République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Polytechnique
Département d'Electronique
Laboratoire Signal et Communications

Thèse de Doctorat en Sciences
Spécialité Electronique

Présentée par
Mr Salim DJEGHIOUR
Magister en Electronique CRSTDLA - Alger

Thème

Reconnaissance Automatique du Locuteur en vue de la Criminalistique dans les Conditions Bruitées

Soutenue publiquement, le : 11/07/2022

Devant le jury composé de :

Président	Mr LARBES Chérif	Professeur	ENP - Alger
Rapporteur	Mme GUERTI Mhania	Professeur	ENP - Alger
Examinateurs	Mme DJERADI Rachida	Professeur	USTHB
	Mme FALEK Leïla	Professeur	USTHB
	Mme CHELALI Fatma Zohra	Maître de conférences	USTHB

ENP 2022

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Polytechnique
Département d'Electronique
Laboratoire Signal et Communications

Thèse de Doctorat en Sciences
Spécialité Electronique

Présentée par
Mr Salim DJEGHIOUR
Magister en Electronique CRSTDLA - Alger

Thème

Reconnaissance Automatique du Locuteur en vue de la Criminalistique dans les Conditions Bruitées

Soutenue publiquement, le : 11/07/2022

Devant le jury composé de :

Président	Mr LARBES Chérif	Professeur	ENP - Alger
Rapporteur	Mme GUERTI Mhania	Professeur	ENP - Alger
Examinateurs	Mme DJERADI Rachida	Professeur	USTHB
	Mme FALEK Leïla	Professeur	USTHB
	Mme CHELALI Fatma Zohra	Maître de conférences	USTHB

ENP 2022

ملخص

يقدم هذا العمل خوارزمية محسنة بهدف تحسين الكلام تسمى الحد الأدنى لمتوسط الخطأ المربع MMSE، استنادا الى طيف تأخير المجموعة المعدل MODGD، للتعرف التلقائي على المتكلمين لغرض جنائي FASR في البيئات الصاخبة. تستخدم هذه الخوارزمية طيف MODGD بدلا من طيف السعة، و هذا لحساب طيف الطاقة للإشارة التالفة بسبب الضوضاء. في المقدر المقترح، يحتفظ MODGD بمعظم معلومات التصاميم، لذلك فإنه يعزز إشارة الكلام الصاخبة بجودة عالية حتى عند مستويات منخفضة للغاية من نسبة الإشارة على الضوضاء SNR. تم اجراء تقييم الخوارزمية المحسنة في سيناريوهات FASR عن طريق إضافة مستويات ضوضاء مختلفة، مستخرجة من قاعدة البيانات NOISEX-92 الى آثار NIST2000 النظيفة. أظهرت النتائج التي تم الحصول عليها أن مقدر MMSE-MODGD المقترح يسمح بإزالة أكبر لمكونات الضوضاء في المناطق ذات نسبة الإشارة على الضوضاء SNR المنخفضة مقارنة بمقدر MMSE، بالإضافة الى ذلك، هناك انخفاض مهم في احتمالية النسبة المتساوية EPP (التحسينات 1.84٪ لضوضاء الثرثرة و 1.25٪ لضوضاء المصنع و الضوضاء البيضاء)، بإستعمال مقدر MMSE-MODGD المقترح مقارنة بالمقدر التقليدي MMSE.

الكلمات المفتاحية: التعرف التلقائي على المتكلم لغرض جنائي، البيئات صاخبة، مقدر الحد الأدنى لمتوسط الخطأ المربع، طيف تأخير المجموعة المعدل، احتمالية النسبة المتساوية.

Abstract

This work presents an improved speech enhancement algorithm called Minimum Mean Square Error (MMSE), based on the MODified Group Delay spectrum (MODGD), for Forensic Automatic Speaker Recognition (FASR) under noisy environments. This algorithm uses the MODGD instead of the amplitude spectrum, to compute the power spectrum of the noise-corrupt signal. In the proposed estimator, the MODGD retains most of the formants information. Therefore, it enhances the noisy speech signal with high quality even at extremely low Signal-to-Noise Ratio (SNR) levels. The evaluation of the improved algorithm in simulated FASR scenarios was performed by adding different noise levels, extracted from the NOISEX-92 database to the clean NIST2000-traces. The results obtained show that the proposed MMSE–MODGD estimator provides greater suppression of noise components in regions of low SNR than the MMSE estimator. In addition, there is an important reduction in Equal Proportion Probability (EPP) (the improvements are 1.84 % for babble noise and 1.25 % for factory and white noises), combining FASR techniques with the proposed MMSE–MODGD estimator than with the conventional estimator.

Keywords: Forensic Automatic Speaker Recognition, MMSE estimator, MODGD spectrum, Noisy environments, Equal Proportion Probability.

Résumé

Ce travail présente un algorithme amélioré, pour le rehaussement de la parole appelé Erreur Quadratique Moyenne Minimum (MMSE), basé sur le Spectre de Retard de Groupe Modifié (MODGD), pour la Reconnaissance Automatique du Locuteur en vue de la criminalistique (FASR) dans des environnements bruités. Cet algorithme utilise le MODGD au lieu du spectre d'amplitude, pour calculer le spectre de puissance du signal corrompu par le bruit. Dans l'estimateur proposé, le MODGD conserve la plupart des informations sur les formants. Par conséquent, il améliore le signal vocal bruité avec une qualité élevée, même à des niveaux de Rapport Signal sur Bruit (SNR) extrêmement bas. L'évaluation de l'algorithme amélioré dans des scénarios FASR simulés a été réalisée en ajoutant différents niveaux de bruit, extraits de la Base de Données NOISEX-92 aux traces NIST2000 propres. Les résultats obtenus montrent que l'estimateur MMSE-MODGD proposé permet une plus grande suppression des composantes de bruit dans les régions à faible SNR que l'estimateur MMSE. De plus, il y a une réduction importante de la Probabilité à Proportions Egales (EPP) (les améliorations sont de 1,84 % pour le bruit de bavardage et de 1,25 % pour les bruits d'usine et blanc), combinant les techniques FASR avec l'estimateur MMSE-MODGD proposé, par rapport à l'estimateur MMSE conventionnel.

Mots clefs : Reconnaissance Automatique Criminalistique du Locuteur, Estimateur MMSE, Spectre MODGD, Environnements bruités, Probabilité de Proportion Egale.

Dédicaces

Je dédie cet humble travail à :

Mes chers Parents ;

Ma chère Femme.

Au nom d'Allah, le Tout Clément, le Tout Miséricordieux

REMERCIEMENTS

Je tiens tout d'abord à exprimer ma haute reconnaissance envers ma Directrice de thèse Mme GUERTI Mhania, Professeur au Département d'Electronique, à l'Ecole Nationale Polytechnique d'Alger, pour avoir bien voulu me proposer un sujet, de diriger cette thèse. Son aide, ses précieux conseils, sa grande disponibilité et sa gentillesse ne sont que quelques unes de ses nombreuses qualités. Je la remercie aussi pour le soin qu'elle a apporté à la finalisation de ce manuscrit, pour ses remarques et orientations qui ont énormément contribué à son amélioration. J'aimerais bien qu'elle sache que ma gratitude va au-delà de ces quelques lignes, pour tout ce qu'elle a fait pour moi !

