétations

III.1 Introduction

Ce chapitre sera consacré aux résultats obtenus en utilisant la méthode adoptée à savoir la technique GMM pour la détection robuste de parole truquée. Dans un premier temps nous allons présenter la base de données utilisée dans ce travail à savoir l'ASV anti-spoof 2017 V2. Ensuite nous allons exposer la stratégie adoptée pour définir les étapes de simulation. Nous exposerons ensuite les résultats obtenus qui seront suivis par les interprétations correspondantes. Le chapitre sera terminé bien sûr par une conclusion.

III.2 Présentation de la base de données utilisée :

La base de données ASVspoof 2017 a été collectée afin de favoriser le développement de contre-mesures pour protéger les systèmes ASV contre les attaques. La même base de données 1.0 originale a été utilisée pour le défi de 2017 avec des changements dans la version 2.0, c'est cette dernière qui est utilisée dans ce travail.

Les sections : III.1 et III.2 sont exclusivement basées sur la référence [31] .

ASVspoof 2017 V2 est une collecte de déclarations (prononciations) appartenant à deux catégories : prononciations naturelles authentiques (prononcées de bonne foi par le vrai utilisateur) et prononciations non authentiques (mal intentionnées) obtenues par différents procédés dans le but de tromper le système. Les déclarations malhonnêtes sont le résultat du repositionnement et de l'enregistrement de déclarations bien intentionnées utilisant une variété de dispositifs hétérogènes et des environnements sonores.

III.1.1 Structure de la base de données ASV spoof 2017 :

Le corpus ASVspoof 2017 est divisé en trois sous-ensembles : apprentissage, développement et évaluation. Un résumé de leur composition est présenté au tableau III.1.

Les données correspondent à 177 différentes sessions de replay et 61 configurations de rejeu distinctes. Le nombre de configurations de rejeu qui composent le sous-ensemble d'évaluation est considérablement plus élevé que celui présent dans les sous ensemble d'apprentissage et de développement. Les prononciations sont dotées d'étiquettes authentiques/falsifiées (non rejeu/rejeu), en plus de l'identité du locuteur

A l'origine, deux parties de cette base sont initialement disponibles aux participants au défi à savoir : la partie apprentissage et la partie développement qui sert à mettre au point le système.

La dernière partie est dévoilée ultérieurement aux participants à fin de mettre à l'épreuve les méthodes qu'ils ont développées.

Subset	# Spk	# Replay sessions	# Replay Config	#Utterances	
				Bona fide	Replay
Training	10	6	3	1507	1507
Devel.	8	10	10	760	950
Eval.	24	161	57	1298	12008
Total	42	177	61	3565	14465

Tableau III. 1: Statistiques actualisées du corpus ASVspoof 2017 2.0

III.1.2 Procédés d'obtention des données de la base :

Nous allons présenter dans cette section les procédés et les environnements acoustiques utilisés pour obtenir les données de cette base.

III.1.2.1 Simulation du rejeu (replay) :

La figure III.1 illustre le procédé d'obtention.

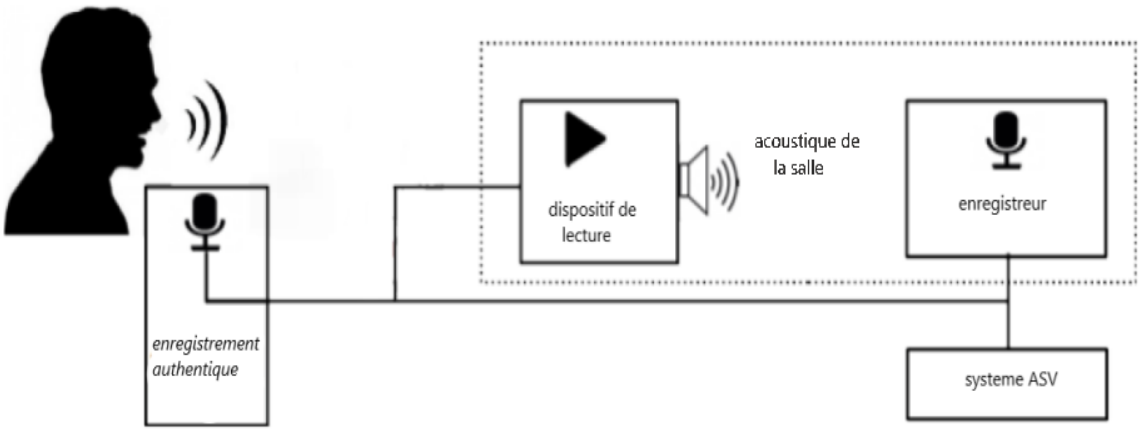


Figure III.1 : Illustration du scénario de rejeu.

