REPUBLIQUE ALGÉRIENNE DEMOCRATIQUEET POPULAIRE

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Université YAHIA FARES DE MÉDÉA



Faculté de Technologie

Département : Génie Électrique

Mémoire de fin d'études de Master

Filière : télécommunications

Spécialité : Système de Télécommunications

ÉTUDE ET IMPLÉMENTATION D'UN MODÈLE D'ATTAQUE CONTRADICTOIRES DANS UN SYSTÈME DE RECONNAISSANCE DU LOCUTEUR DANS UN ENVIRONNEMENT MOBILE

Encadré par :

- Dr. KROBBA Ahmed

Présenté par :

- Mansouri Zohra Abir
- Benkortbi Manel

Année universitaire 2022/2023

Dédicaces

La première des pages qui reste la dernière à écrire ...Et voilà le moment d'achever des merveilleuses années en université, des années pleines avec des magnifiques moments et des bons amis.je dédie ce modeste travail a :

Tout d'abord, je tiens à remercier DIEU de m'avoir donné la force et le courage de mener à bien ce modeste travail.

Je tiens à dédier cet humble travail à : Ma mère qui m'a entouré d'amour, d'affection et qui fait tout pour ma réussite, À mon père qui nous a quitté trop tôt.

À mes très chers frères : ***Amine, AbdelFattah*** *Mes chères sœurs :* ***Nassima et Naila,*** *merci d'être toujours à mes côtés, par votre présence, par votre amour dévoué et votre tendresse, pour donner du goût et du sens à ma vie. En témoignage de mon amour et de ma grande affection, je vous prie de trouver dans ce travail l'expression de mon estime et mon sincère attachement. Je prie Allah, le tout puissant, pour qu'il vous donne bonheur et prospérité.*

A mes amis qui m'ont toujours aidé et encouragé, qui étaient Toujours à mes côtés, et qui m'ont accompagnaient durant mon chemin d'études supérieures, mes aimables amis « Nadjet, Chamia, Asma et en particulier A mon cher binôme Manel.

A mes nièces et mes neveux : Malak, Walid, Maram, Ranim, Céline, Amira et Youcef.

A toute ma famille

Abir

Dédicaces

Tout d'abord je tiens à remercier Dieu le tout puissant de m'avoir donné la santé, la volonté et le courage.

Ensuite, à mon père et ma mère pour leur soutien et leurs encouragements et qu'aucune dédicace ne pourra compenser leur sacrifice pour mon bien être et mon bonheur.

A mon petit cher frère MOHAMED NASSIM, pour lequel j'ai beaucoup d'estimes.

A mes sœurs ASSIA REMAISSA ET KARIMA que je les souhaite une longue vie pleine de joie et de réussite.

A mes oncles, mes tantes et toute ma famille.

A mes amis qui m'ont toujours aidé et encouragé, qui étaient Toujours à mes côtés, et qui m'ont accompagnaient durant mon chemin d'études supérieures, mes aimables amis « AMIRA, INES, NADA, IMEN, AYMEN et en particulier A mon cher binôme ABIR.

A notre encadreur Dr.KROBBA AHMED Et à tous ceux qui nous ont soutenus pour l'achèvement de ce projet avec réussite.

MANEL

Remerciement

Avant tout, nous remercions ALLAH tout puissant de nous avoir donné la Volonté, le courage et la patience pour terminer ce travail.

Notre profond respect et nos sincères remerciements à Monsieur ***A.Krobba****, pour nous avoir fait confiance pour avoir accepté de diriger la réalisation de ce mémoire et pour son soutien inconditionnel.*

Nous adressons nos respectueux remerciements à tous les membres du jury qui nous ont fait l'honneur de participer à l'évaluation de notre travail.

Nos vifs remerciements s'adressent également à tous les enseignants qui ont contribué à notre formation tout au long de nos études.

Nos sentiments de reconnaissance et nos remerciements vont également à l'encontre de toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce travail.

Abir, Manel

Abbreviations

ADN Acide Désoxyribo nucléique

AMR-WB Adaptative Multi-Rate – Wide Band

AMR-NB Adaptive Multirate narrow-band

AWGN Additive White Gaussian Noise

BIM Basic itterative method

DTW Dynamic Time Warping

DNN Deep Neural Networks

EER Equal Error Rate

FA Fausses Acceptations

FAR False Acceptation Rate

FR Faux Rejets

FRR False Rejection Rate

FGSM Fast Gradient Sign Method

IA intelligence artificiel

IAL Identification automatique du Locuteur

IBM International Business Machines

IMT International Mobile Telecommunications

MFCC Mel Frequency Cepstral Coefficients

ML machine Learning

NIST National Institute of Standards and Technology

PGD Projected Gradient Descent

PLDA Probabilistic Linear Discriminant Analysis

RAL Reconnaissance automatique du Locuteur

VAL Vérification Automatique du Locuteur

RAP Reconnaissance automatique de la Parole

SNR Rapport Signal sur Bruit

SRE Speaker Recognition Evaluation

VAL Vérification Automatique du Locuteur

Listes des figures

Figure I.1 : Modèle simplifié de l'appareil phonatoire.....4
Figure I.2 : Le larynx.....5
Figure I.3 : Le signal parole.....5
Figure I.4 : Acte de communication parlée.....6
Figure I.5 : Traitement automatique de la parole.....7
Figure. I.6 : Vue schématique de l'appareil vocal, dans le plan sagittal médian.....8
Figure. I.7 : synthèse et codage de parole.....9
Figure II.1 : Les principales modalités biométriques.....12
Figure II.2 : Différentes tâches du traitement de la parole.....13
Figure II.3 : Schéma de principe de base de l'identification du locuteur.....15
Figure II.4 : Vérification automatique du locuteur.....15
Figure II.5 : Structure générale d'un système RAL.....17
Figure II.6 : Schémas synoptiques de l'extraction de paramètres acoustiques.....19
Figure II.7 : Extraction des paramètres MFCCs.....20
Figure II.8 : Exemple d'attaque contradictoire ciblée contre les « gros appareils de cuisine »23
Figure II.9 : Différents types d'attaques adversaires.....24
Figure II.10 : Exemples d'attaques d'IA accusatoires.....26
Figure II.11 : Illustration de l'attaque ciblée sur la reconnaissance automatique de la parole..27
Figure II.12 : Les Impacts des capacités adverses.....30
Figure III.1 : Schéma synoptique du système de vérification du locuteur au canal de transmission (AWGN) et environnement réel mobile.....35
Figure III. 2 : le signal de la parole original et son spectrogramme.....44
Figure III. 3 : Le signal de perturbation et son spectrogramme.....44
Figure III.4 : Le signal de parole contradictoire et son spectrogramme.....45
Figure III.5 : Les performances du système de vérification en utilisant les paramètres MFCC-FGSM et MFCC-PGD au canal de transmission mobile AWGN.....46
Figure III.6 : Les performances du système de vérification en utilisant les paramètres MFCC-FGSM et MFCC-PGD en environnement mobile bruité (bruit de foule) 48

Liste des tableaux

Tableau III 1 : Les performances du système de vérification en utilisant les paramètres MFCC-FGSM et MFCC-PGD au canal de transmission mobile AWGN.....46
Tableau III 2 : Les performances du système de vérification en utilisant les paramètres MFCC-FGSM et MFCC-PGD en environnement mobile bruité (bruit de foule).....47

Tables des matières

Dédicaces i
Dédicaces ii
Remerciement iii
Abbreviations iv
Listes des figures vi
Liste des tableaux vii
Introduction générale 1
Chapitre I : Généralités sur la parole 3
I.1 Introduction 3
I.2 Définition de la parole 3
I.3 La production de la parole 3
I.4 Caractéristiques du signal de parole 5
I.5 La voix et le langage 5
I.6 Le cerveau et le langage 5
I.7 Traitement automatique de la parole 6
I.8 Mécanisme de production du signal de parole 7
I.9 Variabilité et catégorisation en parole 8
I.10 Synthèse de parole 9
I.10.1 Technique de synthèse de parole 9
I.11 Les sources de variation de la parole 9
I.12 Conclusion 10
Chapitre II : système de reconnaissance du locuteur et les algorithmes d'attaques FGSM, PGD 11
II.1 Introduction 11
II.2 L'authentification biométrique 11
II.3 La reconnaissance du locuteur 12
II.3.1 Identification automatique du locuteur 13
II.3.2 Application IAL 14
II.3.3 Vérification automatique du locuteur 15
II.3.4 Application VAL 16
II.4 La variabilité du signal de parole 16
II.4.1 La variabilité intra-session 16

II.4.2 La variabilité inter-session 16
II.4.3 La variabilité inter-locuteurs 16
II.5 L'architecture d'un système RAL 17
II.5.1 La para-métrisation : 17
II.5.2 Modélisation du locuteur 19
II.5.3 Quantification vectorielle 21
II.6 Décision et mesure des performances 21
II.6.1 L'identification automatique du locuteur 22
II.6.2 Vérification automatique du locuteur 22
II.7 Objectifs et Connaissance contradictoires 23
II.8 Portée de perturbation 24
II.9 Adversaire Attaques on IA systèmes 24
II.9.1 Exemples d'attaques d'IA accusatoires 25
II.10 L'algorithme Fast Gradient Sign Method (FGSM) 26
II.10.1 Fonctionnement d'algorithme FGSM 27
II.11 L'algorithme Projected Gradient Descent (PGD) 28
II.11.1 Fonctionnement d'algorithme PGD 28
II.12 Attaques accusatoires et leurs répercussions 28
II.13 Les méthodes efficaces de défense contre les attaques accusatoires 30
II.13.1 Augmentation des données de formation 30
II.13.2 Régularisation 31
II.13.3 Optimisation 31
II.13.4 Distillation 31
II.14 Conclusion 31
Chapitre III : Simulation et Résultats 33
III.1 Introduction 33
III.2 logiciels et langages utilisés dans cette application 33
III.2.1 Logiciels Matlab 33
III.3 Description de bases de données utilisées 33
III.3.1 La description de la base de données NIST-2008 33
III.4 Étapes de la réalisation des systèmes de vérification du locuteur 34
III.5 La structure générale de Système de vérification du locuteur 34
Ⅲ.5.1 L'approche de modélisation GMM-UBM 36
Ⅲ.5.2 Calcul des statistiques suffisantes (Baum-Welch statistics) 38
Ⅲ.5.3 Variabilité totale (i-vecteur) 38

Ⅲ.5.4 Analyse Discriminante Linéaire (LDA) 39
III.6 Analyse discriminante linéaire probabiliste gaussienne (G-PLDA) 40
Ⅲ.6.1 Apprentissage du modèle GPLDA 40
Ⅲ.6.2 Calcul de score 40
III.7 Extraction des paramètres 41
III.7.1 Para-métrisation des coefficients MFCC 41
III.7.2 Protocole d'évaluation des résultats 41
III.8 Simulations et résultats 42
Ⅲ.8.1 Les différentes normes de codeurs de la parole 42
III.8.2 Attaques contradictoires 42
III.9 Performance du système de vérification du locuteur en bruit AWGN pour les algorithmes d'attaques adversaires FGSM et PGD 45
III.10 Performance du système de vérification du locuteur en bruit de la foule pour les algorithmes d'attaques adversaires FGSM et PGD 47
III.11 Conclusion 48
Conclusion générale 49
Bibliographie 50
Résumé 54

Introduction générale

L'information est un élément constructif et décisif dans tous les domaines. Tous tout au long de l'histoire, les humains ont tenté de transmettre des informations dans une seule direction sûre et confidentielle. Le masquage d'informations a été utilisé comme un outil de sécurité très important pour la stratégie et les échanges militaires données sensibles entre utilisateurs. Besoin de communication sécurisée éprouvé et toujours largement utilisé dans le monde numérique Parce que la plupart de ces informations sont stockées sur votre ordinateur et envoyées à d'autres ordinateurs sur Internet. Pour éviter que les données confidentielles ne tombent dans les mauvaises mains, les chercheurs ont étudié des différentes méthodes et techniques mathématiques reliées aux aspects de sécurité de l'information.

Le RAL (Reconnaissance automatique du Locuteur) consiste à reconnaître l'identité d'un individu en analysant sa voix. Cependant, vous pouvez utiliser des facteurs autres que la langue pour vous authentifier. Les gens aiment les empreintes digitales et l'ADN (Acide Désoxyribose nucléique). Contrairement à la voix, ce dernier élément fait partie du corps humain, Ils ne changent pas avec le temps, ou changent très peu, et vous ne pouvez pas les changer exprès Individuel. En raison de ces caractéristiques, le terme biométrie est souvent utilisé. Définit et met l'accent sur une fiabilité très élevée. Le terme biométrie ou empreinte vocale est ne convient pas à la caractérisation vocale. Cependant, certaines voix demeurent. Seules les applications (services accessibles depuis le réseau téléphonique) sont disponibles authentifiez l'utilisateur.

Les systèmes biométriques sont utilisés pour authentifier une personne basée sur certaines caractéristiques physiques ou comportementales. Vérifier le processus Le haut-parleur automatique (VAL) est une méthode biométrique Il s'agit de reconnaître l'identité d'une personne à travers sa voix. Application de Les systèmes VAL se différencient par le contexte d'application et le niveau de sécurité. Application de téléphonie pour le service interurbain (téléphone mobile). Valider les transactions bancaires (en plus des codes d'accès) confidentiel, améliorant ainsi les services bancaires). Ces systèmes comprennent Architecture client-serveur du réseau mobile contenue dans le serveur Les informations et la langue choisie par le client sont envoyées au serveur sous la langue. Forme appropriée. L'objectif de ce travail consiste en l'étude de solutions assurant la sécurité des systèmes de reconnaissance du locuteur dans les communications mobiles en proposant des algorithmes de protection contre les cyber-attaques, la méthode du signe du gradient rapide (FGSM : Fast Gradient Sign Method) et la descente du gradient projetée (PGD : Projected Gradient Descente) pour des paramètres acoustiques "MFCCs.

Ce mémoire est organisé comme suit :

Le premier chapitre est introductif il présente des notions générales sur la parole. Dans le deuxième chapitre est consacré pour système de reconnaissance du locuteur et les algorithmes d'attaques FGSM, PGD. Dans le troisième chapitre la conception du système de reconnaissance du locuteur sécurisé sera détaillée, enfin l'implémentation de notre application sera discutée avec quelques résultats de la simulation.

Enfin, nous terminons par une conclusion générale.

Chapitre I :

Généralités sur la parole

Chapitre I : Généralités sur la parole

I.1 Introduction

La parole est un signal réel, continu, d'énergie finie, non stationnaire.sa structure est complexe et variable dans le temps : tantôt pse udo-periodique pour les sons voisés, tantôt aléatoire pour les sons non voisés.

Le traitement du signal vocal a pour but de fournir une représentation moins redondante de la parole que celle obtenue par l'onde temporelle tout en permettant une extraction précise des paramètres significatifs.

Donc, Le but de ce chapitre est de présenter le mécanisme de la production de la parole et les différentes taches de traitement de la parole.

I.2 Définition de la parole

La parole est une succession de séquences sonores et de silences, et le seul moyen qui permet de communiquer la pensée par un système de sons articulés. Les humains sont les seuls êtres vivants qui utilisent un tel type des systèmes structurés, et il est le résultat d'une variation de la pression produite par l'émission d'un son par un locuteur [1].

I.3 La production de la parole

La production de la parole est l'opération la plus complexe de l'activité biologique humaine est un système dynamique, dont le comportement à un moment donné dépend de ses états antérieurs. Le système est donc dépendant d'une variable paramétrable.

Par définition, le son est ce que l'oreille perçoit de la vibration d'un corps. Cette vibration est une sorte d'onde (produite par un objet, guitare, piano, tambour, marteau, etc., qui se propage par et à travers des corps physiques (air, eau, métal, bois, etc.), La parole se distingue des autres sons par des caractéristiques acoustiques ayant leurs origines dans le mécanisme de production.

Le signal de parole est généré par l'appareil phonatoire. C'est un organe d'une grande complexité mécanique. Il se compose de deux parties anatomiquement distinctes. Les poumons et le larynx, partie supérieure de la trachée artère, constituent l'essentiel du générateur sonore.

Le larynx : Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée artère, se trouve au sommet supérieur de trachée-artère, où la pression de l'air est modulée avant d'être appliquée au conduit vocal. Le larynx est un ensemble de muscles et de cartilages mobiles. Les cordes vocales sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx et, en s'écartant progressivement, déterminer une ouverture triangulaire appelée glotte. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non-voisés. Les sons voisés résultent au contraire d'une vibration périodique des cordes vocales. Le larynx est d'abord complètement fermé, ce qui accroît la pression en amont des cordes vocales, et les force à s'ouvrir, ce qui fait tomber la pression, et permet aux cordes vocales de se refermer [2].