Aussi, j'exprime ma haute gratitude à M LARBES Chérif, Professeur à l'ENP-Alger, pour l'honneur qu'il me fait en acceptant de présider mon jury de thèse ;

J'adresse également mes remerciements aux membres du jury, Mme HAMAMI Latifa, Professeur à l'ENP-Alger et l'ESDAT, Mme DJERADI Rachida et Mme FALEK Leila, Professeurs à l'USTHB, ainsi qu'à Mme CHELALI Fatma Zohra, Maître de conférences à l'USTHB, pour avoir bien voulu examiner et juger ce travail ;

De plus, je remercie vivement Mr. ASBAI Nassim, maitre de conférences à l'université de Bordj Bou-Arréridj pour ses aides, ses précieux conseils, et son soutien constant ;

Tous ceux qui m'ont aidé de près ou de loin trouvent ici l'expression de mes profonds remerciements.

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABREVIATIONS

ADN	Acide Désoxyribo Nucléique
ANC	Adaptive Noise Cancellation
BD	Base de Données
BSD	Bark Spectral Distorsion
CD	Cepstral Distance
CDF	Cumulative Distribution Function
DCT	Discrete Cosine Transform
DSP	Densité Spectrale de Puissance
DTW	Dynamic Time Warping
EER	Equal Error Rate
EM	Expectation Maximisation
EPP	Equal Proportion Probability
FA	Fausse Acceptation
FASR	Forensic Automatic Speaker Recognition
FAR	False Acceptance Rate
FFT	Fast Fourier Transform
FR	Faux Rejet
FRR	False Reject Rate
GMM	Gaussien Mixture Models
HMM	Hidden Markov Models
IAL	Identification Automatique du Locuteur
IS	Itakura Saito Measure
LLR	Log-Likelihood Ratio
LPC	Linear Prediction Coding
LPCC	Linear Prediction Cepstral Coefficients
LR	Likelihood Ratio
MAP	Maximum A Posteriori
MBSD	Modified Bark Spectral Distorsion
MFCC	Mel Frequency Cepstral Coefficients
ML	Maximum-Likelihood
MMSE	Minimum Mean Square Error
MODGD	MODified Group Delay
MOS	Mean Opinion Score

NSS	Non-Linear Spectral Subtraction
OLA	OverLap and Add
OLS	OverLap and Save
PDF	Probability Density Function
PESQ	Perceptual Evaluation of Speech Quality
$PMEH_0$	Probabiliy of Misleading Evidence in favour of hypothesis H_0
$PMEH_1$	Probabiliy of Misleading Evidence in favour of hypothesis H_1
QV	Quantification vectorielle
RACL	Reconnaissance Automatique Criminalistique du Locuteur
RAL	Reconnaissance Automatique du Locuteur
RNA	Réseaux de Neurones Artificiels
ROC	Receiver Operating Characteristic
SNR	Signal-to-Noise Ratio
SS	Spectral Subtraction
SSOM	Spectral Subtraction with Over subtraction
SVM	Support Vector Machine
TFD	Transformée de Fourier Discrète
TIC	Taux d'Identification Correcte
UBM	Universal Background Model
VAD	Voice Activity Detection
VAL	Vérification Automatique du Locuteur
WSS	Weighted Spectral Slope

INTRODUCTION GENERALE

Dans certaines affaires criminelles, la voix enregistrée à l'aide d'un appel téléphonique est le seul indice dont disposent les enquêteurs. Il y a donc une demande très pressante et parfaitement justifiée de la part de la police judiciaire et des magistrats d'utiliser ces enregistrements audio pour guider l'enquête, et pour établir la culpabilité d'un suspect ou pour prouver son innocence en s'appuyant sur des techniques de reconnaissance criminalistique du locuteur.

Un système de la Reconnaissance Automatique Criminalistique du Locuteur (RACL) ou Forensic Automatic Recognition System (FASR) est considéré comme l'une des disciplines de la Reconnaissance Automatique du Locuteur (RAL), que se soit dans l'Identification ou Vérification du locuteur. Bien que, plusieurs applications RAL aient été développées au cours de dernières années, elles n'ont cependant pas donné les résultats escomptés, en raison de la grande complexité et de la variabilité du signal de parole, les différences entre les conditions de modélisation et de test, principalement dans la vie quotidienne. Les dernières peuvent provenir de sources diverses, telles que la réverbération, l'audio compressé, les canaux dégradés et les bruits de l'environnement qui dégradent énormément les performances du système RACL. Par conséquent, la tâche difficile pour les experts forensiques est de développer des algorithmes efficaces, en vue d'améliorer la parole contre les environnements hautement bruités, tel que le bruit additif.

Pour surmonter ce défi, plusieurs algorithmes d'amélioration de la parole, basés sur le spectre d'amplitude du signal de parole, ont été développés à savoir : Méthode de Soustraction Spectrale (SS), Soustraction Spectrale avec modèle de sur-soustraction (SSOM), Non-Soustraction Spectrale linéaire (NSS) et estimateurs de l'Erreur Quadratique Moyenne Minimum (MMSE).

Dans ce travail, nous proposons un estimateur amélioré appelé MMSE basé sur la fonction de retard de groupe modifiée (MODGDF) dédié pour un système de Reconnaissance Automatique Criminalistique du Locuteur (RACL) dans des environnements bruités, en proposant une modification de l'estimateur MMSE, en remplaçant le spectre de l'amplitude estimé à l'aide de la Transformée de Fourier (TF), par le spectre MODGD. En d'autre terme, nous dérivons les variables aléatoires gaussiennes indépendantes du spectre MODGD, au lieu de leur estimation directe à partir de la Transformée de Fourier Discrète, dans le but d'améliorer l'estimateur MMSE en exploitant les informations contenues dans les spectres de phase.

Notre motivation derrière la modification proposée est double :

- Le conduit vocal d'un locuteur est un système de phase minimum, et pour les systèmes de phase minimum, l'information peut être extraite de la phase ou du spectre d'amplitude. Ainsi, en termes de son analyse, le retard de groupe d'un signal de phase minimum est la somme de retard de groupe de ses composantes de phase minimum. Le MODGD peut être extrait directement du signal de parole. L'application dudit algorithme dans les domaines, tels que l'identification des locuteurs a donné des résultats très satisfaisant, vue la robustesse des propriétés de MODGD (haute résolution des formants) dans le cas où le signal de parole est corrompu par un bruit additif.

  L'approche proposée par Lu et Loizou [1] fonctionne très bien pour les cas où SNR $\geq$ 10. Cependant, pour SNR $<$ 10, il a été constaté l'apparition de bruit musical dans le signal de parole débruité, ce qui est considéré comme un défi du système FASR. Par contre, l'utilité de MMSE-MODGD réside sur la réduction de bruit, notamment le bruit musical, même à des SNR faibles, ce qui permet, par conséquent, d'améliorer les performances du système FASR.
- Il a été démontré que même à valeurs des SNR faibles dans des environnements bruités, le spectre de la fonction MODGD conserve la plupart des informations sur les formants. En d'autres termes, le spectre de MODGD est moins affecté par le bruit que le spectre d'amplitude.