III.1.2.2 Environnements acoustiques :

L'environnement acoustique est l'espace physique dans lequel les données de parole originales sont rejouées et réenregistrées. La base de données ASV spoof 2017 a été recueillie dans 26 environnements différents répertoriés dans le tableau III.2. Les environnements sont désignés par les identifiants de E01 à E26. Les variations entre elles comprennent les types et les niveaux de bruit ambiant additif et de réverbération convolutive.

Le niveau du bruit est supposé être inversement corrélé avec la menace d'attaque posé par rejeu enregistrées dans chaque environnement.

Les codes de couleur dans le tableau III.2 servent d'indicateur de menace d'enregistrement. Le degré de menace allant de faible (vert) à élever (rouge) en passant par moyen (jaune)

Id	Environnement	Id	Environnement
E01	Anechoic room	E14	Office 02
E02	Balcon 01	E15	Office 03
E03	Balcon 02	E16	Office 04
E04	Maison07	E17	Office 05
E05	Maison08	E18	Office 06
E06	cantine	E19	Office 07
E07	Maison01	E20	Office 08
E08	Maison02	E21	Office 09
E09	Maison03	E22	Office 10
E10	Maison04	E23	Studio
E11	Maison05	E24	Fil analogique01
E12	Maison06	E25	Fil analogique02
E13	Office 01	E26	Fil analogique03

Tableau III. 2: Environnements acoustiques utilisés dans la collecte de données

III.2.1 Dispositifs de lecture :

Il existe 26 appareils de lecture, notés P01-P26, comme indiqué dans le tableau III.3.

Des échantillons rejoués sont collectés :

- Qualité inférieure et devrait être détecté avec une relative facilitée (vert) comme les smartphones avec des haut-parleurs compacts et petits.
- Qualité moyenne (jaune) comme les ordinateurs de bureau avec haut-parleurs plus grands.
- Haute qualité (rouge) représente la menace la plus grave, comme les dispositifs de surveillance, les casques de studio et aussi les interfaces audios ou les produits analogiques utilisés dans ensemble d'enregistrements dans des environnements de fils analogiques.

Id	Dispositifs de lecture
P01	Haut-parleurs tout-en-un pour PC

P02	Haut-parleurs créatifs A60
P03	Moniteur studio Genelec 8020C
P04	Moniteur de studio Genelec 8020C (2 haut-parleurs)
P05	Casque Beyerdynamic DT 770 PRO
P06	Haut-parleurs internes de l'ordinateur portable Dell
P07	Conférencier Dynaudio BM5A
P08	Haut-parleurs internes de l'ordinateur portable HP
P09	Haut-parleur VIFA M10MD-39-08
P10	Haut-parleurs internes du netbook ACER
P11	Smartphone BQ Aquaris M5
P12	Haut-parleurs de faible qualité Logitech
P13	Sortie ligne PC de bureau
P14	Haut-parleurs Labtec LCS-1050
P15	Moniteur de studio Edirol MA-15D
P16	Tablette Lenovo Ideatab S6000-H
P17	Haut-parleurs multimédia Logitech S120
P18	Haut-parleurs internes MacBook pro
P19	Enceinte portable Altec lansing Orbit USB iML227
P20	Smartphone Samsung GT-I9100
P21	Tablette Samsung GT-P6200
P22	Moniteur de studio Behringer Truth B2030A
P23	Sortie de ligne d'interface audio Focusrite Scarlett 2i2
P24	Sortie de ligne d'interface audio Focusrite Scarlett 2i4
P25	Moniteur de studio Genelec 6010A
P26	Casque d'écoute AKG K242HD

Tableau III. 3 : dispositifs de lecture utilisés dans la collecte des données

III.1.2.3 Dispositifs d'enregistrements :

Il y a 25 appareils d'enregistrement qui sont illustrés au tableau III.4 ,il existe trois catégories d'appareils enregistrement :

- Les appareils de qualité inférieure (verts) sont facilement détectés, comme les téléphones intelligents et les ordinateurs portatifs dont la batterie fonctionne avec des microphones compacts.
- Dispositifs de qualité moyenne (jaune) comme les ordinateurs de bureau équipés d'écouteurs ou de webcams.