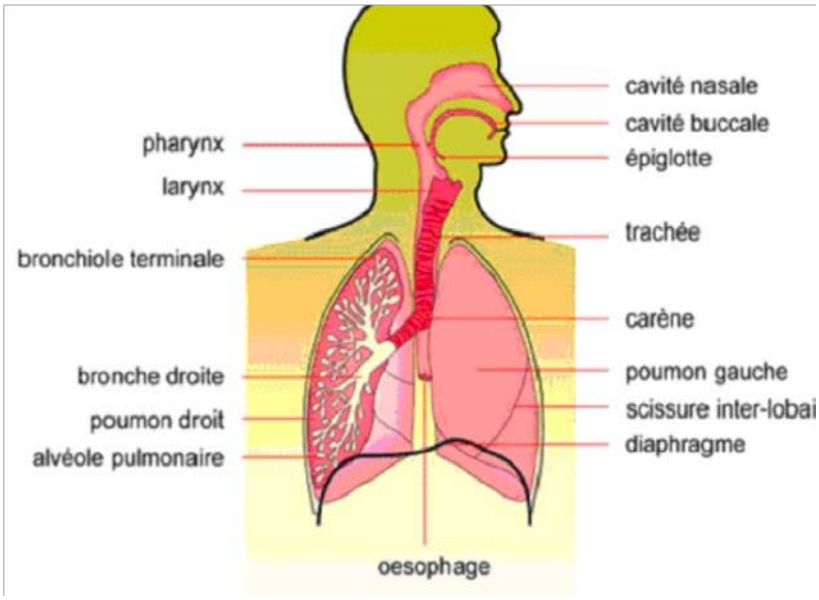


Figure I.1 : Modèle simplifié de l'appareil phonatoire

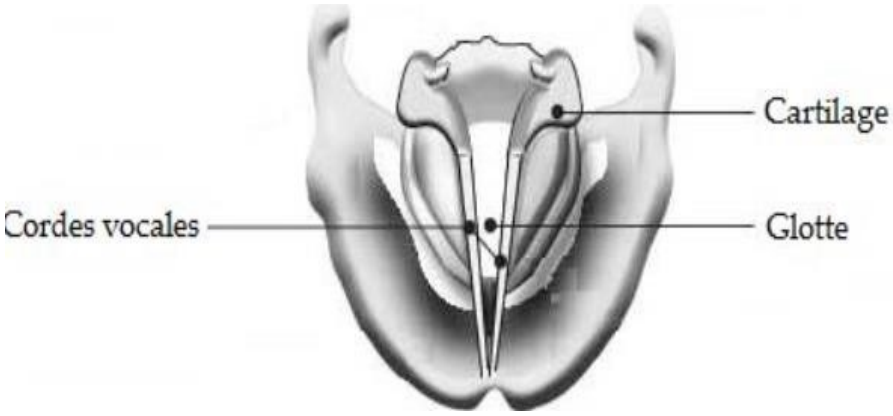


Figure I.2 : Le larynx

I.4 Caractéristiques du signal de parole

La parole est représentée par un signal vocal qui met en évidence des traits acoustiques qui sont : sa fréquence fondamentale, spectre, énergie [3]

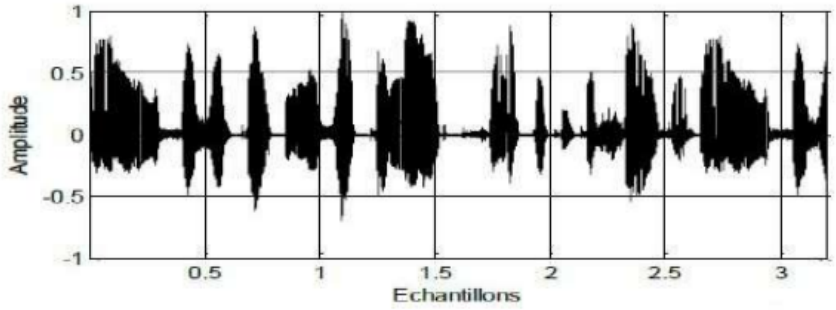


Figure I.3 : Le signal parole

I.5 La voix et le langage

• La voix – ensemble de sons produits par le système phonatoire et articulatoire

• Le langage – fonction d'expression de la pensée et de communication entre les 3hommes, mise en œuvre au moyen d'un système de signes vocaux (parole) et éventuellement de signes graphiques (écriture) qui constitue une langue [4]

I.6 Le cerveau et le langage

Le langage nous aide à structurer le monde en concepts et à réduire la complexité des structures abstraites afin de les appréhender : c'est la propriété de « compression cognitive » • La perception – fonction par laquelle l'esprit se représente les objets [4]

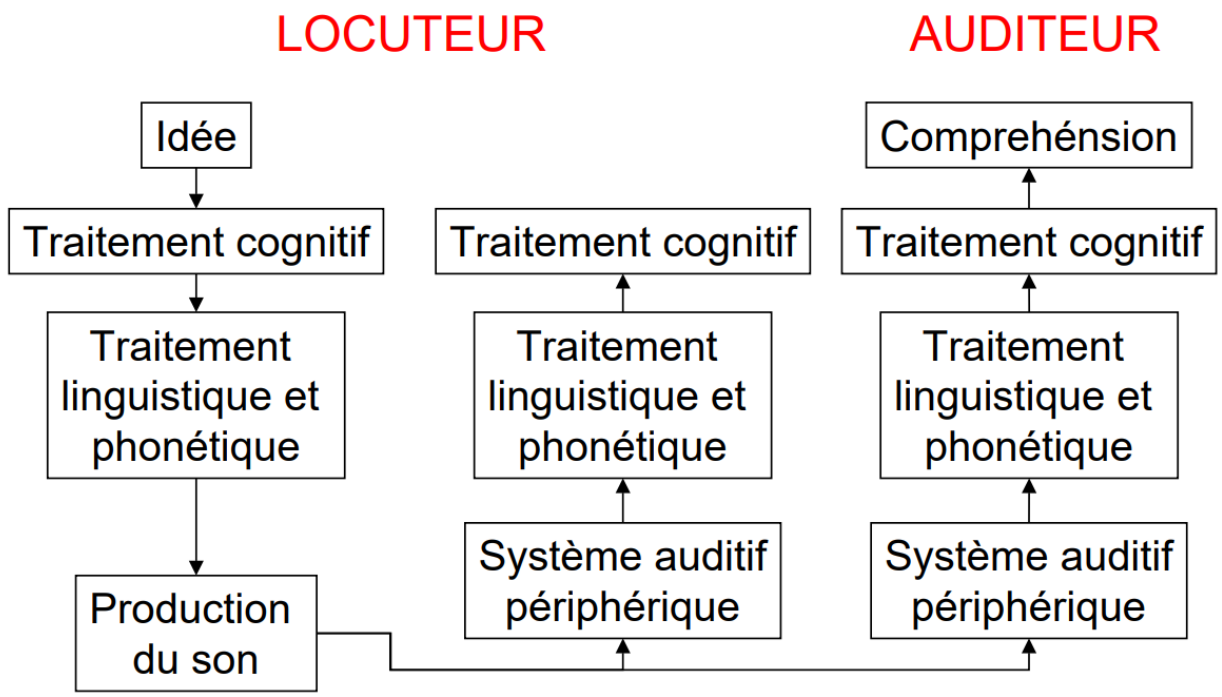


Figure I.4 : Acte de communication parlée

I.7 Traitement automatique de la parole

Est l'ensemble de disciplines dont l'objectif est la captation, la transmission, l'identification et la synthèse de la parole. Ces disciplines rassemblent notamment la reconnaissance de la parole, la synthèse de la parole, l'identification du locuteur ou encore la vérification du locuteur [4]

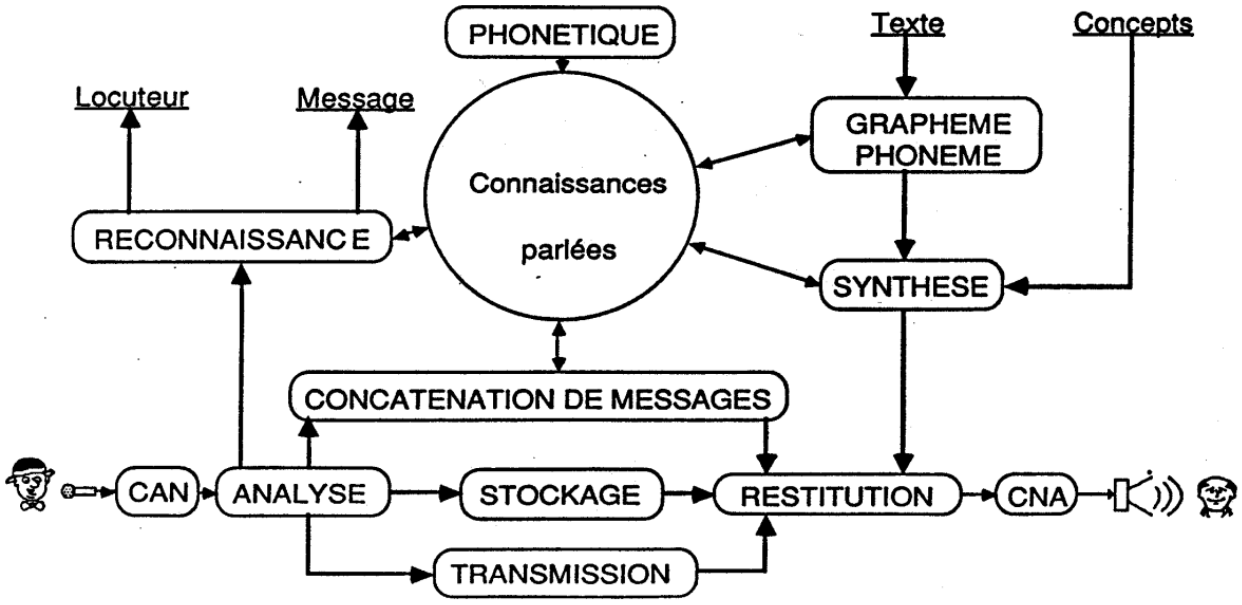


Figure I.5 : Traitement automatique de la parole

I.8 Mécanisme de production du signal de parole

a) L'appareil phonatoire

L'appareil phonatoire nous permet de produire des sons très variés dans un espace fréquentiel et énergétique pourtant limité. La production de la parole est un processus complexe dans lequel sont impliqués de nombreux organes dont les principaux composants responsables de la production de la parole sont illustrés dans le schéma de la figure1.2. L'appareil de phonatoire humain est divisé en trois composantes physiologiques : le système respiratoire (composé des poumons et de la trachée), qui fournit l'énergie et la quantité d'air nécessaire. La deuxième composante représente par le larynx qui permet d'obtenir un flux de nature vibratoire par l'action des cordes vocales. Le conduit vocal, compose des articulateurs, qui permettent de moduler le signal glottique pour obtenir le signal de parole rayonne aux lèvres [5]

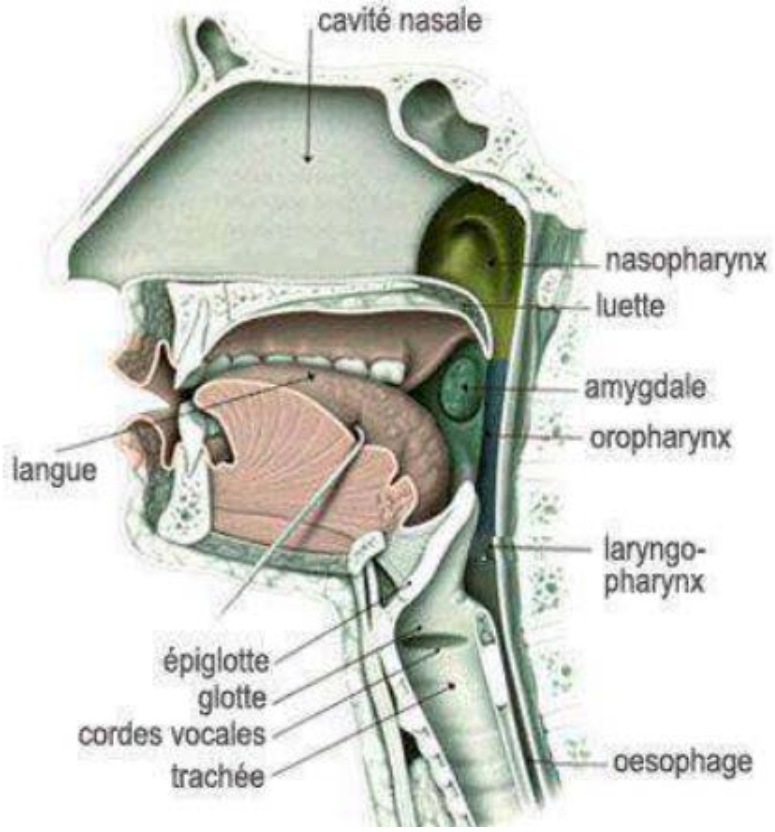


Figure. I.6 : Vue schématique de l'appareil vocal, dans le plan sagittal médian

La parole est une succession d'évènements sonores faisant alternativement apparaître des sons dits voisés caractérisés par la vibration des cordes vocales et des sons non voisés (qui ne font pas intervenir les cordes vocales. Le signal voisé est un signal pseudopériodique présentant des zones fréquentielles plus ou moins importantes. Ces zones fréquentielles d'enveloppe maximale sont appelées des formants [5]

I.9 Variabilité et catégorisation en parole

Précisons d'emblée que notre propos ne consiste en aucun cas à réduire le traitement du langage à un simple processus de catégorisation. Cependant, étant donné l'importance du processus de catégorisation dans tout traitement cognitif, il nous semble indispensable d'analyser sont fonctionnement dans le traitement de la parole. En effet, les catégories sonores du langage sont des catégories dont les caractéristiques physiques sont dépendantes de multiples facteurs. Nous faisons l'hypothèse qu'observer le fonctionnement de ces catégories nous permettra de mieux comprendre le fonctionnement de la parole ainsi que les processus cognitifs en jeu dans le traitement du langage. Depuis quelques années, une synergie de recherches regroupant les sciences humaines et les sciences de la vie s'intéressent aux processus cognitifs présents dans différentes tâches. Les mécanismes mentaux de traitement du langage oral sont, dans cette optique, d'un intérêt capital. [6]

I.10 Synthèse de parole

Dans le cas de la parole le terme "synthèse" implique une combinaison des sons et des bruits et de réalisé des mots ou des phrases.

L'objectif de la synthèse de la parole et de produire des sons de parole à partir d'une représentation phonétique de message [3].

I.10.1 Technique de synthèse de parole

Pour transmettre une information donnée entre un émetteur et un récepteur il faut utiliser des techniques qui permettent une représentation du signal sous forme réduit (codage) tel que, après transmission (stockage),on peut reconstituer le signal original (décodage) à partir de ces informations figure (I.7).

Le principe de synthétiseur est de créer une analogie avec l'appareil phonatoire humain, les différentes techniques offertes par les synthétiseurs de la parole sont :

- Synthétiseur a canaux.
- Synthèse par formant.
- Synthèse basé sur la prédiction linéaire

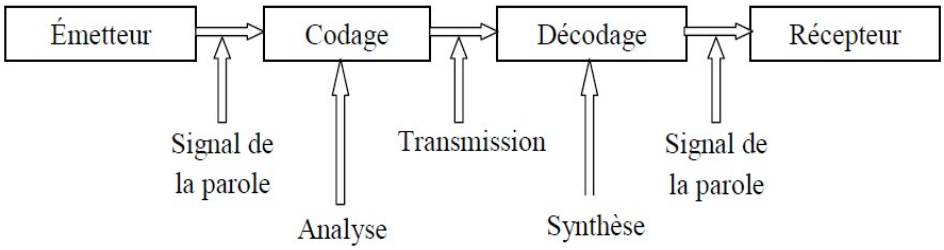


Figure. I.7 : synthèse et codage de parole

I.11 Les sources de variation de la parole

Il nous semble plus juste de parler de "variabilités" au pluriel, plutôt qu'au singulier. En effet, il est hasardeux de regrouper, dans une même catégorie, différentes formes de variations telles que la production de parole dans du bruit, un accent régional, la coarticulation, la situation de parole, etc. Regrouper dans un même ensemble des phénomènes aussi divers peut laisser à penser qu'on leur attribue les mêmes fonctions. Or, s'il est des phénomènes qui sont plutôt inscrits dans le fonctionnement de la parole (spécificité du locuteur, coarticulation, variantes

régionales, sociales, psychologiques), il en est d'autres qui se surajoutent (codage, bruits extérieurs, etc.). Toutefois, quel sens y-a-t-il à dresser une liste des sources de variation de la parole ? Considérer ces phénomènes comme des facteurs de variation, présuppose-t-il qu'il existe un état de parole pour lequel il n'y a pas d'accent régional (il existerait donc une parole standard à partir de laquelle les accents régionaux seraient des dérivés ?), une situation de communication standard (quelle est-elle ?), une production de parole sans bruit (où cela est-il possible, excepté dans une chambre sourde ?), une parole sans coarticulation (par concaténation de phonèmes ?). La communication présuppose un locuteur et un auditeur qui partagent plus ou moins un code commun. Ainsi, ce qui est perçu comme une variation dans un cas (un auditeur breton écoute un locuteur toulousain), ne le sera pas dans une autre (un Toulousain écoute un autre Toulousain). Il n'est pas question ici de dresser une liste des variations possibles à partir d'une hypothétique réalisation canonique des sons du langage. Cette tentative n'aurait probablement aucun sens. [6]

I.12 Conclusion

Dans ce chapitre nous avons exposé des notions de base sur la généralité et la production de la parole, des spécifications du signal vocal. Les objectifs de ce chapitre sont de définir les notions que nous utiliserons dans notre travail.