Par conséquent, la contribution de notre travail est triple :

- Les estimateurs MMSE améliorés basés sur MODGD conviennent mieux aux segments de parole de trace bruités ;
- Exploiter les informations contenues dans la phase ainsi que dans le spectre d'amplitude pour le MMSE-MODGD proposé ;
- Essais approfondis et validation expérimentale du MMSE proposé.

Notre thèse est composée de 4 chapitres :

- le premier, présente des généralités sur la biométrie, aussi nous avons exposé un aperçu sur les différentes technologies biométriques les plus utilisées, et finalement les principaux domaines d'application ;
- le deuxième définit les notions fondamentales sur la RACL et ses grands axes, tout en commençant par les systèmes RAL, nous aborderons aussi l'approche Bayésienne utilisée pour pouvoir évaluer la preuve scientifique ;

- le troisième est consacré aux meilleures techniques de l'amélioration de la parole, notamment le Maximum de vraisemblance (ML), MMSE, log MMSE, Maximum A Posteriori (MAP), Incorporating Speech presence Probability in MMSE (MMSE-ISP), Incorporating Speech presence Probability in log MMSE (log MMSE-ISP), ainsi que l'estimateur Wiener. Finalement, nous aborderons les techniques d'évaluation des méthodes d'amélioration du signal de parole, en ce qui concerne les mesures subjectives et les mesures objectives ;
- Au quatrième chapitre, nous avons abordé notre l'approche proposée MMSE-MODGD et puis nous procédons à la mise en œuvre de ladite approche, pour cela, nous détaillons les différentes étapes de cette réalisation telles que la paramétrisation, la construction du Modèle Universel (UBM) et les modèles des suspects à l'aide de la technique MAP, la phase de test, en utilisant trois types de bruit (bavardage, usine et blanc), ensuite nous allons appliquer l'estimateur conventionnel MMSE et notre estimateur amélioré MMSE-MODGD dans les conditions propres et bruitées avec différentes valeurs du SNR. Finalement, nous terminerons par une comparaison entre les deux techniques utilisées en termes de performances.

Nous terminons cette thèse par une conclusion générale sur les résultats obtenus et des perspectives.

CHAPITRE 1 : GENERALITES SUR LA BIOMETRIE

1.1 Introduction

Avec le développement international de l'évolution technologique et son impact sur le domaine de TIC (Technologie de l'Information et de Communication) et l'ampleur de ce dernier en terme du flux d'informations et ses exubérances (les transactions financières, l'accès aux différents services, etc.), implique nécessairement de s'assurer de l'identité des personnes, soit en identification ou vérification pour faire face à toutes les formes de criminalité et de fraude d'informations.

La sécurité n'a pas de prix, c'est pour cela les gouvernements ont commencé à accorder des budgets colossaux pour assurer la sécurité des personnes et leurs biens. Dans plusieurs pays au monde, les caméras de vidéosurveillance sont installées dans les endroits publics : Autoroutes, aéroports, ports, grands boulevards, etc. les systèmes de reconnaissances faciales ont été déployés au niveau des stades pour identifier les individus perturbateurs, et des distributeurs de billets sont mis à dispositions de leurs clients.

Les scanners biométriques, lecteurs d'empreintes digitales, microphones, caméras intelligentes, etc., sont de plus en plus perfectionnés. Ces dispositifs utilisés pour la capture des informations biométriques génèrent une grande quantité d'informations qui nécessite des ordinateurs puissants et des logiciels appropriés pour pouvoir traiter et analyser rapidement ces informations et prendre la décision au temps opportun.

Dans ce chapitre, nous allons donner quelques notions sur la biométrie telles que : sa définition, son historique, ses caractéristiques, et leurs domaines d'applications, ainsi que son évolution sur le marché international. Ensuite, nous expliquons les systèmes biométriques, le principe général de leurs fonctionnements, ses avantages et inconvénients. Nous terminerons ce chapitre par une comparaison entre les différentes technologies de la biométrie.

1.2 Définition de la biométrie

Le mot biométrie signifie littéralement « mesure du vivant », elle utilise les caractéristiques physiques ou comportementales pour identifier ou vérifier l'identité d'un individu. Le mot français biométrie définit « l'étude mathématique des variations biologiques à l'intérieur d'un groupe déterminé ». Par conséquent, la biométrie recouvre l'ensemble des procédés tendant à identifier un individu à partir de la « mesure » de l'une ou de plusieurs de ses caractéristiques physiques, physiologiques ou comportementales.

Les chercheurs affirment que l'ADN, l'empreinte digitale, la forme de l'oreille et même les contorsions faciales sont des identificateurs uniques utilisés comme des formes les plus connues de la technologie biométrique.

1.3 Bref historique sur Biométrie

La première forme de biométrie a vu le jour dans le début des années 1870. Le fondateur de la police scientifique française Alphonse Bertillon, a mis au point une méthode d'identification des criminels connue sous le nom Bertillonage. Le Bertillonage est une forme d'anthropométrie, un système par lequel des mesures du corps liées à l'anatomie humaine (petit, moyen, grand) sont prises à des fins de classification et de comparaison. Il s'agissait d'enregistrer les formes du corps avec une marque des particularités remarquées à la surface du corps telles que les tatouages, cicatrices, grains de beauté, les couleurs des yeux et les marques de naissance. Néanmoins ce système d'identification criminelle utilisé à cette époque présente des failles car les mesures prises n'étaient pas considérées comme uniques et précises et les caractéristiques sur lesquelles Bertillon a fondé son système d'identification n'étaient pas propres à un seul individu.

En raison des efforts consentis à la collecte minutieuse des mesures fiables, efficaces et précises, le bertillonage a été rapidement remplacé lorsque l'empreinte digitale est apparue sur les lieux comme un moyen d'identification plus efficace.

Les empreintes digitales remontent au 14ème siècle, en Chine. Bien que l'utilisation soit probablement une signature et que les capacités uniques de l'empreinte digitale ne soient pas entièrement connues. Sir Henry Faulds (1843-1930) a remarqué que l'empreinte digitale est une forme d'identification criminelle. Galton (1822-1911) a constaté qu'il n'y avait pas deux empreintes digitales identiques, même pas sur un ensemble de jumeaux identiques, et que l'empreinte digitale resterait fiable et inchangée et pourrait être utilisée pour l'identification tout au long de la vie d'un individu.

En 1897, Edward Henry a développé et mis en service un système de classification pour l'identification des empreintes digitales basés sur des caractéristiques physiologiques. Ce système attribue à chaque doigt une valeur numérique et divise les enregistrements d'empreintes digitales en groupes en fonction des types de motifs. Aussi le système permet de rechercher un grand nombre d'enregistrements digitales en classant les empreintes selon les 3 catégories connues : « arche », « tourbillon » et « boucle ». A partir de l'année 1902, le système d'empreintes digitales développé par Edward Henry est devenu le système le plus utilisé dans les pays anglophones, dans différents domaines (militaire, marine, sécurité).

En 1960, Gunnar Fant a crié le premier modèle de production vocale acoustique décrivant les composantes physiologiques de la production de la parole acoustique pour mieux comprendre les composants biologiques de la parole, un concept important pour la reconnaissance du locuteur.