- Les appareils de qualité supérieure (rouges) sont difficiles à détecter, comme les microphones intensifs ou les enregistreurs manuels.

Id	Appareils d'enregistrement
R01	Zoom H6 enregistreur pratique
R02	Smartphone BQ Aquaris M5
R03	Casque de basse qualité
R04	Nokia Lumia 635 smartphone
R05	NT2 microphone
R06	Rộde smartLav+ microphone
R07	Samsung Galaxy S7 smartphone
R08	PC de bureau microphone entrée
R09	Enregistreur Zoom H6 avec micro Behringer ECM8000
R10	Enregistreur Zoom H6 avec microphone MSH-6
R11	Enregistreur Zoom H6 avec microphone XY
R12	iPhone 5c smartphone
R13	iPhone 7 plus smartphone
R14	iPhone 4 smartphone
R15	Logitech C920 webcam
R16	microphone minidsp UMIK-1
R17	Samsung Galaxy Trend 2 smartphone
R18	Samsung GT-I9100 smartphone
R19	Tablette Samsung GT-P6200
R20	Samsung Trend 2 smartphone
R21	Microphone AKG C3000
R22	microphone SE electronic 2200a
R23	Entrée ligne d'interface Focusrite Scarlett 2i2
R24	Entrée ligne d'interface Focusrite Scarlett 2i4
R25	Enregistreur pratique Zoom HD1

Tableau III. 4: dispositifs d'enregistrement utilisés dans la collecte des données

III.2 Stratégie de l'étude :

Deux GMMs : un pour la parole authentique et un pour la parole truquée sont entrainés par des échantillons de parole appartenant aux deux catégories en question.

La première étape concerne l'étage d'entrée (Front End) et étant donné que les coefficients utilisés dans cette simulation sont du type CQCC, cette étape aura pour objectif d'étudier l'effet du nombre de coefficients utilisés sur les performances du système.

Dans une deuxième démarche, il sera question d'explorer l'aspect dynamique des coefficients CQCC en ajoutant les paramètres dérivatifs d'ordre 1 et 2 (vitesse et accélération).

La troisième étape concerne l'étage de modélisation. Il sera question d'étudier l'effet du nombre de gaussiennes utilisés dans les GMM sur les performances du système

III.3 Résultats obtenus :

Toutes les simulations ont été réalisées sur la même machine :

Intel(R) Core(TM) i5 CPU M 460 @ 2.53 GHz.

RAM : 8 Go.

Le programme utilisé pour la simulation : MATLAB2021

III.3.1 Utilisation des coefficients CQCC statiques :

Dans cette étape le nombre de gaussiennes est fixe (512) et le nombre de coefficients CQCC est variable. Le tableau III.5 donne le taux d'erreur égal EER obtenu. La figure III.1 donne la courbe ROC obtenue.

Nombre de coefficients	20	25	30	35	40	45	50	55	60
EER (%)	12.04	12.89	13.63	13.52	12.38	12.13	14.43	17.71	16.14

Tableau III. 5 : EER en fonction du nombre de coefficients CQCC statiques

La moyenne du temps d'exécution est d'environ 2900 secondes pour chaque colonne du tableau III.5, soit un peu plus de 48 minutes.