Chapitre II :

Système de reconnaissance du locuteur et les algorithmes d'attaques FGSM, PGD

Chapitre II : système de reconnaissance du locuteur et les algorithmes d'attaques FGSM, PGD

II.1 Introduction

La parole est un moyen de communication très efficaces et naturel utilisé par l'humain. Depuis longtemps, il rêve de pouvoir s'adresser par ce même moyen a des machines ce qui les rendre plus intelligentes, on est loin d'atteindre le but désiré d'une machine qui peut comprendre le discours parlé par tous les locuteurs dans tous les environnements. Alors, qu'est-ce que l'on entend par la reconnaissance automatique de la parole (RAP) ? D'une autre part l'évolution de la technologie, des ressources de stockage et de calcul, des études dans le domaine de la compréhension du phénomène de production et de perception de la parole, poussent les chercheurs à reconsidérer les préjugés connus jusque-là et essayer de tirer le maximum de ces informations complémentaires pour améliorer les performances du système de reconnaissance du locuteur. Alors dans ce chapitre, on va essayer de donner une idée générale sur la RAP, de discuter les problèmes de cette dernière, ainsi que les méthodes utilisées pour résoudre ces problèmes aussi nous allons décrier le principe de fonctionnement des algorithmes (FGSM : Fast Gradient Sign Method) et (PGD : Projected Gradient Descent) contre les attaques, afin de sécuriser notre système de reconnaissance du locuteur.

.

II.2 L'authentification biométrique

L'authentification est un moyen de vérifier qu'une personne est bien celle qu'elle prétend, sans laisser de place au doute. L'authentification biométrique réalise cette vérification en s'intéressant à des caractéristiques biologiques (comme l'ADN) ou comportementales (comme la voix) ou morphologiques (comme l'empreinte digitale). [7].

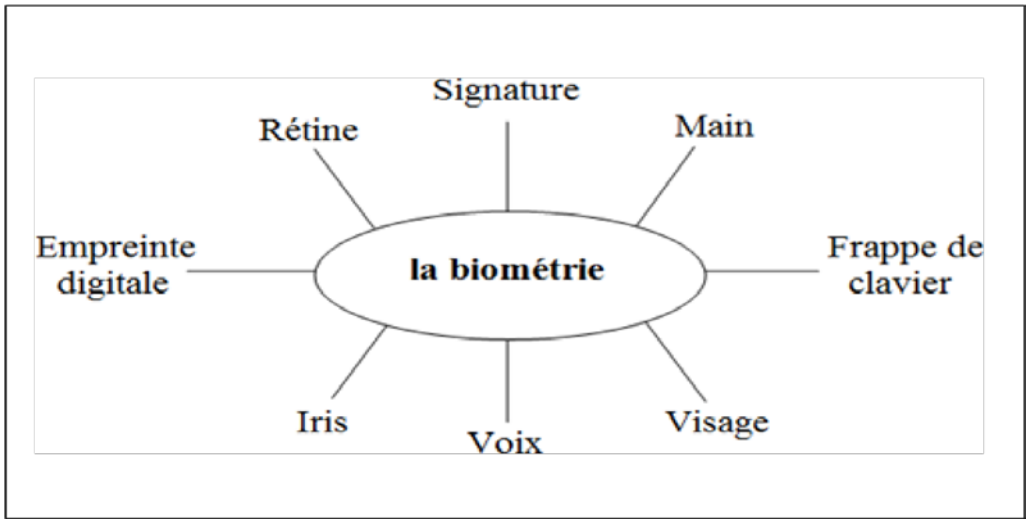


Figure II 1 : Les principales modalités biométriques

Il existe aujourd'hui une panoplie assez large de modalités biométriques et il en apparaît constamment de nouvelles. La **Figure II.1** illustre un exemple de quelque modalité biométriques. . Ces modalités sont basées sur l'analyse des données liées à l'individu et sont généralement classées en trois catégories : La biométrie biologique se base sur l'analyse des données biologiques liées à l'individu (salive, ADN, etc.). La biométrie comportementale se base sur l'analyse de comportements d'un individu (manière de marcher, dynamique de frappe au clavier, la voix, etc.). La biométrie morphologique se base sur les traits physiques particuliers qui, pour toutes personnes, sont permanents et uniques (empreinte digitale, visage, etc.) [7].

II.3 La reconnaissance du locuteur

La reconnaissance du locuteur est la tâche d'authentification de l'identité d'une personne en analysant uniquement sa voix. La reconnaissance automatique du Locuteur –RAL- s'inscrit dans le domaine plus général du traitement de la parole, qui recouvre également un ensemble d'activités (voir **Figure.II.2**), telles que : le codage de la parole, l'analyse, la synthèse et la reconnaissance. Dans la reconnaissance, on peut trouver : la reconnaissance de la parole, la reconnaissance du locuteur et la reconnaissance de la langue.

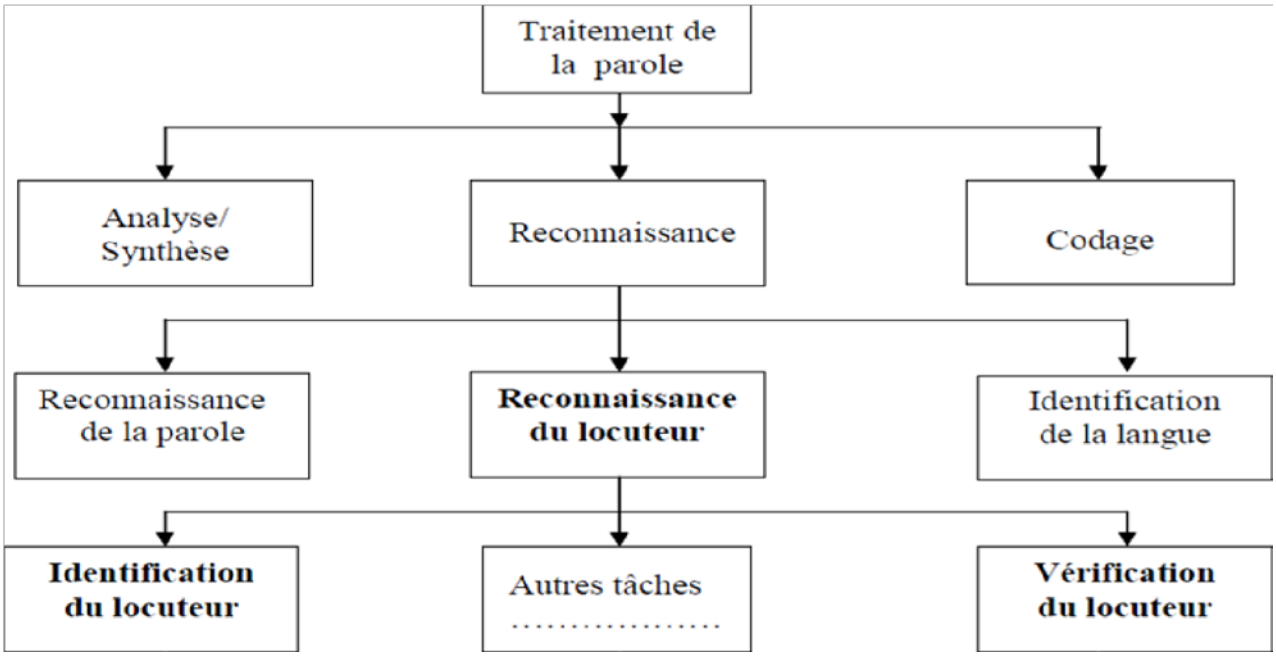


Figure II 2 : Différentes tâches du traitement de la parole [8].

Le discours a également impliqué un large éventail d'activités, telles que : le codage, la compression, synthèse, reconnaissance de la parole et reconnaissance du locuteur. Selon scénarios d'application envisagés. Les scénarios d'application sont divisés en trois catégories principales :

- L'identification de locuteurs.
- La vérification du locuteur.
- L'indexation par locuteur d'un flux audio.

II.3.1 Identification automatique du locuteur

Le principe de l'identification automatique du locuteur consiste à reconnaitre une personne parmi un ensemble de locuteurs en comparant son expression vocale à des références connues. Du point de vue schématiquement, une séquence vocale est donnée à l'entrée du système d'IAL. Pour chaque locuteur connu du système, la séquence vocale est "comparée" à la référence caractéristique du locuteur. L'identité du locuteur dont la référence est la plus "proche" de la séquence de parole est donnée en sortie du système d'IAL [9].

Deux modes sont disponibles dans l'IAL : l'identification de l'ensemble fermé, qui est présumée que la séquence de parole est effectivement prononcée par un locuteur connu et une identification ouverte pour laquelle le locuteur ne peut pas être connu. En mode ouverte «Open

set», le système doit prendre une décision sur la crédibilité de son jugement en acceptant ou rejet de l'identité trouvée. Selon le principe - pour déterminer l'identité parmi identités potentielles - les performances de l'IAL se dégradent lorsque la population de locuteurs augmente [10].

II.3.2 Application IAL

Il n'y a presque pas d'applications d'IAL. On peut continuer par exemple à utiliser système IAL pour s'adapter au locuteur du système de Reconnaissance automatique de la Parole (RAP). De plus, cela peut être intéressant pour les applications, les entreprises associent le même mot de passe à quelques intervenants (Membres de la famille, membres de la société). Dans ces circonstances, les collections ouvertes et liées au texte peuvent être utilisées pour contrôler l'accès aux données sensible [10] et cette application consiste à déterminer l'identité du locuteur parmi une population (ou base) composée de N locuteurs. L'entrée un échantillon de parole d'un locuteur inconnu. La sortie correspond à l'identité du locuteur de la base de référence qui est le plus "proche" du signal de parole d'entrée qu'on trouve ça sur la figure II.3:

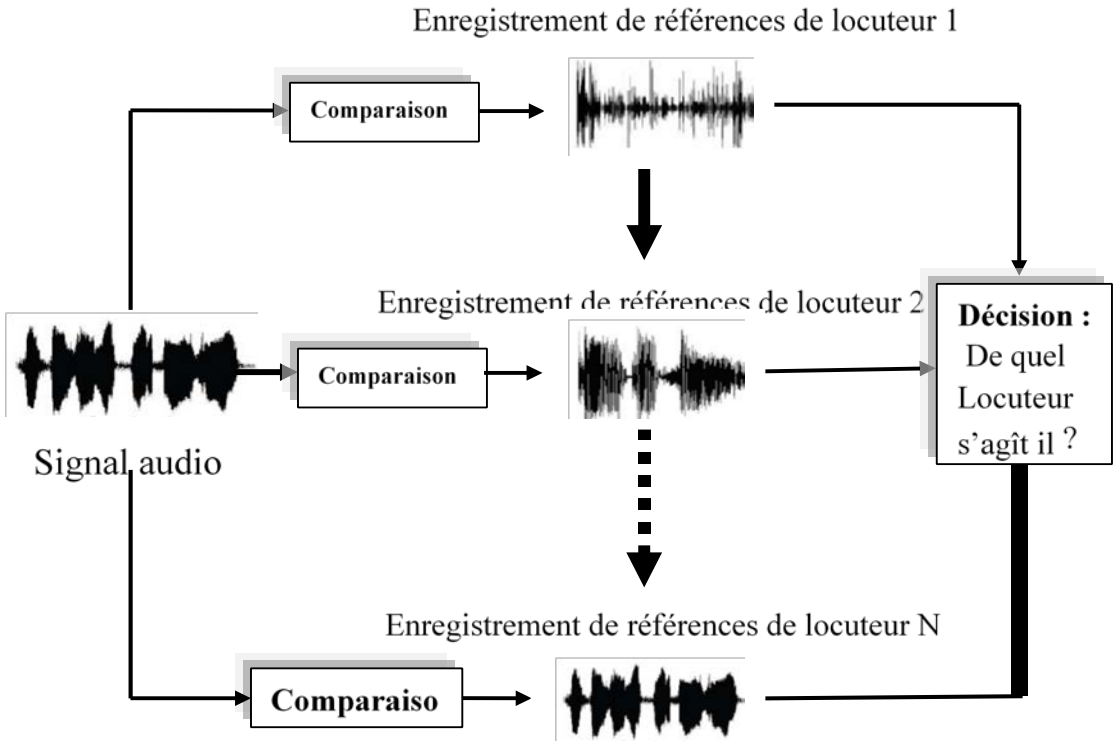


Figure II-3 : Schéma de principe de base de l'identification du locuteur [11]

II.3.3 Vérification automatique du locuteur

La vérification automatique du locuteur (VAL) consiste à accepter ou refuser l'identité proclamée par un locuteur, en se basant sur un modèle qui lui est associé.

Elle traite le scénario où le système prend en entrée un énoncé de test ainsi qu'une identité proclamée. La tâche consiste alors à prendre une décision binaire qui va confirmer ou infirmer le fait que l'enregistrement de test est effectivement prononcé par le locuteur proclamé [12] :

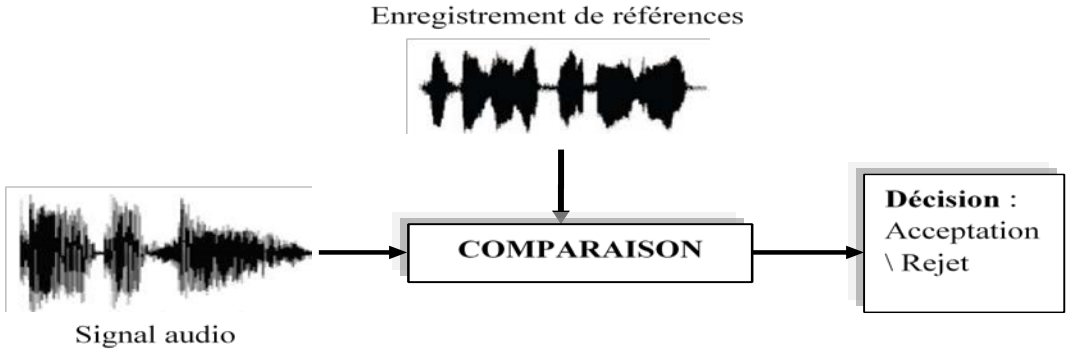


Figure II-4 : Vérification automatique du locuteur [12]

II.3.4 Application VAL

VAL a de nombreuses applications, principalement commerciales par exemple [12]

- ✓ Serrures vocales pour le contrôle d'accès à des locaux.
- ✓ Authentification pour l'accès à distance.
- ✓ Protection de matériel contre le vol (téléphones portables, voitures, etc.).

II.4 La variabilité du signal de parole

Il est important d'introduire les types de variabilités de la parole. En plus du contenu linguistique, le signal vocal véhicule aussi des informations sur : l'identité du locuteur, son genre, son état émotionnel, son état de santé, sa tranche d'âge, etc. Dans une base de données de plusieurs locuteurs où chaque locuteur a au moins un enregistrement, les enregistrements d'un même locuteur peuvent être effectués dans différents lieux (par exemple : au bureau, à la maison, à l'extérieur, dans une chambre insonorisée, etc.) et en utilisant différents types de canaux (par exemple : téléphone fixe, téléphone portable, microphone, etc.). Usuellement, nous pouvons limiter toute cette gamme d'informations ou de variabilités dans ces trois catégories suivantes [13]

II.4.1 La variabilité intra-session : il s'agit de la variabilité perçue au cours d'un même enregistrement et qui peut être causée par un changement de l'état émotionnel, de lieu, d'emplacement par rapport au microphone ou bien du microphone lui-même, etc.

II.4.2 La variabilité inter-session : il s'agit de la variabilité perçue entre l'ensemble des enregistrements d'un même locuteur, elles ainsi des eaux mêmes causes quel

'intra-session, mais elle est susceptible d'être plus importante vu les différences possibles entre les dates et les circonstances d'enregistrement.

II.4.3 La variabilité inter-locuteurs : Les signaux de parole véhiculent plusieurs types d'informations. Parmi eux, la signification du message prononcé est d'importance primordiale. Cependant, d'autres informations telles que le style d'élocution ou l'identité du locuteur jouent un rôle important dans la communication orale. Écouter un interlocuteur permet d'avoir des indications concernant son sexe, son état émotionnel et bien souvent de l'identifier si on l'a déjà entendu. Dans notre vie quotidienne, ces informations sont très utiles. Elles nous permettent, par exemple, de différencier les divers messages que nous entendons selon le locuteur et leur

degré d'importance. Si toutes les voix étaient perçues de la même façon, il serait par exemple impossible de suivre une émission radio faisant participer des personnes différentes. La grande variabilité entre les locuteurs est due, d'une part, à l'héritage linguistique et au milieu socioculturel de l'individu, et d'autre part aux différences physiologiques des organes responsables de la production vocale. L'expression acoustique de ces différences peut être traduite par une variation de la fréquence fondamentale, dans l'échelle des formants (plus haute chez les femmes et les enfants que chez les hommes) et dans le timbre de la voix (richesse en harmoniques due à la morphologie du locuteur et au mode de fermeture des cordes vocales).