Le système d'identification automatique des empreintes digitales (AFIS) est une méthode d'identification biométrique qui utilise la technologie d'imagerie numérique pour obtenir, stocker et analyser les données d'empreintes digitales. Depuis l'année 1969 le Bureau Fédéral des Investigations (FBI) aux états unis utilise le système AFIS pour élucider les affaires criminelles et lutter contre la fraude.

Avec le développement des ordinateurs et de la technologie numérique dans les années 60, on a commencé à automatiser l'identification des individus par les différentes techniques telles que la reconnaissance faciale en 1970, l'iris de l'œil en 1980, etc.

1.4 Marché de la Biométrie

La biométrie est toujours en pleine croissance depuis des dizaines d'années, elle est devenue en quelques années une solution grand public et se développe à grande vitesse. L'intégration de capteurs d'empreintes digitales, d'iris et de reconnaissance faciale dans les Smartphones l'a rendue accessible au plus grand nombre de sociétés financières en France, telles que : Société Générale, Banque Postale, BPCE, Crédit Nord, etc.

Dans son rapport intitulé « Sensors for Biometry and recognition 2016 », l'institut d'études Yole Développement estime que les technologies d'empreintes digitales dominantes évolueront progressivement vers des solutions multimodales. La conclusion la plus importante souligne que le secteur des applications smartphone constitue le moteur majeur du développement de la biométrie à près de 66% du marché total de la biométrie. La biométrie pour le consommateur bénéficiera sans doute d'une croissance de l'ordre de 10% de 2016 à 2021, selon les analystes de Yole.

Selon Yole, 525 millions d'unités de capteurs auraient été vendues en 2015 et ce chiffre devrait atteindre 1500 millions d'unités d'ici à 2021. A côté de la détection des empreintes digitales pour le déverrouillage et le paiement mobile, il faut compter avec les technologies de reconnaissance visuelle pour la sécurité basées sur les images combinées de l'œil et du visage. En outre, les assistants vocaux développés par Amazon et Google mettent en jeu des modules de reconnaissance vocale enregistrée.

Selon un rapport de TechSchi Research, le marché mondial de l'identification biométrique automobile devrait atteindre 303 millions de dollars américains d'ici 2024.

Le rapport devise le marché en cinq types de saisie de données biométriques (empreintes digitales, faciales, iris, voix et multimodal) et valorise la marché à 2019 à 138 millions de dollars américains.

1.5 Parts de marché par technologie

Les empreintes digitales continuent à être la principale technologie biométrique en termes de part de marché, près de 50% du chiffre d'affaires total (hors applications judiciaires). La reconnaissance du visage, avec 12% du marché (hors applications judiciaires), dépasse la reconnaissance de la main, qui avait avant la deuxième place en termes de source de revenus après les empreintes digitales.

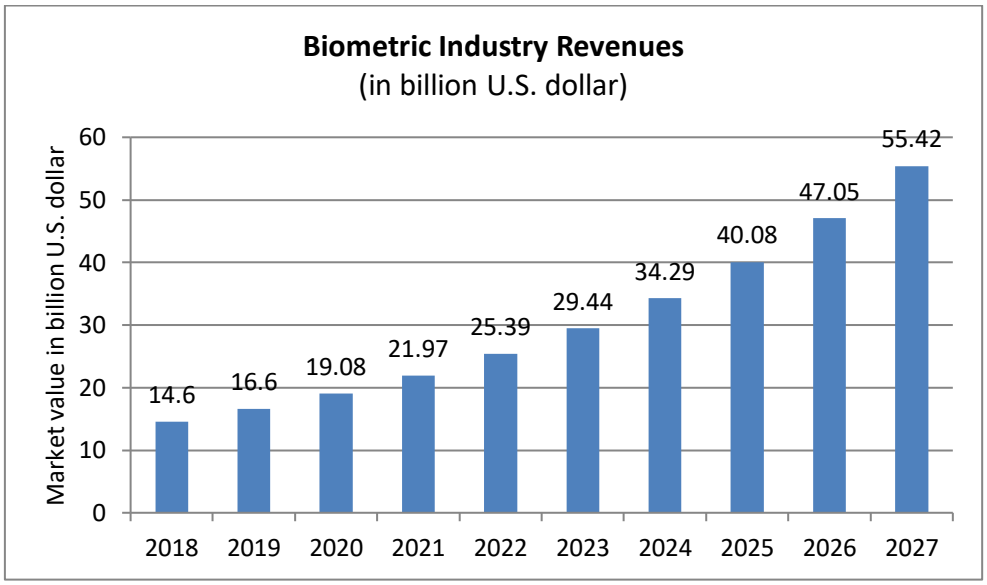


Figure 1.1: Evolution de l'industrie de la biométrie dans le marché international [2]

Selon un rapport de TechSchi Research, le marché mondial de l'identification biométrique automobile devrait atteindre 303 millions de dollars américains d'ici 2024. Les performances des futurs capteurs disposant d'un SNR (rapport signal bruit) élevé et de mémoires de comportement vont modifier le marché. Des applications émergentes potentielles émergent dans le bâtiment ou le secteur automobile, assurant la détection et l'identification des personnes grâce à des capteurs d'empreintes digitales flexibles disposés sur la poignée de porte ou sur le volant.

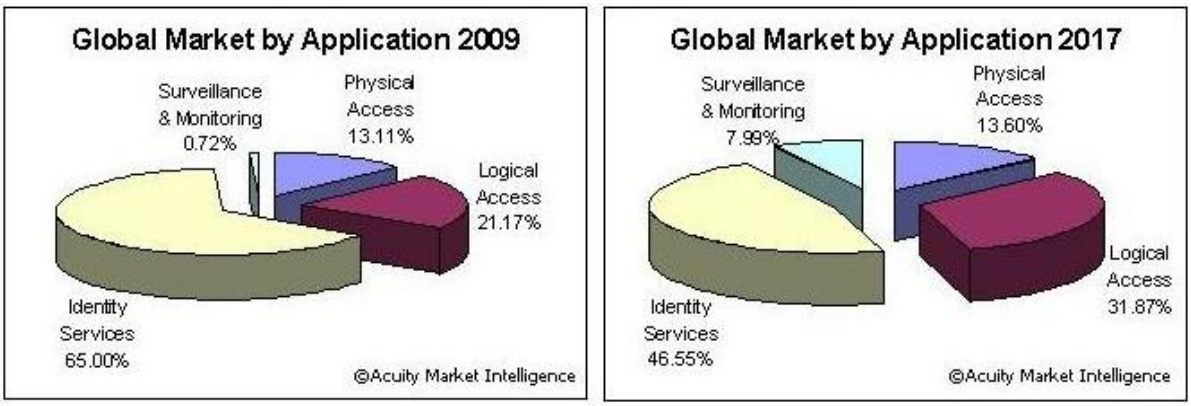


Figure 1.2 : Part de marché biométrique par application [2]

1.6 Caractéristiques de la Biométrie

La biométrie est basée sur l'analyse de données liées à l'individu et peut être classée en trois grandes catégories :

- Analyses biologiques (ADN, sang, urine, salive, etc.) ;
- Analyses morphologiques (empreinte digitale, voix, iris de l'œil, forme de la main, trait de visage, dessin de réseau veineux, etc.) ;
- Analyses comportementales (voix, frappe sur un clavier d'ordinateur, dynamique de la signature).