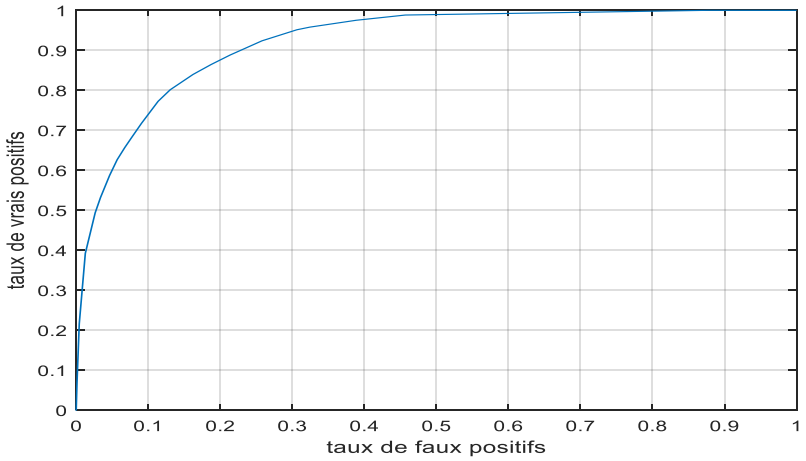


Figure III.2 : Courbe ROC obtenue dans le cas de CQCC statiques

III.3.2 Utilisation des coefficients dynamiques :

Dans cette étape le nombre de gaussiennes est fixe (512) et les coefficients CQCC SDA puis ESDA (Energy + Statiques + Dynamiques + Accélération) sont utilisés.

Le tableau III.6 donne le taux d'erreur égal EER obtenu.

Nombre de coefficients	20	25	30	35	40	45	50	55	60
EER (SDA)	12.64	12.56	12.21	11.26	11.65	11.34	11.39	11.81	12.17
EER (ESDA)	12.12	11.66	9.97	10.54	10.46	11.58	10.14	12.48	14

Tableau III. 6: EER en fonction du nombre de coefficients CQCC dynamiques

La figure III.2 donne la courbe ROC obtenue :

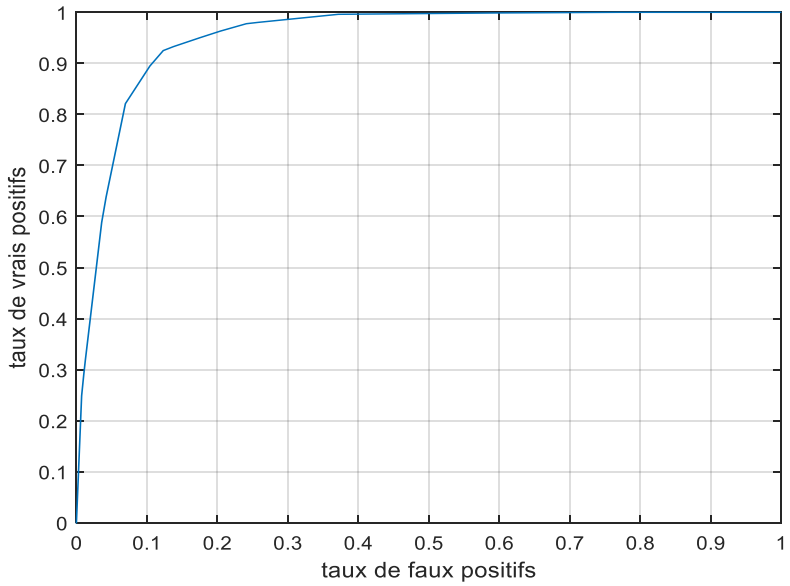


Figure III.3 : Courbe ROC obtenue dans le cas de CQCC dynamiques

III.3.3 Utilisation coefficients MFCC de type SDA (statiques + delta + accélération)

Nombre de coefficients	13	20	25	30
EER (%)	11.89	14.87	13.47	14.53

Tableau III.7: Résultats pour coefficients MFCC de type SDA

III.3.4 Effet du nombre de gaussiennes :

Dans cette étape, le nombre de coefficients CQCC est fixe (30 statiques plus leurs dérivées d'ordre 1 et 2) et le nombre de gaussiennes est variable. Le tableau III.7 donne le EER obtenu pour chaque nombre de gaussiennes.

Nombre de gaussiennes	128	256	512	1024
EER %	13.63	11.10	9.97	11.37

Tableau III.8: EER en fonction du nombre de gaussiennes

La moyenne du temps d'exécution est d'environ 4054 secondes par colonne du tableau III.7.