II.5 L'architecture d'un système RAL

Un système reconnaissance automatique du locuteur tel qu'il illustré par la Figure I.5, se compose de trois processus principaux, qui sont : la para-métrisation (analyse acoustique), la modélisation et la décision. En plus, un système de RAL possède deux phases [14] :

a) Phase d'apprentissage, où un modèle est estimé pour chaque locuteur « client » du système puis servira de référence pour les tâches de reconnaissance à venir.
b) P hase du test, où l'étape de reconnaissance (vérification, identification...) est effectuée. En sortie de cette phase, les sémèmes une réponse, une identité pour la tâche d'identification ou une décision accès/rejet pour la vérification.

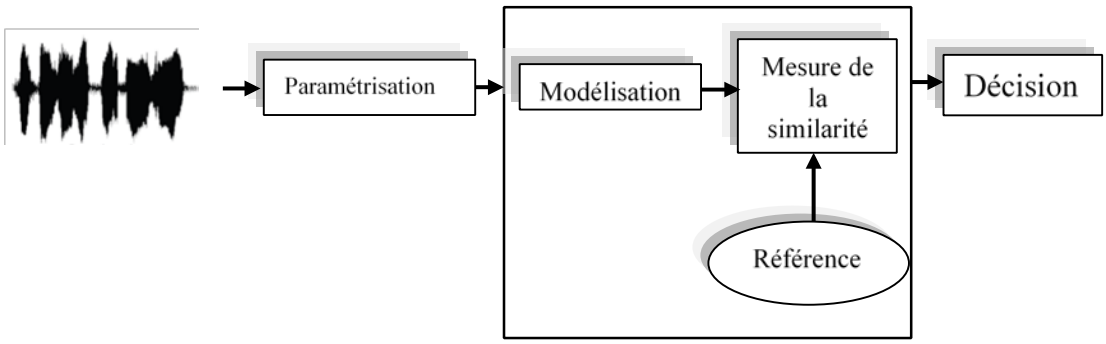


Figure II-5 : Structure générale d'un système RAL [14]

II.5.1 La para-métrisation

Dans cette étape en décompose le signal parole en trames de longueur variantes de (10 à 30ms) car le signal vocal est supposé stationnaire dans cette durée (invariance par changement

d'observation). Pour chaque fenêtre on fait l'extraction des paramètres acoustiques qui peuvent fournir une représentation plus simple du signal parole. Il existe trois grandes catégories de paramètres [15] :

- **Les paramètres prosodiques :** Ces paramètres illustrent le débit, la durée et les fréquences de pauses, pitch et l'énergie [15].
- **Les paramètres acoustiques :** Ces paramètres sont utilisés en RAL. Ils représentent l'influence des caractéristiques physiques de l'appareil phonatoire de chaque individu. Ces paramètres sont les GFCCs et MFCCs [16].
- **Les paramètres dynamiques :** Le vecteur des paramètres issus des para-métrisations précédentes peut être complété par le vecteur correspondant aux dérivées du premier et second ordre de ces paramètres qui sont calculées à partir de plusieurs trames adjacentes. Elles permettent d'introduire une information concernant le contexte temporel d'une trame courante [16].

II.5.1.1 Extraction des paramètres

Le signal de parole présente de la redondance et il est porteur de plusieurs types d'informations, ce qui justifie la recherche d'une représentation spécifiquement pertinente pour la reconnaissance. L'extraction des paramètres consiste à associer au signal de parole une série de vecteurs de paramètres acoustiques. Cette extraction débute par une mise en forme du signal de parole suivie par le calcul de coefficients acoustiques comme indiqué sur la figure I.7 suivante [17] :

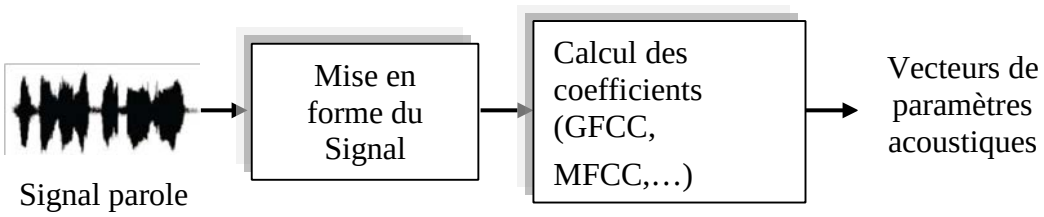


Figure II-6 : Schémas synoptiques de l'extraction de paramètres acoustiques [17].

II.5.1.2 Les paramètres acoustiques

L'analyse spectrale de la parole présente des avantages au niveau de la perception, car l'oreille humaine effectue ce genre d'analyse. On introduit en traitement du signal parole les outils suivants [16] [18] :

a) **Coefficients MFCCs :** Le coefficient MFCC (*Mel Frequency Cepstral Cofficients*) est une technique très utilisée entrainement de la parole. Il est basé sur la variation des bandes critiques de l'oreille humaine avec la fréquence, les filtres espacés linéairement aux basses fréquences et .Ces filtres sont modélisés par une échelle non-linéaire issue de connaissances sur la perception humaine : l'échelle Mel [16].Pour les MFCCs, on utilise la fenêtre de Hamming durant la transformation du domaine et temporel au domaine fréquentiel. Cette transformation est faite en utilisant la transformée de Fourier. Un filtrage, est appliqué ensuite, par banc de filtres triangulaires qui vont chacun couvrir une fréquence, permettant de mieux simuler le fonctionnement de l'oreille humaine.Par la suite les filtres sont espacés selon l'échelle de Mel. Cette échelle reproduit la sélectivité de l'oreille qui diminue avec l'accroissement de la fréquence. Après le calcul de log, une transformée en cosinus discrète est appliquée pour assurer un retour au domaine temporel [17].

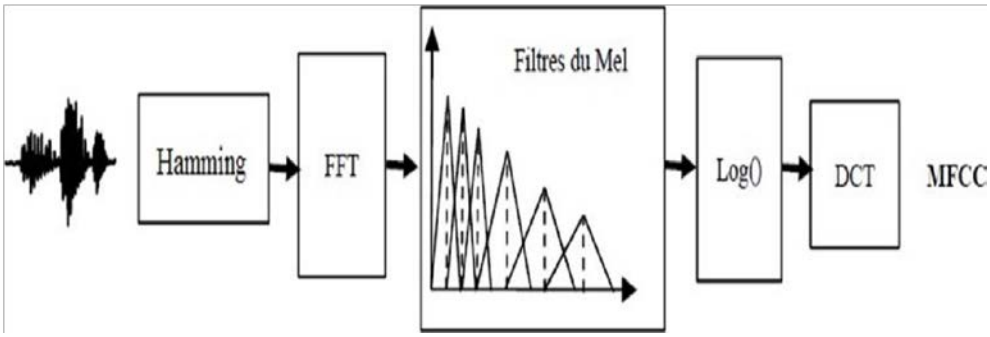


Figure II-7 : Extraction des paramètres MFCCs [14].

II.5.2 Modélisation du locuteur

De manière à modéliser des caractéristiques qui dépendent du locuteur, nous utilisons des algorithmes capables de capturer les points communs entre différentes représentations de motifs spectraux issus du même locuteur (constituant ainsi un modèle du locuteur), tout en ayant la

possibilité de s'adapter aux variations d'échelles fréquentielles et temporelles liées au signal de parole.

Différentes approches ont été développées, néanmoins on peut les classer en trois grandes familles [18] :

- ✓ L'approche vectorielle : le locuteur est représenté par un ensemble de vecteurs de paramètres dans l'espace acoustique. Ses principales techniques sont la reconnaissance à base de DTW (Dynamics Time Warping) et par quantification vectorielle.
- ✓ L'approche statistique : consiste à représenter chaque locuteur par une densité de probabilités dans l'espace des paramètres acoustiques. Elle couvre les techniques de modélisation par les modèles de Markov cachés, par les mélanges de gaussiennes et par des mesures statistiques du second ordre.
- ✓ L'approche connexionniste : consiste principalement à modéliser les locuteurs par des réseaux de neurones.

a)-L'approche connexionniste : Les réseaux de neurones ont été assez largement utilisés en reconnaissance du locuteur. Ils offrent en effet une bonne alternative au problème de la discrimination entre les locuteurs. Ces outils de classification permettent de séparer des classes, dans un espace de représentation donné, de façon non linéaire. L'inconvénient important de l'application de cette technique en identification du locuteur est le coût important lié à l'ajout d'un nouveau locuteur dans la base de référence (ce n'est pas le cas en vérification du locuteur). On peut aussi utiliser les réseaux de neurones en les couplant à d'autres techniques, par exemple les modèles de Markov cachés, on parle alors des méthodes hybrides [17].

b)- L'approche vectorielle (Reconnaissance du locuteur a base DTW) : La reconnaissance par l'alignement temporel par programmation dynamique DTW repose sur le principe que chaque mot est représenté par une prononciation de référence. Compte tenu des décalages temporels entre les différentes prononciations d'un même mot, l'algorithme met en correspondance des séquences de paramètres par distorsion temporelle (Time warping). La programmation dynamique permet d'aligner temporellement une phrase de test avec une phrase d'apprentissage ce qui signifie que c'est une technique exclusivement utilisée en mode dépendant du texte [18]. Ce type de technique a été peu à peu abandonné au profit des modèles séquentiels statistiques (comme les modèles de Markov cachés) qui sont moins "rigides" et donc plus robustes vis-à-vis de la variabilité inhérente au signal de parole.

II.5.3 Quantification vectorielle

La quantification vectorielle est une méthode non paramétrique qui permet de décrire un ensemble de données par un faible nombre de vecteurs formant un dictionnaire associé aux données. Le dictionnaire est en général calculé de telle façon que la distance moyenne entre un vecteur issu des données et son plus proche voisin dans le dictionnaire soit la plus petite possible. La quantification vectorielle est une technique de groupage qui est d'autant plus adapté que les données présentent naturellement des "points d'accumulation" autour duquel la densité de vecteurs issus des données est importante.

Pour la reconnaissance du locuteur, la mesure de similarité entre deux ensembles de vecteurs acoustiques consiste à évaluer la distance moyenne d'un des deux ensembles de vecteurs acoustiques en utilisant le dictionnaire optimisé pour l'autre ensemble de vecteurs acoustiques par quantification vectorielle [16].
La caractérisation de la distribution des données obtenue par la quantification vectorielle est en fait voisine de celle fournie par un modèle de mélange de distributions gaussiennes. Les performances des deux types de systèmes sont donc assez proches. Lorsque les données disponibles pour l'apprentissage sont suffisantes, il semble que le modèle de mélange de densités gaussiennes soit plus robuste.

II.6 Décision et mesure des performances

Un module de reconnaissance fournit en sortie un score. La nature de ce score varie selon les modules de reconnaissance utilisés. Il s'agit la plupart du temps d'une distance, d'une probabilité ou d'une vraisemblance. Un module de décision doit, à partir de ce score, fournir une décision qui constituera la réponse finale du système de reconnaissance. Le module de décision décrit dans le paragraphe précédent reçoit, en entrée et pour chaque test, un score. Celui-ci résulte de la comparaison entre les caractéristiques biométriques de l'utilisateur testé et la référence apprise lors de la phase d'enrôlement. Un score élevé signifiera que la probabilité pour que l'utilisateur testé corresponde à l'identité qu'il annonce est élevée et un score faible signifiera que cette probabilité est faible. La décision binaire qui constitue la sortie du module résulte de la comparaison de ce score avec un seuil défini à l'avance. Si le score est supérieur au seuil, l'utilisateur est accepté et s'il est inférieur au seuil, l'utilisateur est rejeté.

II.6.1 L'identification automatique du locuteur

Consiste à reconnaître un locuteur parmi un ensemble de locuteurs en comparant son identité vocale à des références connues**.** Les performances du système d'identification sont données en termes de taux d'identification correcte *Ic*ou incorrecte *Ii*, soit :

$$I_c = \frac{Nombre\ de\ tests\ correctement\ identifiés}{Nombre\ total\ de\ tentatives} \times 100$$

$$I_i = \frac{Nombre\ de\ test\ sma\ lidentifiés}{Nombre\ total\ de\ tentatives} \times 100$$

$$\text{Avec} I_i + I_c = 100\%.$$

Alors qu'en ensemble ouvert ce sont les faux rejets et les fausses acceptations qui sont mesurées [8].

II.6.2 Vérification automatique du locuteur

Dans un système de vérification automatique du locuteur, une décision d'acceptation ou de rejet de l'identité d'un locuteur susceptible d'être la source de l'enregistrement vocal est prise En d'autres termes est un processus de comparaison entre la voix d'une personne déterminée et l'échantillon de propre voix cette méthode vérifie si la voix de la personne proclamant identité connue correspond bien à la voix de référence de cette personne les performances d'un système de VAL s'évaluent par les critères suivants [13]:

a)-Les fausses acceptations FAR et les faux rejets FRR : Pour évaluer les performances d'un système de vérification, deux types d'erreurs peuvent être commis premièrement le client est rejeté alors que l'identité proclamée est la sienne. Dans ce ça on parle de FR la deuxième erreur survient lorsqu'un imposteur est accepté alors que l'identité annoncée n'est pas la sienne Ce cas est appelé FA Un système biométrique est évalué par les taux des FR et FA (en anglais, False Rejection Rate (FRR) et False Acceptation Rate (FAR)) un bon système de vérification minimise les deux taux d'erreurs FAR et FRR [8] :

$$FAR = \frac{nombre\ d'imposteur\ acceptés}{nombre\ d'imposteur\ totale} \times 100$$

$$FRR = \frac{nombre\ de\ client\ rejetées}{nombre\ totale\ du\ client} \times 100$$

b)- Taux d'Égal Erreur EER

Le plus souvent, les performances du système RAL sont exprimées en taux d'égale erreur Equal Error Rate, (EER) qui correspond à la valeur où le FRR est égal au FAR Plus la valeur de FEER est faible, plus le système est meilleur LEER est utilisé pour comparer les systèmes RAL entre eux.

II.7 Objectifs et Connaissance contradictoires

Les attaques ciblées sont les attaques où l'attaquant tente pour orienter le modèle vers une classe particulière autre que la vraie classe et Les attaques non ciblées sont les attaques où l'attaquant essaie de mal guider le modèle pour prédire l'une des mauvaises classes aussi on distingue Les attaques de boîtes blanches supposent que l'attaquant sait tout sur l'architecture du modèle, les paramètres, ensemble de données de formation, etc. Aussi Les attaques de boîtes noires supposent que l'attaquant peut interroger le modèle mais n'a pas accès aux architectures, paramètres, jeu de données de formation, [19]



Figure II-8 : Exemple d'attaque contradictoire ciblée contre les « gros appareils de cuisine » [21]

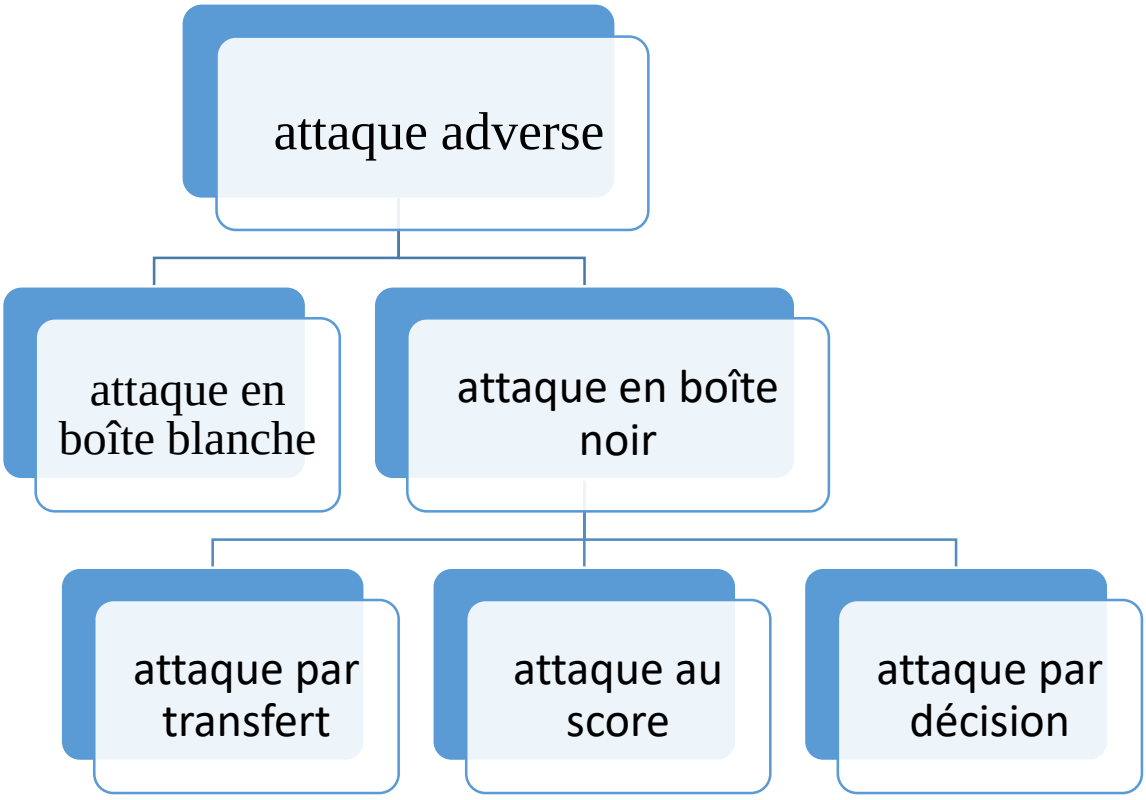


Figure II-9 : Différents types d'attaques adversaires

II.8 Portée de perturbation

- Les attaques individuelles se font en résolvant une optimisation problème pour chaque échantillon. Perturbation adversaire pour chaque échantillon est différente.
- Les attaques universelles se font en résolvant une optimisation problème pour l'ensemble des données. Ajout du même bruit aux différents échantillons d'entrée peut produire une mauvaise classe pour chacun des échantillons d'entrée. [19]

II.9 Adversaire Attaques on IA systèmes

Les attaques d'IA antagonistes posent un défi de sécurité croissant à l'utilisation de l'intelligence artificielle (IA) et de l'apprentissage automatique (ML) dans les entreprises. Ces attaques, également connues sous le nom d'Adversaire ML, trompent les modèles d'IA en leur fournissant de fausses données. Cette forme de cyberattaque a surtout été observée dans la classification des images et la détection des pourriels.