1.6.1 Biométrie morphologique

Nous pouvons citer :

- **Visage** : Le visage est certainement la caractéristique biométrique la plus utilisée pour s'identifier entre les êtres humains, ce qui explique pourquoi elle est en général très bien acceptée par les utilisateurs. L'identification à partir du visage se base sur les caractéristiques jugées significatives comme l'écart entre les yeux, la forme de la bouche, le tour du visage, la position des oreilles. Les équipements utilisés généralement pour les systèmes d'identification du visage sont : un appareil photo, une caméra de vidéosurveillance et un ordinateur.

  Dans un environnement contrôlé, des paramètres tel que : l'angle de prise de photo, l'intensité de ressource lumineuse, arrière plan, la distance de la caméra au sujet sont des paramètres maîtrisés par le système. Dans un environnement non contrôlé, il faut tout d'abord détecter la présence ou l'absence de visage dans l'image, puis on doit effectuer des traitements. Enfin, si nous travaillons sur un flux vidéo, le système doit suivre le visage d'une image à l'autre.

  - ✓ avantages :
    - absence de contact avec le capteur ;
    - technique non coûteuse ;
    - technique non encombrante.
  - ✓ inconvénients :
    - taille du lecteur biométrique est très volumineuse ;
    - les vrais jumeaux ne sont pas différenciés ;
    - technique moins fiable en cas de : déguisement, port de lunette, chapeaux, etc ;
    - sensible au changement de la source d'éclairage.

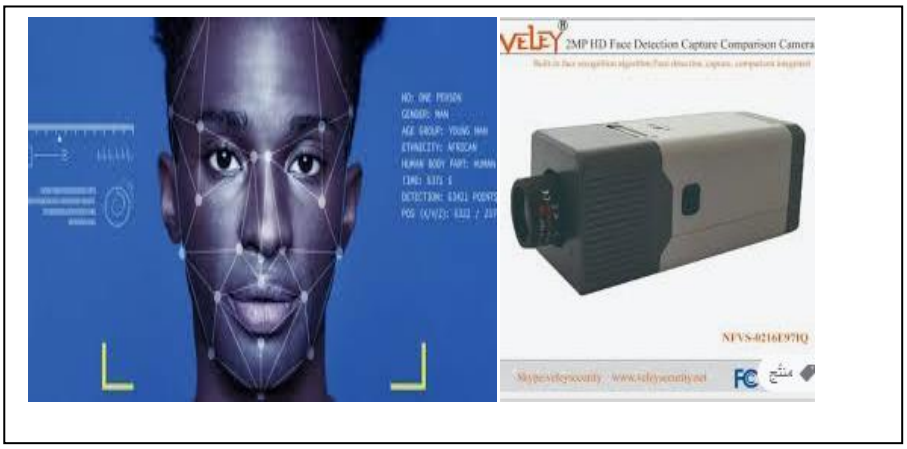


Figure 1.3 : Capture et détection du visage [3]

- **Empreinte digitales** : les systèmes biométriques utilisant l'empreinte digitale sont les plus utilisés et les plus anciens. Cependant les empreintes digitales sont une technique biométrique mal acceptée par les utilisateurs. Les lecteurs d'empreintes digitales scannent puis relèvent des éléments permettant de différencier les empreintes digitales qui sont formées par les crêtes (ridge) et les vallées (furrow) présentes sur la surface du bout des doigts.

  Nous constatons de plus en plus l'implémentation des lecteurs empreintes digitales sur des ordinateurs ou des Smartphones pour sécuriser leurs utilisations et cela devient un peu acceptable par le public.

  - ✓ avantages**:**
    - taille du lecteur biométrique non volumineuse ;
    - technique non coûteuse ;
    - fiable.
  - ✓ inconvénients **:**
    - indispensabilité de la coopération de l'individu ;
    - acceptante d'un moulage de doigt ou un doigt coupé.

Figure 1.4 : Exemple d'un appareil d'acquisition des empreintes digitales

- **Géométrie de la main :** la reconnaissance s'effectue à partir de la géométrie de la main dans l'espace 3D : longueur de doigts, largeur et épaisseur de la paume, dessins des lignes de la main. Les systèmes de reconnaissance de la géométrie de la main sont simples d'usage. L'utilisateur doit poser la paume de sa main sur une plaque qui possède des guides afin de l'aider à positionner ses doigts.

  Ces appareils peuvent être difficiles à utiliser pour certaines catégories de population pour lesquelles étendre la main est un problème, telles que les personnes âgées ou celles qui ont de l'arthrite. Cette technique est bien adaptée pour les systèmes à moyenne sécurité telle que le contrôle d'accès physique ou logique.

  - ✓ avantages **:**
    - faible volume de stockage ;
    - bonne acceptante.
  - ✓ inconvénients **:**
    - système encombrant ;
    - technique coûteuse.

Figure 1.5 : Exemple de capture de la géométrie de la main

- **Rétine :** les caractéristiques de la rétine sont liées à la configuration géométrique des vaisseaux sanguins. Par conséquent, les motifs formés par les veines sous la surface de la rétine sont uniques et stables dans le temps. Ils ne peuvent être affectés que par certaines maladies. Pour ces raisons, la reconnaissance de la rétine est actuellement considérée comme une des méthodes biométriques les plus sûres. Cette technologie utilise un matériel spécialisé et un rayon qui illumine le fond de l'œil. Les systèmes identifient jusqu'à cent quatre vingt douze points de repères. Quelques risques pour la santé ont été révélés et limitent l'utilisation de cette technique à des locaux de haute sensibilité.

Figure 1.6 : Photo numérique de la rétine avec un appareil d'acquittions [4]

- ✓ avantages**:**
  - o système fiable ;
  - o cartographie reste la même tout au long de la vie ;
  - o unicité même chez les vrais jumeaux.
- ✓ inconvénients **:**
  - o technique coûteuse ;
  - o nécessité de placer le capteur à une distance des yeux ;
  - o difficile de contrôler d'une population importante (exp : aéroport).
  - o sensible au changement de la source d'éclairage.

- **Iris :** cette technique est récente puisqu'elle ne s'est véritablement développée que dans les années 80, grâce aux travaux de J. Daugman [5]. L'iris est un organe visible de l'extérieur qui est protégé par un modèle épi-génétique unique et reste stable tout au long de la vie d'un être humain. Les iris sont uniques et les deux iris d'un même individu sont différents. La reconnaissance de l'iris est considérée comme une des méthodes biométriques les plus fiables. La capture de l'iris se fait par une caméra standard. Du fait des contraintes sur l'éclairage de l'œil, le capteur doit être assez proche de celui-ci (un mètre maximum) ce qui restreint les applications d'une telle technologie. L'éclairage de l'œil doit être uniforme et il faut éviter les reflets.
  - ✓ avantages **:**
    - o système fiable ;
    - o l'iris ne varie presque pas au cours d'une vie ;
    - o les iris sont uniques et différents même pour les vrais jumeaux.
  - ✓ inconvénients **:**

- contraintes d'éclairage ;
- acceptabilité très faible.