III.4 Interprétations :

- Le tableau III.5 montre que l'utilisation des seuls coefficients statiques n'est pas bénéfique, en outre on remarque une hausse du EER lorsqu'on augmente le nombre de coefficients.
- Le tableau III.6 montre que l'utilisation de coefficients dynamiques conduit à une baisse du EER ce qui est bénéfique. La valeur la plus basse du EER obtenue étant celle qui correspond à l'utilisation de 30 coefficients ESDA
- La différence de performance est notable lorsqu'on examine les deux courbes ROC obtenues dans les deux cas statique et dynamique. En effet la courbe de la figure III.2 est beaucoup lus étirée vers le haut que celle de la figure III.1. En d'autres termes, la surface sous la courbe de la figure III.2 est plus importante que celle sous la courbe de la figure III.1 ce qui implique de meilleures performances de classification.
- Comparés avec les résultats précédents, les coefficients CQCC donnent de meilleures performances. C'est d'ailleurs la raison pour laquelle les CQCC sont plus utilisés dans ce domaine que les MFCC ou bien ils sont utilisés conjointement selon une stratégie nommée : fusion de paramètres. Cette stratégie trouve son intérêt dans le fait que chaque type de paramètres donne de bons résultats dans certaines conditions et pas dans d'autres. La fusion de paramètres de plusieurs types permet de doter le système des avantages propres à chaque type de paramètre pris séparément.

- Concernant l'effet du nombre de gaussiennes utilisées, il apparait clairement que 512 gaussiennes (généralement adoptés dans la littérature) suffisent à avoir de bons résultats, un nombre plus faible de gaussiennes conduit à une dégradation de résultats. On pourrait être tentés d'augmenter le nombre de gaussiennes afin d'améliorer les performances de classification mais il ne faut pas oublier que cela conduit à augmenter considérablement le nombre de paramètres des GMMs conduisant ainsi au problème de sous-apprentissage (under training), ce qui se traduit à son tour par une dégradation de performance de classification.
- Les enregistrements sonores ayant été collectés par différentes sources, il aurait certainement été bénéfique d'effectuer une normalisation de la moyenne et de la variance des coefficients (CMVN) afin de palier aux différences entre canaux. Etant rattrapés par le temps, cette étape pourrai servir à améliorer le présent travail.

III.5 Conclusions :

Nous avons exposé dans ce chapitre les résultats de simulation de la technique GMM dans le contexte de la classification de parole naturelle et truquée. L'étage d'entrée utilise des coefficients de type CQCC. Il a été question d'étudier l'effet d'un certain nombre de paramètres sur les performances de classification à savoir la nature et le nombre de coefficients utilisés, ainsi que le nombre de gaussiennes des GMMs.

Conclusion générale

Arrivés à terme de ce travail, un certain nombre de constatations et de conclusions sont à soulever.

La plus grande source de difficulté à laquelle nous avons fait face dans ce travail était le caractère vaste et pluridisciplinaire de ce domaine, conjugué à un intervalle de temps assez limité pour appréhender la base de connaissances nécessaire.

Ce domaine de recherche suscite un intérêt de plus en plus croissant de la part de la communauté de recherche dans le domaine du traitement de la parole, étant donné qu'il concerne l'aspect sécurité des systèmes offrant un certain nombre de services à un certain nombre d'utilisateurs qui proclament leur identité chaque fois qu'ils sollicitent les services du système en question. L'usurpation d'identité est une vraie menace pour de tels systèmes, et cette menace s'intensifie avec l'avancée de la technologie. II est donc capital de doter les systèmes en question de mécanismes de protection contre les attaques.

Nous avons fait dans ce travail une étude de l'une des techniques les plus utilisées dans ce domaine à savoir la technique GMM. D'autres méthodes sont certainement en cours d'utilisation et de développement, à leur tête on trouve l'apprentissage profond (deep learning). Ils peuvent être proposés comme perspective d'amélioration pour le présent travail.

Aussi, si le temps l'aurai permis, il aurait été intéressant d'essayer un autre type de paramètres en entrée, à savoir les fameux coefficients MFCC. Pour aller plus loin, l'idée d'une fusion des décisions provenant des deux types de paramétrisation serait intéressante à explorer.

Enfin, il est utile de noter que ce travail concerne la seule classification de parole en deux classes : naturelle et truquée et que cette classification est utilisée en conjonction avec un étage de reconnaissance du locuteur.

1. Nefti, S., *Segmentation automatique de parole en phones. Correction d'étiquetage par l'introduction de mesures de confiance*. 2004, Université Rennes 1.