Les risques posés par les attaques d'IA contradictoires comprennent :

• Ingénierie des médias sociaux : Ce type d'attaque « gratte » automatiquement les profils d'utilisateurs sur les plateformes de médias sociaux et crée du contenu automatisé pour attirer le profil cible.

• Deepfakes : Utilisés principalement dans la fraude bancaire, les faux profonds peuvent causer du tort par le chantage et l'extorsion, nuisant à la crédibilité d'une personne, la fraude documentaire et le harcèlement sur les médias sociaux.

• Dissimulation de logiciels malveillants : Les auteurs de menaces utilisent le langage ML pour dissimuler les logiciels malveillants dans un trafic apparemment normal, ce qui rend leur détection difficile pour les utilisateurs réguliers.

• Mots de passe : Les attaques d'IA antagonistes peuvent analyser un grand nombre de mots de passe et générer des variations, rendant le craquage de mots de passe plus efficace. En outre, ML peut être utilisé pour résoudre CAPTCHAs. [20].

II.9.1 Exemples d'attaques d'IA accusatoires

Les attaques d'IA accusatoires permettent aux pirates de manipuler les modèles d'IA pour prendre des décisions incorrectes. Ceci est accompli en introduisant une version corrompue d'une entrée légitime au modèle, on peut voir des exemples d'attaques d'IA accusatoires dans la figure II.3 :

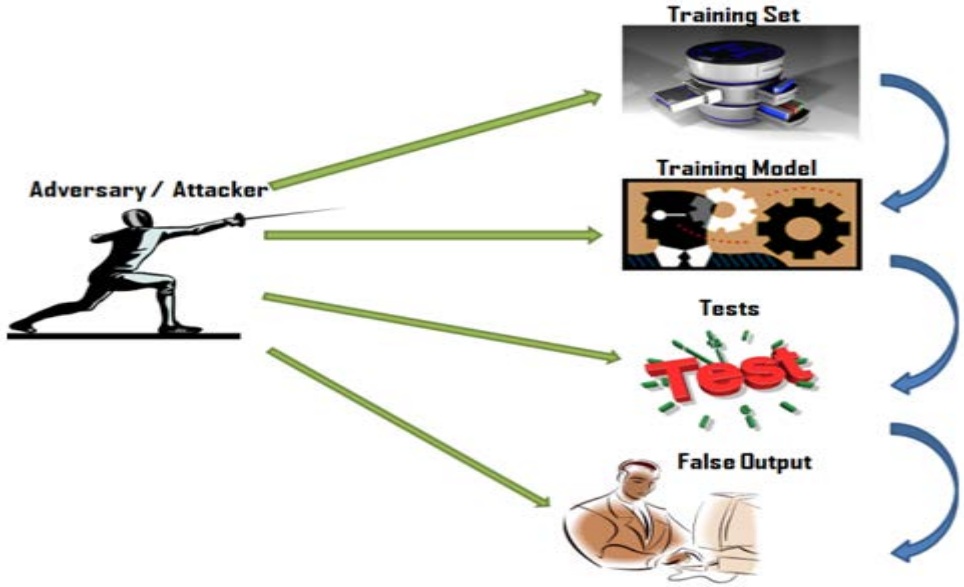


Figure II-10 : Exemples d'attaques d'IA accusatoires

II.10 L'algorithme Fast Gradient Sign Method (FGSM)

La méthode FGSM est une des premières attaques accusatoires introduites dans la littérature [21], et une méthode à une étape. L'attaque tente de changer les prévisions d'un système en trouvant une perturbation qui augmente la perte Lsys. Pour réaliser ceci, un facteur d'échelle λ est utilisé pour effectuer un pas d'échelle dans la direction de gradient de la fonction de perte, c.-à-d.,

$$\delta = \lambda * sign\left(\nabla_x L_{sys}(f(x), y\right) \quad \text{(II.1)}$$

Avec :

- δ : Perturbation contradictoire
- ∇_x : Gradient w.r.t. *x*
- Lsys : fonction de perte spécifique au système (ex., perte d'entropie croisée)
- X: signal original
- Y : Étiquette Ground-truth

Le facteur d'échelle λ contrôle la facilité de trouver une perturbation adversative (plus grande λ) par rapport à son imperceptibilité (plus petite λ). [23]

II.10.1 Fonctionnement d'algorithme FGSM

Nous utilisons l'algorithme FGSM pour attaquer un modèle. D'abord, nous définissons une valeur de magnitude moyenne, puis nous utilisons une attaque ciblée, qui peut améliorer la transférabilité des exemples contradictoires générés par cet algorithme. Nous pouvons utiliser l'algorithme FGSM pour effectuer une attaque de boîte blanche sur le modèle de réseau neuronal. La première étape est de définir une valeur fixe. Ensuite, parce que l'algorithme FGSM est une attaque non ciblée, lors de l'ajout d'une condition de fin d'attaque itérative, nous pouvons utiliser l'algorithme FGSM pour effectuer une attaque ciblée, de sorte que les exemples adversaires générés peuvent être identifiés comme des balises spécifiques par le modèle de réseau neuronal [22].

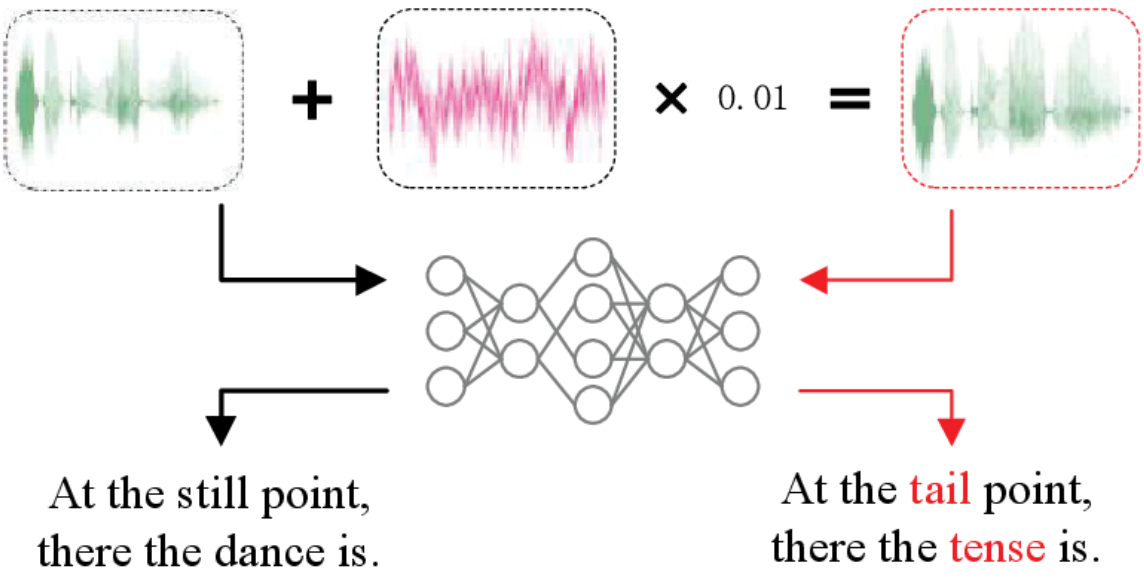


Figure II-11 : Illustration de l'attaque ciblée sur la reconnaissance automatique de la parole [24]

L'attaque FGSM est très efficace pour tromper les modèles d'apprentissage profond. Il est rapide et facile à mettre en œuvre, ce qui en fait un choix populaire pour les chercheurs et les attaquants. Cependant, l'attaque FGSM est limitée dans la mesure où elle ne considère que les perturbations en une étape et ne tient pas compte de la non-linéarité des modèles

d'apprentissage profond. Pour surmonter ces limites, les chercheurs ont proposé des variantes de l'attaque FGSM, comme la méthode itérative de base (BIM) et la méthode de descente de gradient projetée (PGD) et nous on va parler de :

II.11 L'algorithme Projected Gradient Descent (PGD)

La descente en pente projetée (PGD) est une technique largement utilisée pour l'apprentissage automatique contradictoire aussi c'est une attaque de boîte blanche qui signifie que l'attaquant a accès aux gradients du modèle, La taille de ces modèles est déterminée par le facteur η, conduisant à [27].

$$\delta_{ep+1} = clip_{\epsilon}(\delta_{ep} + \eta * sign(\nabla_{\delta_{ep}} L_{sys}(f(x + \delta_{ep}), y))) \quad \text{(II.2)}$$

Avec:

δep : Perturbation during iteration ep.

η : Multiplication factor for updates. [23].

II.11.1 Fonctionnement d'algorithme PGD

PGD fonctionne en modifiant itérativement les entrées dans la direction des gradients de la fonction de perte, pour maximiser la perte pour la classe cible. Aussi c'est une technique puissante pour les attaques adversatives, car il peut provoquer des réseaux neuronaux profonds pour faire des prédictions incorrectes avec une grande confiance. Il est également plus efficace que d'autres attaques basées sur les gradients et il permet de mieux contrôler l'ampleur des perturbations, ainsi qu'il est plus facile de trouver les perturbations optimales pour un modèle cible et une classe cible spécifiques. [24]

II.12 Attaques accusatoires et leurs répercussions

La découverte d'exemples contradictoires met en évidence les limites et la fragilité des réseaux neuronaux profonds (DNNs). Il montre que les DNN ne comprennent pas nécessairement le concept d'un objet comme un humain, ce qui soulève des questions sur leur fiabilité et leur fiabilité dans des applications réelles.

De plus, la vulnérabilité des DNNs aux attaques accusatoires pourrait être exploitée par des acteurs malveillants pour tromper intentionnellement le système. Les chercheurs ont mis en

évidence plusieurs scénarios possibles où des attaques contradictoires peuvent avoir de graves conséquences, notamment :

- Voitures autonomes trompeuses : Eykholt et al. ont mis au point une technique pour créer des "graffitis" artificiels qui peuvent confondre les DNNs généralement utilisés dans les voitures autonomes en une mauvaise interprétation des panneaux de signalisation. Cela pourrait causer l'accident d'une voiture autonome.
- Rendre une personne invisible aux caméras de sécurité : Thys et al. ont créé des "patches adversarial" qui peuvent tromper le réseau neuronal convolutif YOLOv2, qui est utilisé pour la surveillance visuelle. Le patch adversarial rend le DNN incapable de détecter une personne dans l'image.
- Fausses transcriptions d'enregistrements vocaux : Cisse et al. ont attaqué l'application Google Voice-to-text en créant des échantillons audio contradictoires. Les auteurs ont établi que les sujets humains ne pouvaient pas faire de distinction entre l'audio original et l'audio accusatoire, et lorsqu'ils utilisaient Google Voice, le système créait de fausses transcriptions de l'audio accusatoire qui perdaient leur sens.
- Les implications des attaques accusatoires sont de grande portée, et elles montrent l'importance de développer des DNNs robustes et sécurisés pour prévenir les dommages potentiels dans des applications réelles. [22]

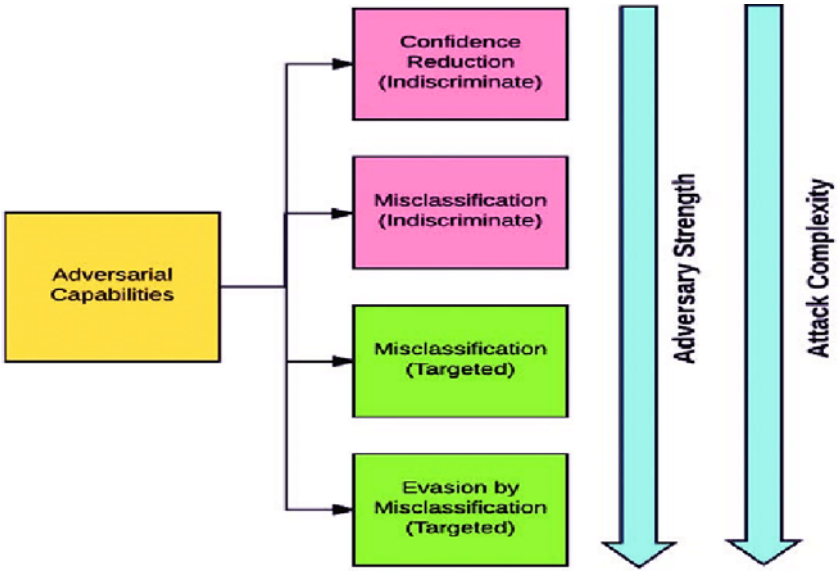


Figure II-12 : Les Impacts des capacités adverses [25]

II.13 Les méthodes efficaces de défense contre les attaques accusatoires

La découverte d'exemples accusatoires a révélé une vulnérabilité dans les réseaux neuronaux profonds (DNN) qui les rend susceptibles à une classification erronée. Les exemples accusatoires sont des intrants fabriqués artificiellement qui sont conçus pour induire en erreur un DNN, l'amenant à faire une prédiction incorrecte. Cette vulnérabilité n'est pas seulement une préoccupation théorique, mais elle a des répercussions réelles, y compris la possibilité que des acteurs malveillants utilisent des attaques accusatoires pour induire intentionnellement en erreur les DNNs à des fins malveillantes Pour se défendre contre les attaques antagonistes, les chercheurs ont développé une variété de méthodes de formation antagonistes visant à doter les DNN de modèles plus robustes qui sont moins vulnérables aux attaques antagonistes. Nous nous référerons ici à quatre de ces méthodes : Augmentation des données de formation, Régularisation, Optimisation et Distillation.

II.13.1 Augmentation des données de formation

Une façon de se défendre contre les attaques antagonistes consiste à générer des exemples antagonistes et à les ajouter aux données d'entraînement d'un DNN. Ce processus est connu sous le nom d'augmentation des données de formation. L'idée est qu'en exposant le DNN à des exemples contradictoires pendant la formation, le modèle apprendra à reconnaître et à classer ces exemples correctement, ce qui le rendra moins vulnérable aux attaques adverses.

II.13.2 Régularisation

Une autre approche de la défense contre les attaques adverses est la régularisation. Cette méthode consiste à ajouter des termes dans la fonction de perte d'entraînement pour orienter la descente en pente vers des paramètres plus résistants aux attaques adverses. Une forme courante de régularisation est la décroissance du poids, qui pénalise les valeurs de poids élevé et aide à prévenir le débordement. Un autre est le décrochage, qui met à zéro une partie des activations du réseau pendant la formation, réduisant ainsi le nombre de fonctionnalités que le réseau utilise pour faire des prédictions.

II.13.3 Optimisation

La troisième approche pour se défendre contre les attaques adverses est l'optimisation. Dans cette méthode, l'objectif de formation passe de minimiser la perte à minimiser la perte maximale réalisable par une attaque accusatoire. Cela force le DNN à apprendre une représentation plus robuste des données qui est moins sensible aux attaques adverses.

II.13.4 Distillation

La méthode finale dont nous discuterons est la distillation. Cette méthode consiste à former un modèle plus petit et plus compact pour reproduire les prédictions d'un modèle plus grand et plus complexe. Le modèle plus petit est ensuite utilisé pour faire des prédictions, réduisant ainsi la taille et la complexité du réseau et le rendant moins vulnérable aux attaques adverses. En supposition, le développement d'attaques antagonistes a soulevé des préoccupations quant à la sécurité et la robustesse des DNNs. Pour se défendre contre ces attaques, les chercheurs ont développé une variété de méthodes de formation antagonistes, y compris l'augmentation des données de formation, la régularisation, l'optimisation et la distillation. En combinant ces méthodes, les organisations peuvent créer des DNN plus robustes qui sont moins sensibles aux attaques antagonistes et assurer la sécurité de leurs systèmes ML. [26]

II.14 Conclusion

A travers ce chapitre, nous avons introduit le principe de la reconnaissance automatique du locuteur ainsi que les différentes étapes d'un système RAL. Le système de reconnaissance automatique du locuteur est probablement la méthode la plus ergonomique pour résoudre les problèmes d'accès. Cependant, la voix ne peut être considérer comme une caractéristique

biométrique d'une personne compte tenu de la variabilité intra et inter -locuteur. Ce système se résume en trois étapes principales : l'analyse acoustique du signal de parole, la modélisation du locuteur et une dernière étape de décision qui se base sur les deux processus d'identification et /ou de vérification de locuteur, et cela quelle que soit l'application ou la tâche visée aussi nous avons présentés deux types des algorithmes d'attaques FGSM et PGD avec une explication détaillée sur son principe et fonctionnement, et nous allons entamer dans le chapitre suivants l'implémentation des deux algorithmes et améliorer la sécurité du système de reconnaissance du locuteur.