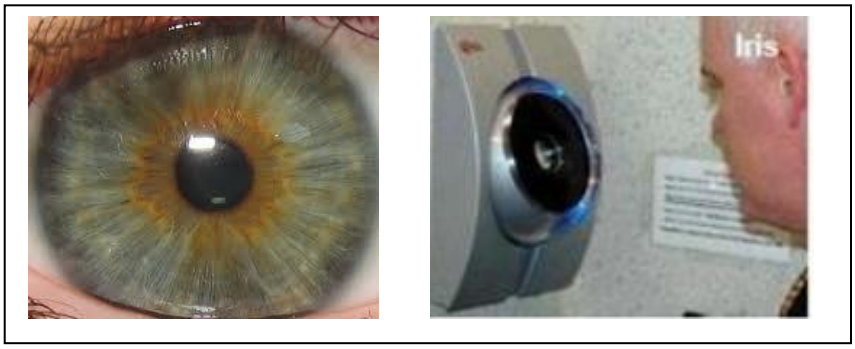


Figure 1.7 : Photo numérique de l'Iris avec un appareil d'acquittions [6]

1.6.2 Biométrie comportementale

Nous pouvons citer :

- **Signature manuscrite** : nous pouvons identifier une personne par une signature qui est propre à lui. Ce système fonctionne avec un capteur et un stylo. Le capteur est une tablette graphique. La reconnaissance de la signature est basée sur deux modes :
  - Mode statique : le mode statique utilise seulement la forme géométrique de la signature ;
  - Mode dynamique : le mode dynamique utilise les informations géométrique et dynamique telles que l'accélération, pression, vitesse de la signature, variation du rythme du stylo, calcul de la distance pendant laquelle la plume est suspendue entre deux lettres [7].

✓ avantages **:**

- très acceptable par les utilisateurs ;
- facile à utiliser ;
- action qui implique (responsabilité) le demandeur.

✓ inconvénients **:**

- dépendance de l'état de santé ou émotionnel ;
- forte possibilité de fraude ;
- faible stabilité à long terme.

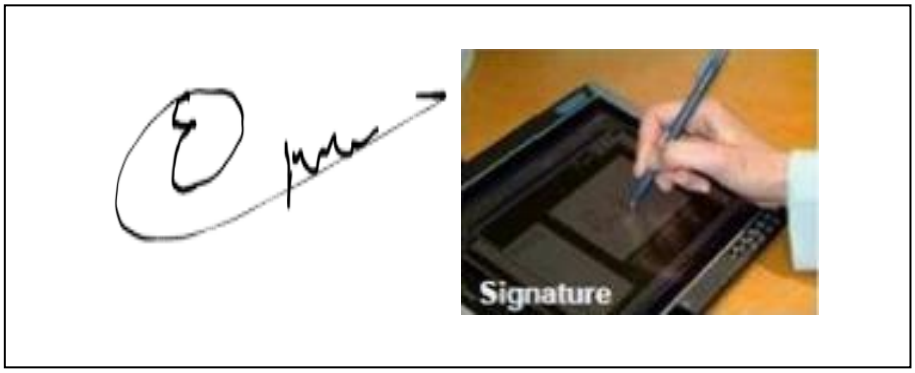


Figure 1.8: Exemple de capture d'une signature

- **Frappe au clavier** : c'est une technique de reconnaissance d'un individu basée sur la dynamique de frappe de touche sur un clavier en analysant la façon dont on tape du texte sur un clavier, en utilisant un dispositif matériel et logiciel qui calcule la vitesse de frappe, la suite des lettres, mesure des temps de frappe, pause entre chaque mot. L'avantage de ce système n'exige pas trop de matériel, il suffit de disposer d'un clavier et un logiciel de reconnaissance biométrique installé sur un serveur.
  - ✓ avantages **:**
    - o très acceptable par les utilisateurs ;
    - o moyen non intrusif qui exploite un geste naturel.
  - ✓ inconvénients **:**
    - o n'est pas permanente en fonction de l'âge, l'état de santé ou émotionnel.

Figure 1.9 : Exemple de représentation de la dynamique de frappe au clavier [8]

- **Voix** : la voix est le moyen de communication le plus naturel chez l'être humain, ses caractéristiques varient d'un individu à l'autre. Elle est classée à la frontière entre la caractéristique comportementale et caractéristique physique. La caractéristique de la voix déterminée par le conduit vocal et les cavités buccales et nasales. De nos jours, l'usage massive des téléphones portables et micro portables dotés de microphones a

énormément aidé le développement de cette technologie. Il existe différentes modalités de reconnaissance du locuteur :

- Indépendant du texte : le locuteur est libre de se prononcer ce qu'il veut, les corpus d'apprentissage et de test sont différents. Ce mode est utile lorsque l'on veut reconnaître un locuteur sans sa coopération. Le mode indépendant du texte est choisi chaque année par l'Institut National des Standards et Technologies (NIST), pour conduire les évaluation en reconnaissance du locuteur [7][9][10] ;
- dépendant du texte : le modèle du locuteur est entraîné avec un vocabulaire (mots et phrases) qui sera utilisé au moment du test ;
- Texte prompté : un texte est différent à chaque session et pour chaque personne, est imposé au locuteur et déterminé. Le corpus d'apprentissage et de test peuvent être différent.

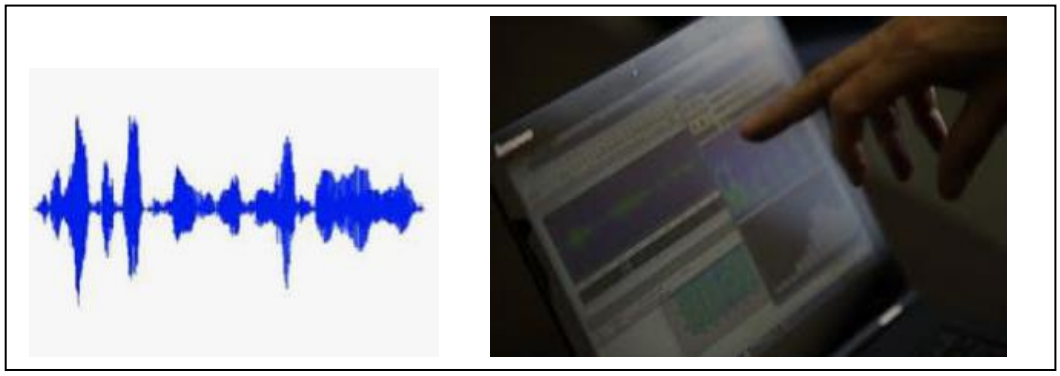


Figure 1.10 : Exemple de reconnaissance de la voix et son analyse

- ✓ avantages **:**
  - très acceptable par les utilisateurs ;
  - elle n'est pas intrusive ;
  - technologie biométrique facile à mettre en œuvre ;
  - elle permet de faire la reconnaissance de l'individu par téléphone.
- ✓ inconvénients **:**
  - sensibilité aux bruits lors de l'acquisition ;
  - caractéristiques comportementales changent avec le temps ;
  - faible niveau de différenciation entre deux voix.