2. Sturmel, N., *Analyse de la qualité vocale appliquée à la parole expressive*. 2011, Université Paris Sud-Paris XI.

3. Beigi, H. and H. Beigi, *Speaker recognition*. 2011: Springer.

4. Buniet, L., *Traitement automatique de la parole en milieu bruité: étude de modèles connexionnistes statiques et dynamiques*. 1997, Université Henri Poincaré-Nancy 1.

5. Zitouni, I., *Modélisation du langage p les systèmes de reconnaissance de la parole destinés aux grands vocabulaires: application à MAUD.* 2000.

6. AJGOU, R., *Reconnaissance Automatique du Locuteur à Travers les Canaux Digitaux*. 2016, Université Mohamed Khider–Biskra.

7. Didiche, M.K.A., *Modélisation neuro-prédictive pour la classification phonétique*. 2014, Université Mohamed Khider-Biskra.

8. Alouache Lyes, L.Y., *Compensation de la variabilité du canal en reconnaissance du locuteur*. 2019, université d'Akli Mohand oulhadj-bouira.

9. KEBE, A.B., *Linguistique générale*

10. Drygajlo, D.A., *TRAITEMENT DE LA PAROLE* p. 233.

11. LARAK, I. and S. AFFOUN, *Systèmes de Reconnaissance Automatique de la Parole*. 2017, Université Akli Mouhand Oulhadj-Bouira.

12. Ranjan, R. and A. Thakur, *Analysis of feature extraction techniques for speech recognition system.* International Journal of Innovative Technology and Exploring Engineering, 2019. **8**(7C2): p. 197-200.

13. Yang, J., R.K. Das, and H. Li. *Extended constant-Q cepstral coefficients for detection of spoofing attacks*. in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2018. IEEE.

14. Furui, S., *Digital speech processing.* Synthesis, and Recognition, 1989.

15. Hollien, H., *The expert witness: Ethics and responsibilities.* Journal of Forensic Sciences, 1990.

16. Rabiner, L. and B.-H. Juang, *Fundamentals of speech recognition*. 1993: Prentice-Hall, Inc.

17. Rabiner, L.R., *Digital processing of speech signals*. 1978: Pearson Education India.

18. Preti, A. *Surveillance de réseaux professionnels de communication par reconnaissance du locuteur*. 2008. Avignon.

19. Sayoud, H., *Reconnaissance automatique de locuteur approche connexioniste*. 2003, Université des Sciences et de la Technologie d'Alger, Houari Boumediène.

20. Boite, R., *Traitement de la parole*. 2000: PPUR presses polytechniques.

21. Jousse, V., *Identification nommée du locuteur: exploitation conjointe du signal sonore et de sa transcription*. 2011, Université du Maine.

22. Reynolds, D.A. and R.C. Rose, *Robust text-independent speaker identification using Gaussian mixture speaker models.* IEEE transactions on speech and audio processing, 1995. **3**(1): p. 72-83.

23. Wayman, J.L., et al., *Biometric systems: Technology, design and performance evaluation*. 2005: Springer Science & Business Media.

24. Available from: https://www.researchgate.net/publication/43057044_Reconnaissance_Biometrique_par_Fusion_Multimodale_du_Visage_et_de_l'Iris.

25. Available from: https://www.researchgate.net/figure/Courbe-ROC-pour-un-systeme-de-verification-de-scripteurs_fig6_280979425.

26. Wu, Z., et al., *Spoofing and countermeasures for speaker verification: A survey.* speech communication, 2015. **66**: p. 130-153.

27. Wu, Z., et al. *A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case*. in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. 2012. IEEE.

28. Wu, Z., E.S. Chng, and H. Li. *Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition*. in *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

29. Kamble, M.R., et al., *Advances in anti-spoofing: from the perspective of ASVspoof challenges.* APSIPA Transactions on Signal and Information Processing, 2020. **9**: p. e2.

30. Ergünay, S.K., et al. *On the vulnerability of speaker verification to realistic voice spoofing*. in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. 2015. IEEE.

31. Delgado, H., Todisco, M., Sahidullah, M., Evans, N., Kinnunen, T., Lee, K. A., & Yamagishi, J. (2018, June). ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements. In *Odyssey 2018-The Speaker and Language Recognition Workshop*