Chapitre 3 :

Simulations et Résultats

Chapitre III : Simulation et Résultats

III.1 Introduction

Dans ce chapitre, nous allons présenter les résultats obtenus après l'évaluation de notre système de vérification du locuteur. Notre objectif étant de développer un système de reconnaissance sécurisé basé sur les paramètres acoustiques MFCC par des algorithmes d'attaques FGSM et PGD dans un canal de transmission et environnement réel mobile. Les protocoles de développement et d'évaluation de cette technique de para-métrisation sont décrits dans cette section. Ils mettent en jeu des modules d'extraction des paramètres acoustiques à savoir MFCC_FGSM et MFCC_PGD.

III.2 logiciels et langages utilisés dans cette application

III.2.1 Logiciels Matlab

Nous avons utilisé un logiciel MATLAB (Matrix Laboratory) version 22.02 sous Windows10, Matlab est un langage de programmation émulé par un environnement de développement, il est utilisé à des fins de calcul numérique et commercialisé par la société américaine the Math Works. Ce langage est basé essentiellement sur le calcul matriciel à syntaxe simple, avec des fonctionnalités mathématiques spécialisées et graphiques étendues.

III.3 Description de bases de données utilisées

Les simulations ont été effectuées en utilisant les bases de données NIST-2008.

III.3.1 La description de la base de données NIST-2008

L'évaluation de système de reconnaissance du locuteur fait une partie des séries continue d'évaluations annuelles menées par le NIST (National Institute of Standards and Technology).En effet, depuis 1996 une série d'évaluations SRE (Speaker Recognition Evaluation), dans le but de promouvoir la recherche en reconnaissance automatique du locuteur. La base de données contient des fichiers audio (hommes et femmes) sous format (WAV file), avec une fréquence d'échantillonnage égale à 8 KHz.

La durée des fichiers varie de 10 s À trois minutes. Le répertoire « WAV » est divisé en 2 sous répertoire « Apprentissage » pour créer un modèle du locuteur et le (test) pour la reconnaissance.

- La base de données est passée ensuite à travers le codec standard AMRWB pour aboutir à la base de données transcodée mobile et le canal de transmission est simulé par un bruit additif AWGN avec différent niveau SNR (-5dB, 0dB, 5dB, 10dB).
- La base de données bruitée qui représente environnement réel mobile cette base, qui a été bruitée par le bruit de la foule avec différent niveau SNR (-5dB, 0dB, 5dB, 10dB).

III.4 Étapes de la réalisation des systèmes de vérification du locuteur :

Différentes étapes sont nécessaires à la construction et à la validation de notre système :

- Créer l'ensemble d'apprentissage : la base de données que nous avons utilisée est constituée de 1600 fichiers, prononcés par 267 locuteurs de différents sexes qui répètent la même phrase 6 fois.
- Faire des opérations par l'algorithme tel que : FGSM et PGD.
- Effectuer l'analyse acoustique : l'ensemble des fichiers de la base de données ont été convertis en une série de vecteurs (MFCC d'apprentissage et le test.)
- Faire l'apprentissage des modèles du locuteur cible.
- L'apprentissage d'un sous-espace de variabilité totale à partir des statistiques.
- Créer un modèle GPLDA Gaussien avec des i-vecteurs de développement.
- Calcule le score entre les modèles d'i-vecteurs de développements et le test.
- Faire la reconnaissance et évaluer la performance des systèmes.

III.5 La structure générale de Système de vérification du locuteur

La Figure 3.1 représente le synoptique du système de vérification du locuteur basé sur les différentes techniques de para-métrisation mises en œuvre (MFCC-FGSM, MFCC-PGD,). Dans cette structure, on distingue deux phases : la phase d'apprentissage et la phase de vérification (test)

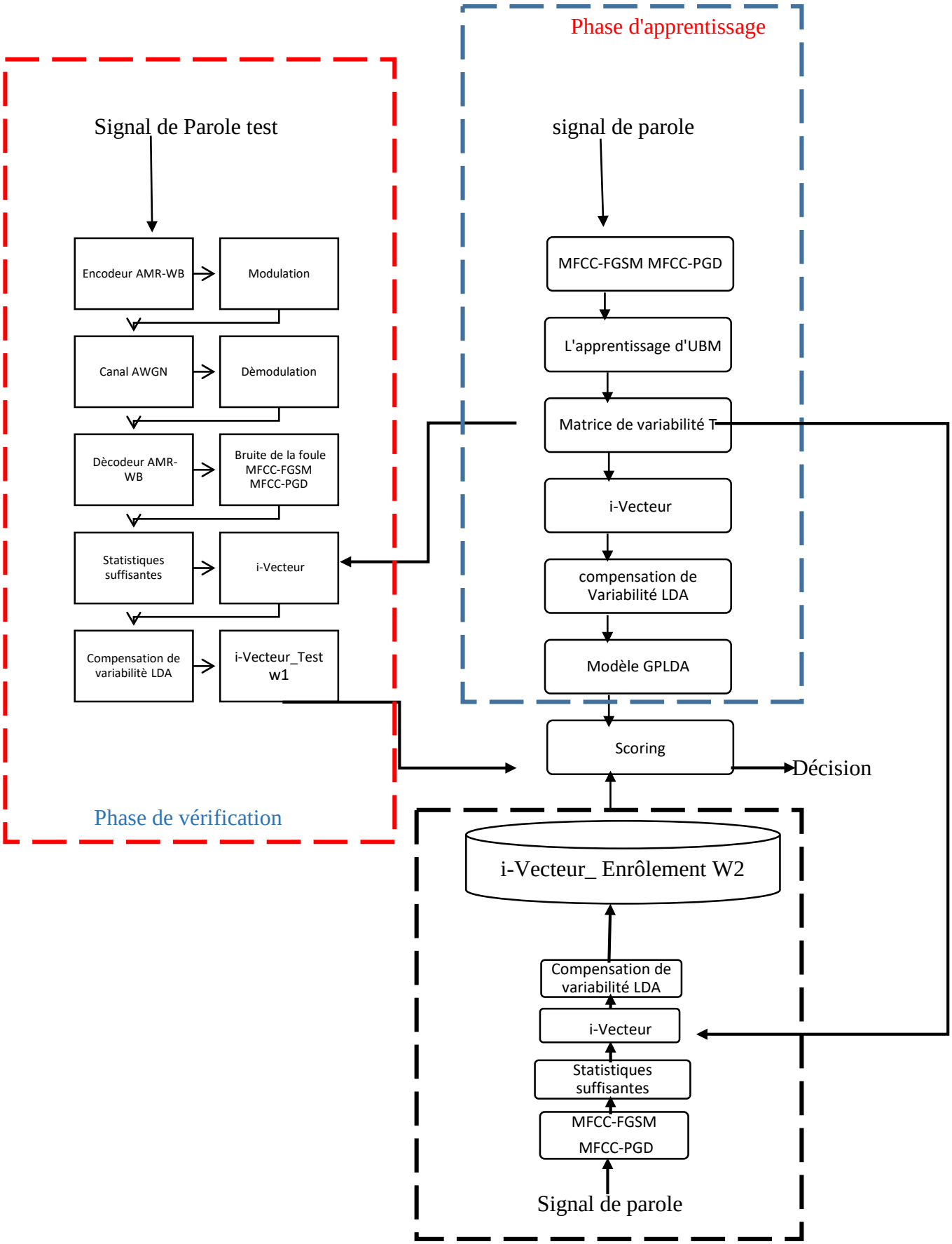


Figure III 1 : Schéma synoptique du système de vérification du locuteur au canal de transmission (AWGN) et environnement réel mobile

III.5.1 L'approche de modélisation GMM-UBM

La modélisation par les mélanges de gaussiennes est devenue la méthode de base pour toutes les approches de modélisation dans le domaine de la reconnaissance du locuteur. Soit $X = \{x_1, x_2, ..., x_T\}_{t=1,...T.}$ une suite de T vecteurs acoustiques (MFCC, GFCC,...etc.) représentant le signal vocal d'un locuteur l et soit λ_{GMM} le modèle GMM composé de k composantes gaussiennes modélisant un locuteur l. Le modèle de mélange gaussien (GMM), chaque locuteur est représenté par des modèles de mélange gaussien.

$$\lambda_{GMM} = \{\mu_k, \sum_k, w_k\}_{k=1,2.....K} \tag{3.1}$$

Où μ_k est le vecteur moyen, $\sum_k$ est la matrice de la covariance, w_k est le poids de la $k^{ième}$ composante du mélange. L'apprentissage d'un GMM selon le principe du maximum de vraisemblance (Maximum Likelihood, *ML*) consiste tout simplement à estimer les paramètres qui maximisent la vraisemblance totale $P(X / \lambda_{GMM})$

$$P(X / \lambda_{GMM}) = \prod_{t=1}^{T} P(x_t / \lambda_{GMM}) = \prod_{t=1}^{T} \sum_{k=1}^{K} w_k^l N(x_t, \mu_k, \sum_k) \tag{3.2}$$

L'estimation par maximum de vraisemblance nécessité une grande quantité de données, pour estimer les paramètres d'un modèle GMM. Dans le cas où la quantité de données n'est pas suffisante pour un apprentissage, on utilise la méthode d'adaptation de modèles ou le modèle du monde (UBM). Cette approche est plus communément appelée GMM-UBM. La méthode d'adaptation la plus répandue dans la littérature est une méthode bayésienne connue sous l'appellation MAP (*Maximum A Posteriori*). Le modèle du monde est donné par

$$\lambda_{UBM} = \{\mu_{k_{UBM}}, \sum_{k_{UBM}}, w_{k_{UBM}}\}_{k=1,2,.....K}$$

Les étapes d'adaptation itérative des paramètres du modèle du monde sont données comme suite :

1. Pour chaque instant t et chaque gaussienne k, nous calculons la probabilité *a posteriori* :

$$P(k / x_t, \lambda_{UBM}) = \frac{w_k N(x_t, \mu_k, \sum_k)}{\sum_{k=1}^{K} w_k N(x_t, \mu_k, \sum_k)} \tag{3.3}$$

2. Nous calculons des statistiques d'ordre zéro, un et deux comme suit :

$$P(k / \lambda_{UBM}) = \sum_{t=1}^{T} P(k / x_t, \lambda_{UBM}) \tag{3.4}$$

$$E[x] = \frac{1}{p(k / \lambda_{UBM})} \sum_{t=1}^{T} P(k / x_t, \lambda_{UBM}) x_t \tag{3.5}$$

$$E[xx'] = \frac{1}{p(k / \lambda_{UBM})} \sum_{t=1}^{T} P(k / x_t, \lambda_{UBM}) x_t x_t' \tag{3.6}$$

3. Ces statistiques servent à estimer les paramètres adaptés

$$\hat{w}_{k_{UBM}} = \left(\alpha_k^w \frac{P(k / \lambda_{UBM})}{T} + (1 - \alpha_k^w) w_k^{\lambda_{UBM}} \right) \gamma \tag{3.7}$$

$$\hat{\mu}_{k_{UBM}} = \alpha_k^w E_k[x] + (1 - \alpha_k^{\mu}) \mu_k^{\lambda_{UBM}} \tag{3.8}$$

$$\hat{\sum}_{k_{UBM}} = \alpha_k^{\sum} E_k[xx']_k + \left(1 - \alpha_k^{\sum}\right) \left(\sum_k^{\lambda_{UBM}} + \mu_k^{\lambda_{UBM}} \mu_k^{\lambda_{UBM}} \right) - \mu_k^{\lambda_{UBM}} \hat{\mu}_k^{\lambda_{UBM}} \tag{3.9}$$

Les paramètres du model du monde sont adaptés aux nouveaux paramètres

$$\hat{\lambda}_{UBM} = \left\{ \hat{w}_{k_{UBM}}, \hat{\mu}_{k_{UBM}}, \hat{\Sigma}_{k_{UBM}} \right\}.$$

Où le paramètre γ assure que la sommation des poids estimés de toutes les gaussiennes du mélange soit égale à un, le facteur de pondération α_k^w sert à régulariser les paramètres du modèle adapté.

III.5.2 Calcul des statistiques suffisantes (Baum-Welch statistics)

Les statistiques suffisantes sont connu par des statistiques d'ordre zéro, un et deux obtenues par un alignement, de types Viterbi ou Baum-Welch pour une suite de *T* vecteurs acoustique $X = \{x_1, x_2, ..., x_T\}_{t=1,....T.}$ et les paramètres de modèle du monde UBM. Les statistiques suffisantes sont représentées respectivement par N_k, Fk et S_k 0

$$N_k = \sum_{t=1}^{T} P(k / x_t, \lambda_{UBM}) \tag{3.10}$$

$$F_k = \sum_{t=1}^{T} P(k / x_t, \lambda_{UBM}) x_t \tag{3.11}$$

$$S_k = \sum_{t=1}^{T} P(k / x_t, \lambda_{UBM}) x_t x_t' \tag{3.12}$$

Où $P(k / x_t, \lambda_{UBM})$ la probabilité *a posteriori* générée par la $k^{\text{ième}}$ *gaussienne* du modèle UBM.

III.5.3 Variabilité totale (i-vecteur)

Les extracteurs d'i-vecteurs sont devenus la technique la plus avancée dans le domaine de la reconnaissance du locuteur. Un espace des i-vecteurs représente des segments de parole sous forme de vecteurs de caractéristiques de petite dimension appelés i vecteurs. La technique par i-vecteur est inspirée par la méthode de modélisation JFA, qui est définie par deux espaces séparés, à savoir l'espace du locuteur et l'espace du canal. Dehak et al, ont observé que les facteurs du canal dans JFA contiennent également des informations dépendantes du locuteur. Cet espace, à l'inverse des deux sous-espaces de JFA, ne fait aucune distinction entre les différents types de variabilité (c.-à-d. la variabilité interlocuteur et celle due au canal). La définition la plus élémentaire d'un i-vecteur est de le considérer comme étant la projection d'un

super-vecteur dans un espace de dimension réduite. Mathématiquement, un super-vecteur *M* d'un segment audio dépendant du locuteur et du canal peut se réécrire sous la forme suivante :

$$M_{s,c} = m + Tw \quad (3.133$$

Où *m* est le super-vecteur moyen de dimension *CF* dépendant du modèle du monde UBM, *T* la matrice rectangulaire qui est un vecteur de faible dimension appelé l'i-vecteur de dimension $CF \times D$ *et* w est un vecteur caché de dimension *D* qui suit une distribution normale standard.

L'approche i-vecteur contient des informations non seulement sur l'identité du locuteur, mais aussi dans une large mesure sur d'autres facteurs telles que les émotions du locuteur, les canaux de transmission et les bruits d'environnement. Il est donc nécessaire d'appliquer des techniques de compensation du canal afin de minimiser ces effets.

III.5.4 Analyse Discriminante Linéaire (LDA)

L'analyse discriminante linéaire ou bien l'analyse de Fisher est une méthode supervisée largement connue dans le domaine de la reconnaissance des formes. Elle est utilisée pour réduire la dimensionnalité d'un espace de variables. L'avantage principal de cette méthode est le fait de tenir compte des étiquettes des classes afin de former un sous-espace (c.-à-d. de dimension réduite) optimal en termes de la classification. L'idée de la LDA est de trouver un ensemble d'axes orthogonaux pour minimiser la variation inter-classes et maximiser la variabilité intra-classes. La fonction objective de la LDA est la suivante :

$$S_b v = \lambda S_w v \quad (3.14)$$

Où S_b et S_w sont les matrices de dispersion inter et intra-classes définies comme suit :

$$S_b = \sum_{s=1}^{L} n_s (\overline{w}_s - \overline{\overline{w}})(\overline{w}_s - \overline{\overline{w}})^T \quad (3.15)$$

$$S_w = \sum_{s=1}^{L} \sum_{i=1}^{n_s} (w_{s,i} - \overline{w}_s)(w_{s,i} - \overline{w}_s)^T \quad (3.16)$$

Où L est le nombre total de locuteurs, n_s le nombre d'énoncés du locuteur et $\overline{\overline{w}}$ la moyenne, elle est définie par la relation suivante :

$$\overline{\overline{w}} = \frac{1}{N} \sum_{s=1}^{L} \sum_{i=1}^{n_s} w_{s,i} \quad (3.17)$$

III.6 Analyse discriminante linéaire probabiliste gaussienne (G-PLDA)

L'approche PLDA est considérée comme un modèle discriminatif à caractère probabiliste visant à minimiser la variabilité intra-individu et maximiser la variabilité inter-individus. Dans le domaine de la reconnaissance du locuteur, la PLDA décompose chaque i-vecteur *w* en une composante dépendante du locuteur, une composante dépendante de la session et une composante résiduelle à savoir :

$$w_{s,r} = \eta + Hz_k + H_1 y_{s,r} + \varepsilon_{s,r} \tag{3.18}$$

Où η est un vecteur de dimension *N* qui correspond à une moyenne globale sur l'espace des i-vecteurs, il représente la composante indépendante du locuteur et du canal.