1.6.3 Biométrie biologique

Nous pouvons citer :

- **ADN** : la molécule ADN, connue sous le nom d'acide désoxyribonucléique, se trouve dans les noyaux de toutes les cellules humaines, elle contient les instructions propres à la cellule et détermine comment les traits d'un individu seront transmis d'une génération à l'autre. L'ADN contient toutes les informations nécessaires au développement et au fonctionnement du corps. L'information génétique d'une personne est unique, car aucun membre de l'espèce ne possède la même combinaison de gènes codés dans l'ADN. C'est pour cela l'analyse des empreintes génétiques est une technique d'identification des individus très fiable, cette notion fut introduite par les services de police pour identifier les criminels, les cadavres déchiquetés et résoudre certaines affaires judiciaires.
  - ✓ avantages **:**
    - identifier les personnes avec une grande précision ;
    - permanent durant la vie de l'individu.
  - ✓ Inconvénients **:**
    - technique coûteuse ;
    - lente pour avoir les résultats.

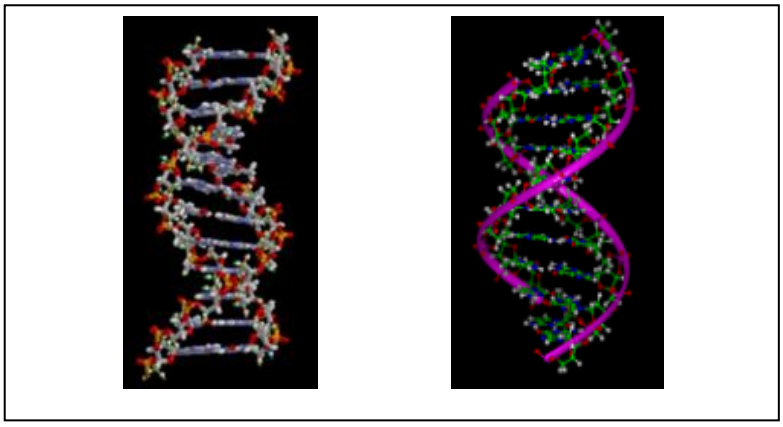


Figure 1.11 : Exemple de l'ADN dans la biométrie

- **Veines de la main** : les veines de la main sont des réseaux qui varient d'une personne à une autre. L'analyse de cette différence permet de maintenir des points pour différencier une personne à une autre. Des capteurs infrarouges peuvent être utilisés pour enregistrer l'image des veines et déterminer la structure de la veine de la main.

- ✓ avantages **:**
  - ne nécessite pas de contact ;
  - difficile à falsifier.
- ✓ inconvénients **:**
  - technique est très coûteuse.

Figure 1.12 : Capture des réseaux veineux de la main [11]

1.7 Comparaison entre les différents modalités biométriques

Le Tableau 1.1, donne une comparaison entre les différentes biométriques, que nous avons abordés dans la section 1.6, avec H = élevé, M = moyenne, L = faible.

Tableau 1.1 : Tableau de comparaison entre les différentes modalités biométriques

Biométrie	Performance	Coût	Acceptabilité	Stable au long terme	Universelle	Unique
Visage	M	M	H	M	H	L
Empreinte digitale	H	M	M	H	M	H
Géométrie de la main	M	H	M	M	M	M
Rétine	H	M	L	H	H	H
Iris	H	H	L	H	H	H
Signature	M	M	H	L	L	L
Frappe clavier	M	M	M	L	L	L
ADN	H	H	L	H	H	H
Voix	M	L	H	M	M	L
Veines de la main	M	L	M	M	M	M

1.8 Caractéristiques communes des systèmes biométriques

Généralement, les caractéristiques communes des systèmes biométriques :

1.8.1 Unicité

Pour identifier une personne au sein d'une population donnée, il est nécessaire que la donnée biométrique utilisée soit unique à cette personne. L'ADN, l'empreinte digitale, la rétine et l'iris sont considérées comme des caractéristiques uniques au sein de grandes populations. Elles doivent aussi permettre la différenciation d'un individu par rapport à un autre.

1.8.2 Universalité

Ces caractéristiques sont reconnues juridiquement et doivent être possédées par chaque individu.

1.8.3 Permanence

La caractéristique biométrique doit être invariante et stable dans le temps.

1.8.4 Enregistrement

Possibilité de mesurer les caractéristiques biométriques d'un individu à l'aide d'un capteur approprié.

1.8.5 Mesure

Un système biométrique est mesuré par certains paramètres qui représentent les performances dudit système.

1.9 Application de la biométrie

Les applications de la biométrie servent beaucoup plus le domaine de la sécurité, comme le contrôle d'accès physique et virtuel, authentification des transactions, criminalistique, nous pouvons citer :

1.9.1 Contrôle d'accès

Le contrôle d'accès peut être lui-même subdivisé en deux catégories :

1.9.1.1 Contrôle d'accès physique

- Salle machine
- Lieux sensibles (bâtiments, centrales nucléaires).

1.9.1.2 Contrôle d'accès virtuel

- Accès au réseau informatique ;
- Logiciels utilisant les mots de passes ;
- Accès aux sites Web ;
- Lancement des systèmes d'exploitation ;

1.9.2 Authentification des transactions

- Transferts de fonds ;

- Paiement à distance sur internet et par téléphone;
- Commerce électronique ;
- Retrait d'argent aux guichets des banques.

1.9.3 Administration

- Documents d'identité (passeport, carte d'identité, permis de conduire) ;
- Services sociaux (sécurité sociale);
- Système de vote électronique ;
- Contrôle des frontières.

1.9.4 Equipement divers

- Véhicule anti-démarrage ;
- Coffre fort avec serrure électronique ;
- Contrôle des temps de présence ;
- Distributeur automatique des billets.

1.9.5 Criminalistique

- Identification de corps ;
- Recherche criminelle ;
- Recherche des véhicules volés.

1.10 Choix de la technique biométrique

La Figure 1.13 appelée analyse de Zephyr présente une évaluation des avantages et inconvénients applicatifs de différentes méthodes de reconnaissance biométriques basés sur des critères d'évaluation à savoir : le coût, la précision, l'ergonomie (effort) et le Caractère intrusif.

Les principales contraintes liées à la biométrie sont dues à l'ergonomie et à l'acceptabilité de certaines modalités. Si la reconnaissance d'iris ou d'empreintes digitales sont généralement mal acceptés par le public, il existe d'autres modalités, moins intrusives, comme la Reconnaissance Automatique du Locuteur (RAL) et les biométries du visage. D'une façon générale, les propriétés souhaitées pour tout système biométrique correspondent aux points suivants.

1.10.1 Robustesse

La caractéristique biométrique doit être la plus stable possible au cours du temps et la plus difficilement altérable par le contexte d'utilisation ;

1.10.2 Distinctibilité

La caractéristique biométrique doit être la plus fortement dépendante de l'utilisateur.

1.10.3 Accessibilité

Elle doit être facilement et efficacement mesurable par un capteur.

1.10.4 Acceptabilité

Elle ne doit pas être perçue comme intrusive par l'utilisateur. Cette propriété relativement subjective dépend du contexte culturel dans lequel le système d'identification biométrique est mis en œuvre.

1.10.5 Disponibilité

Pour chaque utilisateur, une quantité suffisante de mesure de la caractéristique biométrique doit être simplement disponible.