H Est une matrice désignant les sous-espaces de voix propres (eigen-voices) et H_1 est une matrice de rang faible représentant une base sur le sous-espace canal (eigen-channel), z et y sont les facteurs latents dépendants du locuteur et du canal, respectivement. $\varepsilon_{s,r}$ est une variable aléatoire représentant le bruit résiduel.

Dans ce travail, nous nous intéressons à une variante simplifiée de la PLDA, couramment utilisée en reconnaissance du locuteur appelée PLDA Gaussienne (GPLDA), dans laquelle la variabilité intra-locuteur est modélisée seulement par un bruit résiduel $\varepsilon_{s,r}$. Le modèle génératif du nouvel espace projeté de la variabilité i-vecteurs est donné par l'équation suivante :

$$w_{s,r} = \eta + Hz_s + \varepsilon_{s,r} \tag{3.19}$$

Où le nouveau $\varepsilon_{s,r}$ est un terme résiduel représentant la variabilité intra-locuteur.

III.6.1 Apprentissage du modèle GPLDA

L'apprentissage du modèle GPLDA gaussien donné par l'équation (3.30) consiste à estimer le vecteur moyen η, la matrice des voix propres et la matrice de covariance Λ de la composante du bruit.

III.6.2 Calcul de score

Le score est une décision produite par le système de RAL pour mesurer le degré de similitude entre deux segments vocaux (apprentissage/test) (Ben Kheder W, 2017). Le scoring dans l'espace des i-vecteurs est effectué en utilisant le rapport de vraisemblance qui teste la validité des deux hypothèses (les hypothèses client et imposteur), qui sont données par deux i-vecteurs *w1* et *w2*.

Le scoring par log-ratio de vraisemblance des hypothèses complémentaires s'écrit (Kanagasundaram, A el al, 2014) :

$$score(w_1, w_2) = \log \frac{P(w_1, w_2 / H_s)}{P(w_1, w_2 / H_d)} \quad (3.20)$$

- **Phase de développement :** cette phase consiste en la génération des modèles acoustiques du locuteur (les modèles GPLDA) à partir de la matrice des variabilités totales (i-vecteur) et l'apprentissage de l'UBM à 512 composantes (et à matrices de covariance diagonales) et une matrice de variabilité totale de rang égal à 400, la LDA est appliqué pour réduire la dimension des i-vecteurs à 200.

- **Phase d'enrôlement :** dans cette phase les i-vecteurs d'enrôlement sont extraits à l'aide des statistiques suffisantes et la matrice des variabilités totales**.**

- **Phase de vérification :** la phase de vérification du locuteur via les modèles GPLDA consiste à évaluer un score mesurant le degré de la ressemblance entre deux i-vecteurs (enrôlement et test)

III.7 Extraction des paramètres

III.7.1 Para-métrisation des coefficients MFCC

Le signal de parole est paramétré en utilisant les coefficients MFCC. Ces paramètres sont extraits toutes les 10ms, sur des fenêtres de 25 ms. La fenêtre de pondération utilisée est la fenêtre de Hamming et puis appliquer Douze (20) coefficients MFCC, ils sont calculés pour chaque trame dans l'échelle fréquentielle Mel. Les coefficients différentiels du premier et du second ordre (ΔΔ).

III.7.2 Protocole d'évaluation des résultats

Pour évaluer notre système, nous avons structuré les bases de données comme suivant :

- **La base d'apprentissage :** elle est composée de 1600fichiers vocaux prononcés par 267 locuteurs comprenant les deux sexes. Chaque locuteur répète la phrase 6 fois.

- **La base de test :** elle est composée de 600fichiers vocaux prononcés par 200 locuteurs, comprenant les deux sexes Chaque locuteur répète la phrase 3 fois.

III.8 Simulations et résultats

Pour évaluer la robustesse des systèmes de vérification du locuteur à base de coefficients tels que : MFCC-FGSM, MFCC-PGD, nous avons réalisé des expériences de vérification du locuteur par un type du bruit (canal de transmission AWGN) et le bruit de foule, le bruit AWGN et la foule ont été seulement ajouté pour la base de test, en utilisant les niveaux SNR suivants : -5dB, 0dB, 5dB, 10dB. Pour l'entrainement des modèles acoustiques a été réalisé avec la base de données NIST-2008 non bruitée. Dans la phase de vérification (test), les données de test sont codées par un codeur de la parole AMR-WB.

III.8.1 Les différentes normes de codeurs de la parole

Le but de codage du signal parole est de représenter le signal par le moins de bits possible, mais maintenir un certain niveau de qualité du signal parole décodé. En général, les techniques de codage sont basées sur la réduction de la redondance du signal, afin que le signal résiduel (après réduction de redondance) puisse être encodé avec moins de bits que l'originale, en distingue plusieurs normes de codeur de la parole :

- **Codeur EFR** : Le codeur EFR (Enhanced Full rate) Un exemple important d'un codeur ACELP, il apporte une certaine amélioration de la qualité par rapport aux codeurs FR et HR, notamment dans les communications mobiles. Le codeur EFR utilise 12.2 kb/s pour le codage de la parole et 10.6 kb/s pour la gestion des erreurs.

- **Codeur AMR :** Le codeur AMR (Adaptive Multi rate-AMR) est une famille de codeurs ACELP avec la même structure que l'EFR, mais travaillant à différents débits. À un instant donné. Les codeurs AMR utilisent une longueur de trame de 20 ms et un jeu de coefficient LSP est calculé une fois par image (sauf à 12,2 kbit/s, EFR) en appliquant une fenêtre asymétrique qui met l'accent sur les échantillons plus récents. La fiche d'AMR a également été adoptée par l'UMTS (spécification 3GPP TS 26.090).

III.8.2 Attaques contradictoires

L'attaque contradictoire est représentée par une perturbation imperceptible de la forme du signal de la parole afin modifier la décision du locuteur. Supposons que x ∈T est un signal de parole propre d'origine de longueur T. Un exemple un signal de parole contradictoire de $\check{x}$ est donné par $\check{x} = x + \delta$, où δ est la perturbation contradictoire. Pour résoudre le problème d'optimisation, dans ce travail, nous avons adopté la méthode Fast Gradient Sign (FGSM) et la méthode Projected Gradient Descent (PGD). Les attaques contradictoires sont menées comme suit :

1. Nous construirons un modèle de contre-mesure d'usurpation d'identité en utilisant l'apprentissage LSTM.
2. Nous évaluons les performances anti-usurpation d'identité et générons des exemples contradictoires à l'aide de l'ensemble d'évaluation.
3. Nous ajoutons une perturbation dans les vecteurs de paramètres acoustiques en utilisant la méthode FGSM ou la méthode PGD.
4. Les paramètres acoustiques perturbés sont reconstruits en forme du signal de parole d'original pour attaquer du modèle de contre-mesure.

Nous avons utilisé les réseaux de longue mémoire à court terme (LSTM : Long Short-Term Memory) ou dans la boite à outils (Deep Learning Toolbox). La taille d'entrée la taille de coefficients MFCC c'est-à-dire, num Features. Nous avons utilisé deux couches LSTM bidirectionnelles cachées avec une taille de sortie de 150 et générez une séquence. La Figure.3 représente le signal de parole original et son spectrogramme

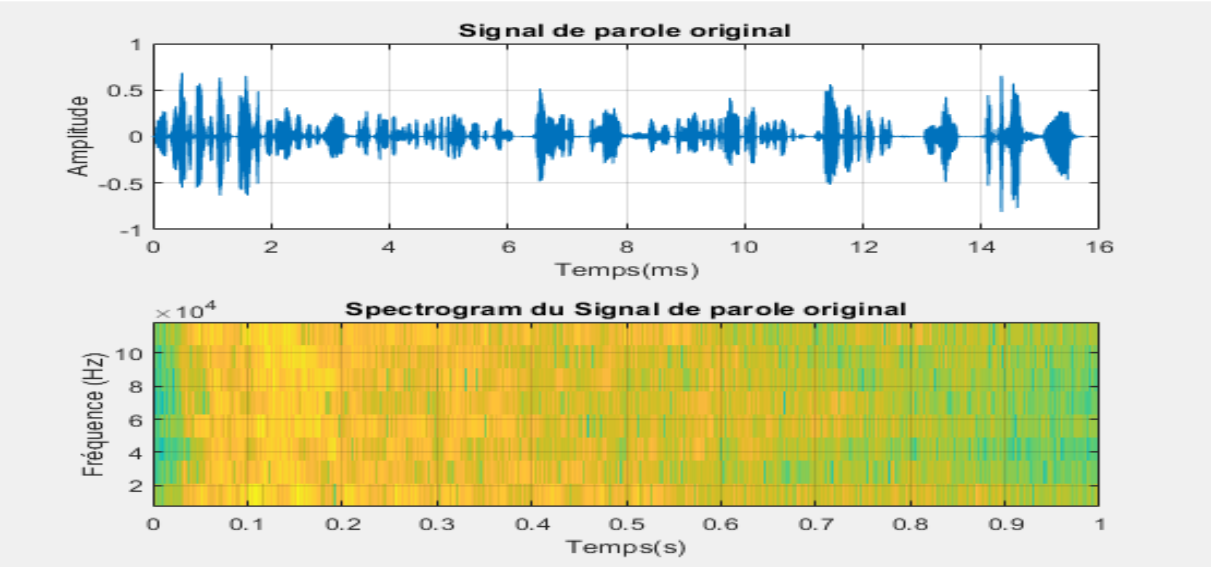


Figure III 2 : le signal de la parole original et son spectrogramme

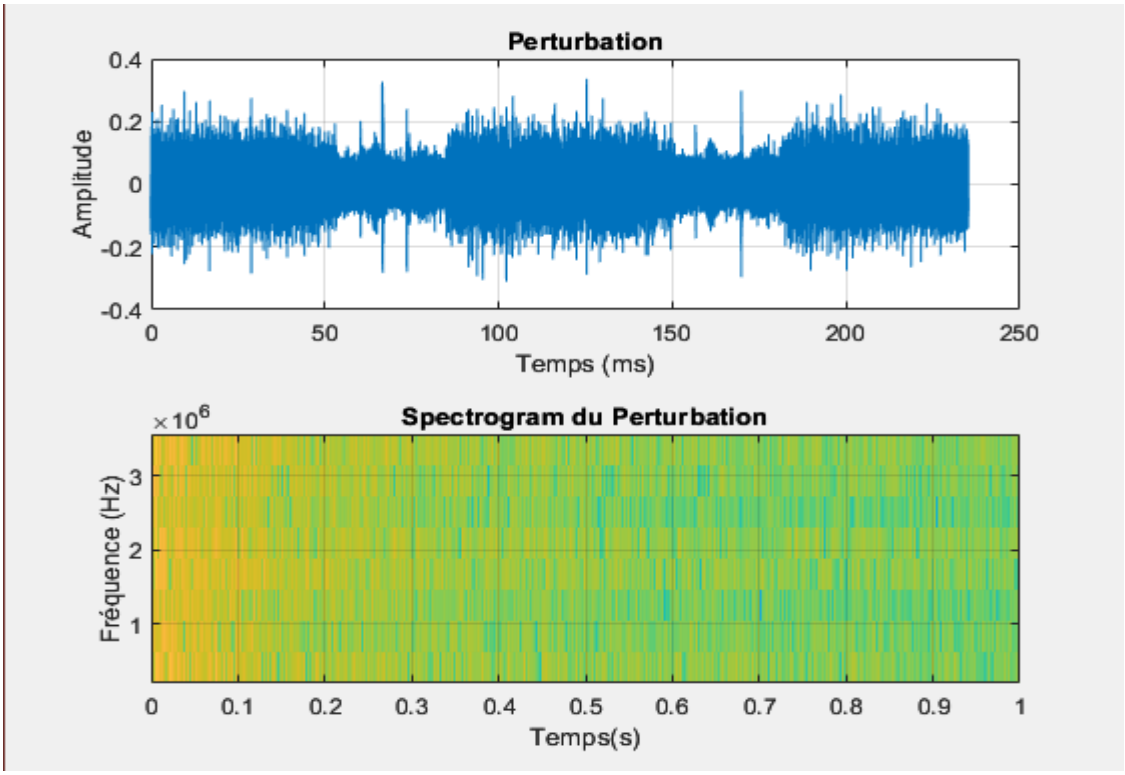


Figure III 3 : Le signal de perturbation et son spectrogramme

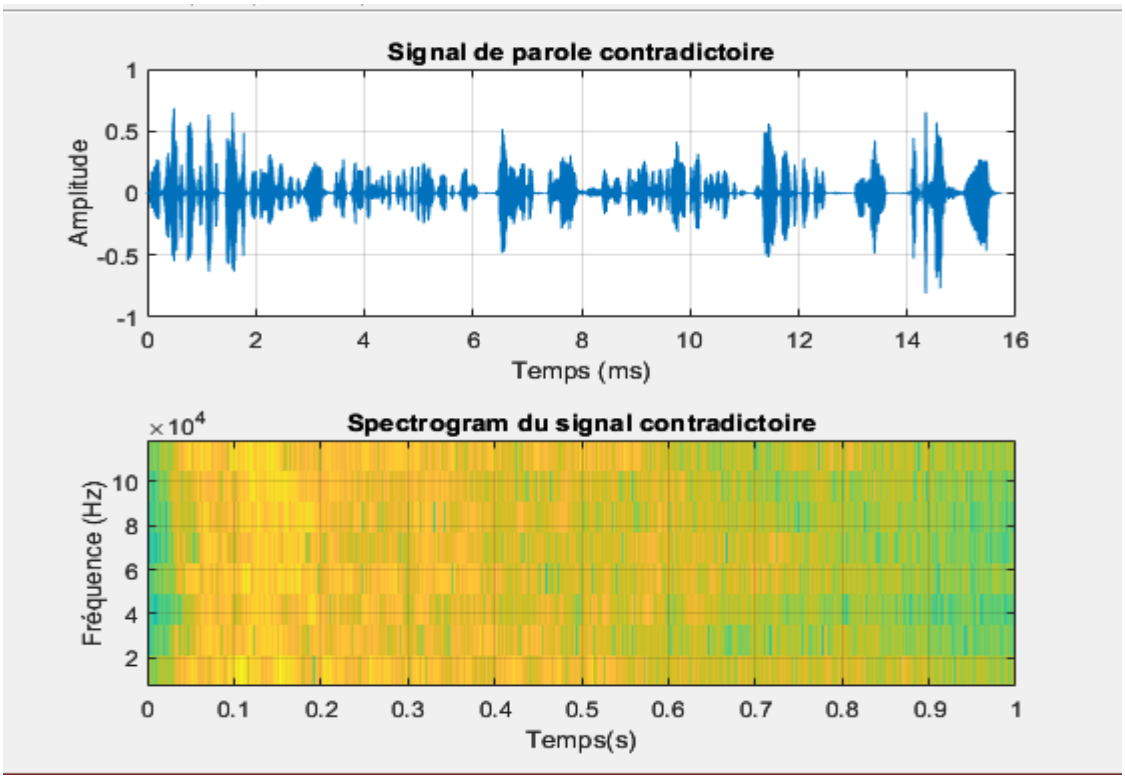


Figure III 4 : Le signal de parole contradictoire et son spectrogramme

III.9 Performance du système de vérification du locuteur en bruit AWGN pour les algorithmes d'attaques adversaires FGSM et PGD

Dans cette partie, nous étudions l'effet du canal AWGN avec différents paramètres (MFCC-FGSM et MFCC-PGD) sur les performances globales du système. Nous évaluons et comparons les performances des différents paramètres en utilisant différents niveaux de SNR (-5, 0, 5, 10dB).). Le Tableau.III.1 et la Figure.III.4 présentent les résultats obtenus par l'algorithme FGSM pour MFCC-FGSM au canal AWGN.

EER (%)				
SNR (AWGN)	-5	0	5	10
MFCC-FGSM	30.10	25.6	20.12	15.60
MFCC-PGD (10 ite)	23.4	21.43	18.75	13.10
MFCC-PGD (20 ite)	20.05	18.40	16.8	14.20
MFCC-PGD (30 ite)	19.5	17.45	13.42	11.82

Tableau III 1 : Les performances du système de vérification en utilisant les paramètres MFCC-FGSM et MFCC-PGD au canal de transmission mobile AWGN

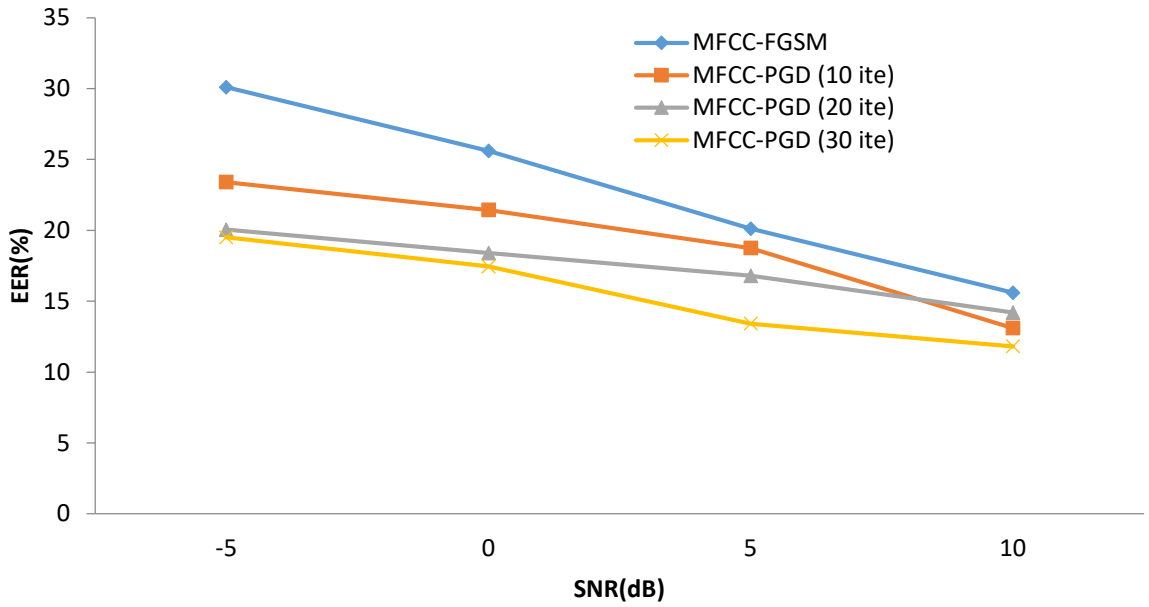


Figure III 5 : Les performances du système de vérification en utilisant les paramètres MFCC-FGSM et MFCC-PGD au canal de transmission mobile AWGN

Les résultats des Tableau.III.1et la Figure.III.5, montrent que les paramètres MFCC-FGSM et MFCC-PGD donnent de meilleurs résultats de reconnaissance en termes de taux EER (%) par rapport à l'algorithme FGSM.

III.10 Performance du système de vérification du locuteur en bruit de la foule pour les algorithmes d'attaques adversaires FGSM et PGD

Dans cette partie, nous étudions la vérification du locuteur en bruit de la foule pour les algorithmes d'attaques adversaires FGSM et PGD avec différents paramètres (MFCC-FGSM et MFCC-PGD) sur les performances globales du système. Nous évaluons et comparons les performances des différents paramètres en utilisant différents niveaux de SNR (-5, 0, 5, 10dB).). Le Tableau.III.2 et la Figure. III.6 présentent les résultats obtenus par les algorithmes d'attaques adversaires FGSM et PGD.

EER(%)				
SNR	-5	0	5	10
MFCC-FGSM	45.12	42.56	34.72	24.56
MFCC-PGD (10 ite)	40.48	37.89	30.45	28.14
MFCC-PGD (20 ite)	39.12	35.72	27.40	24.87
MFCC-PGD (30 ite)	38.45	33.45	22.12	18.79

Tableau III 2 : Les performances du système de vérification en utilisant les paramètres MFCC-FGSM et MFCC-PGD en environnement mobile bruité (bruit de foule)

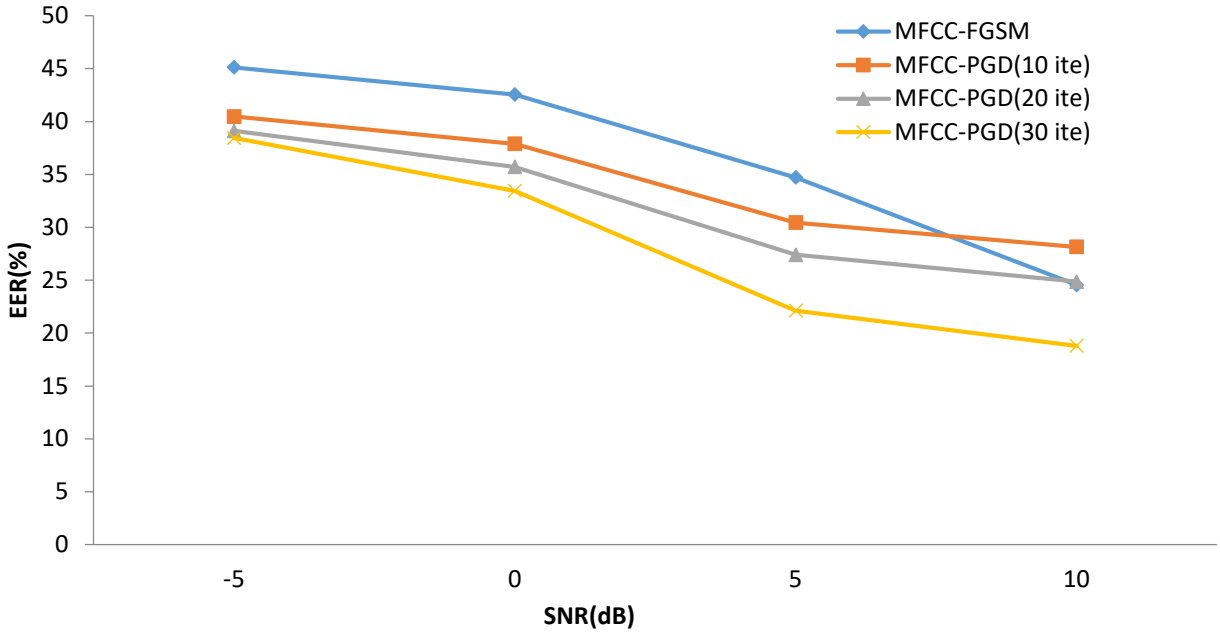


Figure III 6 *:* Les performances du système de vérification en utilisant les paramètres MFCC-FGSM et MFCC-PGD en environnement mobile bruité (bruit de foule).

Les résultats des Tableau.III.2, et la Figure.III.6 montrent que les paramètres MFCC- FGSM et MFCC-PGD donnent de meilleurs résultats de reconnaissance en terme de taux EER (%) par rapport au paramètres de base FGSM et PGD, plus particulièrement à faible niveau de SNR (-5dB et 0 dB). Les résultats obtenus via les paramètres FGSM et PGD sont meilleurs que ceux de les MFCC-FGSM, MFCC-PGD.

III.11 Conclusion

Dans ce chapitre, on a proposé une technique de para-métrisation basée sur les différents algorithmes d'attaque adversaire du canal de transmission FGSM, PGD. La technique proposée améliore les performances du système de vérification du locuteur en utilisant la modélisation i-vecteur-GPLDA. Les performances du système ont été évaluées à l'aide de la base de données NIST 2008. Le système est analysé en présence du bruit de transmission du canal : le bruit gaussien blanc (AWGN) à divers niveaux du rapport signal-bruit (SNR : -5dB à 10dB). L'analyse des résultats obtenus montre que l'utilisation des paramètres proposés pour les bruits des canaux mobiles permet

d’améliorer la performance du système de vérification du locuteur par rapport aux coefficients conventionnels MFCC.

Conclusion générale

Conclusion générale

Dans ce travail nous avons parlé de la parole et ses différentes caractéristiques ainsi qu'on a traité la performance et la sécurité du système de vérification du locuteur dans un environnement mobile à base des paramètres acoustiques attaqué par de différents algorithmes d'attaque. La performance du système de vérification du locuteur est étroitement liée à la qualité du signal parole, les signaux transmis par le réseau mobile montrent une dégradation plus importante à cause de l'effet du canal de transmission.

Pour réaliser le travail dans un environnement mobile, nous avons utilisé une base de données NIST 2008. Cette base de données s'est codée par un codeur AMRWB (Adaptive Multi Rate-Wide Band) pour obtenir une base de données des signaux mobile. Le bruit du canal de transmission a été lui simulé par un bruit blanc additif AWGN (Additive White Gaussien Noise) à différents niveaux SNR (Signal to Noise Ratio), -5dB, 0dB, 5dB, 10dB. Nous avons utilisé comme un vecteur d'entré pour le système de reconnaissance les paramètres : MFCCs avec différents algorithmes d'attaque adversaire MFCC_FGSM et MFCC_PGD.

Les résultats obtenus en utilisant les paramètres acoustiques par des algorithmes d'attaque adversaire FGSM et PGD montrent que l'utilisation de ces paramètres proposés pour les bruits des canaux mobiles permet d'améliorer la performance et la sécurité du système de vérification du locuteur.

Références bibliographique

Bibliographie

[1] P. Taylor and R. Caley, Book, “The architecture of the Festival speech synthesis System”, Edition 1, Jean1998.

[2] N. Schnell ET all. “Synthesizing a choir in real-time using Pitch Synchronous Overlap Add (PSOLA)”. In ICMC. 2000.

[3] Y. Aziza. Modélisation AR et ARMA de la parole pour une vérification robuste du locuteur dans un milieu bruité en mode dépendant du texte. Mémoire de Magister. Université FERHAT ABBAS Setif.2013.

[4] Traitement de la parole Dr. Andrzej Drygajlo andrzej.drygajlo@epfl.ch ELE 233

[5] MEIGNIER, S. Indexation en locuteurs de documents sonores : Segmentation d’un document et Appariement d’une collection. These de doctorat, Université d’Avignon et des Pays de Vaucluse, 2002.

[6] MANUEL, S.Y. AND KRAKOW, R.A. (1984), "Universal and Language Particular Aspects of Vowel-to-Vowel Coarticulation", Haskins Laboratory Status Report on Speech Research, SR-77/78, p. 69-78

[7] M. Mohamad El Abed, « **Evaluation de systèmes biométriques** ». Thèse de doctorat, Spécialité Informatique et applications, Université de Caen/Basse-Normandie.

[8] Yassine Mami, « **Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence** ».Thèse de doctorat, spécialité d’images et de signal, École nationale supérieure des Télécommunications Paris Tech, 2003.

[9] Alexandre Preti. «Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur ». Thèse de doctorat, spécialité d'informatique, Université d'Avignon et des pays de Vaucluse 2008.

[10] Sayoud Halim, « Reconnaissance automatique du Locuteur-Approche connexionniste», Thèse de Doctorat spécialité d'électronique et Électrotechnique, Faculté d'Électronique et D'Informatique USTHB, 2003.

[11] AjgouRiadh, « Reconnaissance automatique du Locuteur à travers les canaux digitaux», Thèse de Doctorat, spécialité d'électronique, Université Mohamed Khider Biskra, 2016.

[12] Antonio M. Peinado, José C. Segura, «Reconnaissance de la parole sur les canaux Numériques : Robustesse et normes », 2006.
www.onlinelibrary.wiley.com/book/10.1002/0470024720, consulté le (14/04/2023)

[13] **Moohammed** Senoussaoui, « **Amélioration de la robustesse des systèmes de reconnaissance automatique du locuteur dans l'espace des i-vecteurs** ». Thèse de doctorat, spécialité génie électrique, École de technologie supérieure université du Québec 2014.

[14] Ben Mahria Bilal, «**L'approche hybride GMM-SVM pour l'identification des locuteurs en mode indépendant du texte en milieu fermé**», thèse de master, spécialité de système intelligent et réseau, Université Sidi Mohamed Ben Abdallah, 2014-2015.

[15] Fredouille, Corinne, « Approche statistique pour la reconnaissance automatique du locuteur : informations dynamiques et normalisation bayésienne des vraisemblances ». Thèse de doctorat, spécialité d'informatique, Université d'Avignon et des Pays de Vaucluse Donneur de grade 2000.

[16] **Ben** Mahria Bilal, «L'approche hybride GMM-SVM pour l'identification des locuteurs en mode indépendant du texte en milieu fermé», thèse de master, spécialité de système intelligent et réseau, Université Sidi Mohamed Ben Abdallah, 2014-2015

[17] Laurent Besacier, «**Un modèle parallèle pour la reconnaissance automatique du locuteur** » thèse de doctorat, Spécialité mathématique et Informatique l'Université d'Avignonet des Pays de Vaucluse, 1998

[18] Claude Barras, « **Reconnaissance automatique des locuteurs pour la structuration en tours de parole** ». Thèse de master, spécialité d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, Université Paris-Sud11, 2011.

[19] **Junde** Wu, Rao Fu, "Universal, transferable and targeted adversarial attacks," arXiv preprint arXiv: 1908.11332, 2019, unpublished.

[20] Enoch Anbu Arasu Ponnuswamy, Securing AI Systems from Adversarial Threats, *Science and Technology*, Vol. 13 No. 1, 2023, pp. 6-11.

[21] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[22] Carlini, N., and Wagner, D. A. (2018). Audio adversarial examples: Targeted attacks on speech-to-text. In Proc. of the IEEE Security and Privacy Workshops, pages 1–7. IEEE

mailto:https://doi.org/10.1109/SPW.2018.00009, consulté le (05/05/2023)

[23] M. Cisse, Y. Adi, N. Neverova, and J. Keshet. Houdini: Fooling deep structured prediction models. arXiv preprint arXiv:1707.05373, 2017.

[24] Kurakin, A.; Goodfellow, I.J. & Bengio, S. Adversarial examples in the physical world. In the International Conference on Learning Representations (ICLR), Toulon, France, 2017.

[25] Sonal Joshi, Jesús Villalba, Piotr Zelasko, Laureano Moro-Velázquez, Najim Dehak. Study of Pre-Processing Defenses against Adversarial Attacks on State-of-the-Art Speaker Recognition Systems. IEEE Transactions on Information Forensics and Security, Vol. 16, 2021.

[26] Xu Li , Jinghua Zhong , Xixin Wu, Jianwei Yu, Xunying Liu , Helen Men Adversarial Attacks on Gmm I-Vector Based Speaker Verification Systems. ICCASP 2020.

Résumé

Ce travail appartient au domaine de la sécurité du système de reconnaissance du locuteur (RAL) appliqué dans un réseau mobile. La sécurité des réseaux mobiles fait référence aux mesures conçues pour protéger l'information sensible stockée et transmise par les smartphones contre les cyber-attaques. Les contradictoires sont conçues pour tromper le système de reconnaissance du locuteur par l'ajout de bruit imperceptible dans les signaux vocaux. Cette opération représente une grave menace pour la sécurité des systèmes de reconnaissance du locuteur en environnement mobile. Il devient ainsi vital d'étudier leurs vulnérabilités. Dans ce travail, nous avons utilisé deux algorithmes de protection contre les cyber-attaques, la méthode du signe du gradient rapide (FGSM : Fast Gradient Sign Method) et la descente du gradient projetée (PGD : Projected Gradient Descente) pour des paramètres acoustiques "MFCCs. Nous comparons les performances de ces paramètres afin de trouver le paramètre le plus robuste avec algorithmes de protection contre les cyber-attaques dans un environnement mobile.

Mots clés :

Reconnaissance du locuteur (RAL), FGSM (Fast Gradient Sign Method), descente du gradient projetée (PGD), Paramètres acoustiques "MFCCs, Environnement mobile.

Abstract

This work belongs in the field of Speaker Recognition System (SRS) security applied in a mobile network. Mobile network security refers to measures designed to protect sensitive information stored and transmitted by smartphones from cyber-attacks. Adversaries are designed to fool the speaker recognition system by adding imperceptible noise to speech signals. This operation represents a serious threat to the security of speaker recognition systems in a mobile environment. In this work, we used two algorithms to protect against cyber-attacks, the Fast Gradient Sign Method (FGSM:) and the Projected Gradient Descent (PGD) for speaker recognition based on feature MFCCs We compare the performance of these parameters in order to find the most robust feature with algorithms to protect against cyber-attacks in a mobile environment.

Keywords:

Speaker Recognition (SRS), FGSM (Fast Gradient Sign Method), Projected Gradient Descent (PGD), Acoustic Parameters "MFCCs, Mobile environment.

ملخص

يشير أمان شبكة الهاتف .المطبق في شبكة الهاتف المحمول (RAL)ينتمي هذا العمل إلى مجال أمان نظام التعرف على السماعات .المحمول إلى الإجراءات المصممة لحماية المعلومات الحساسة التي يتم تخزينها ونقلها بواسطة الهواتف الذكية من الهجمات الإلكترونية تمثل هذه .تم تصميم التناقضات لخداع نظام التعرف على السماعات عن طريق إضافة ضوضاء غير محسوسة إلى إشارات الكلام .وبالتالي يصبح من الضروري دراسة نقاط ضعفهم .العملية تهديدًا خطيرًا لأمن أنظمة التعرف على السماعات في بيئة الهاتف المحمول طريقة تسجيل :FGSM)في هذا العمل ، استخدمنا خوارزميتين للحماية من الهجمات الإلكترونية ، طريقة تسجيل التدرج السريع

هذه المعلمات من أجل العثور على أكثر .للمعلمات الصوتية (هبوط التدرج المتوقع :PGD)ونسب التدرج المتوقع (التدرج السريع
.المعلمات قوة مع الخوارزميات للحماية من الهجمات الإلكترونية في بيئة الهاتف المحمول

:الكلمات الدالة

، المعلمات الصوتية (PGD)، هبوط التدرج المتوقع (طريقة إشارة التدرج السريع) FGSM، (RAL)التعرف على السماعات
"MFCCs ،. بيئة الهاتف المحمول