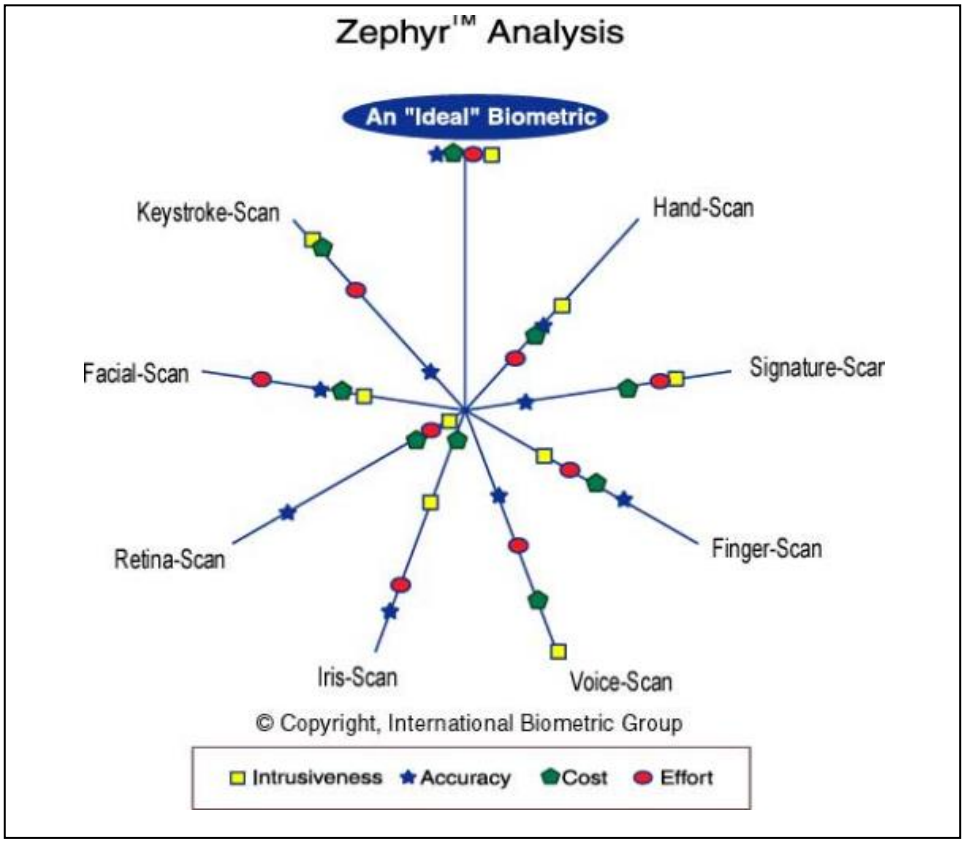


Figure 1.13 : Avantages et inconvénients applicatifs de différentes Méthodes de Reconnaissance Biométriques [2]

1.11 Architecture générale d'un système biométrique

Tout système biométrique (Figure 1.14), fonctionne soit en mode d'identification ou en mode vérification, il comporte deux processus qui se chargent de réaliser les opérations d'enregistrements et de tests :

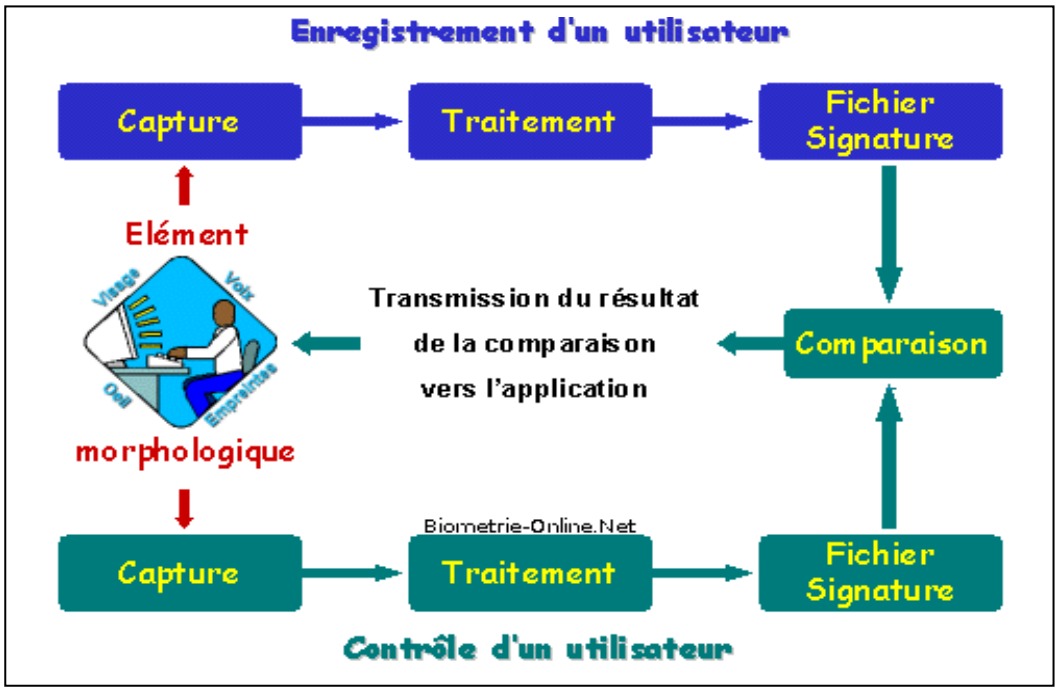


Figure 1.14 : Architecture d'un système biométrique

- le processus d'enregistrement : Ce processus a pour but d'enregistrer les caractéristiques des utilisateurs dans la base de données.
- le processus de tests (identification/vérification) : Ce processus réalise l'identification ou la vérification de l'individu.

Dans chacun des deux processus précédents le système biométrique exécute quatre opérations.

1.11.1 Module d'acquisition

Pour acquérir les données biométriques de l'individu, on utilise un capteur qui peut être un appareil photo, un lecteur d'empreintes digitales, une caméra de vidéosurveillance, etc.

1.11.2 Module d'extraction de caractéristiques

Après l'acquisition, on fait le traitement des données acquises pour extraire les caractéristiques pertinentes (Par exemple : extraire les formants, l'énergie et le pitch dans le cas de l'identification vocale) dont le système de reconnaissance a besoin et les présenter sous formes de vecteur caractéristique appelé modèle.

1.11.3 Module de classification

En examinant les modèles stockés dans la base de données (vecteurs), les caractéristiques biométriques extraites sont comparées avec ces vecteurs et en marquant de degré de similitude (scores de correspondances).

1.11.4 Module de test

Vérifie l'identité affirmée par un utilisateur ou détermine l'identité d'une personne basée sur le degré de similitude entre les caractéristiques extraites et le(s) modèle(s) stocké(s).

1.12 Multimodalité

Le but de la biométrie multimodale consiste à combiner plusieurs systèmes biométriques tels que l'iris, l'empreinte digitale et le visage dans un seul système, elle permet de réduire certaines limitations des systèmes basés sur une seule modalité (technique) tout en améliorant leur performances à savoir : la fiabilité, sécurité et robustesse.

Il existe des méthodes biométriques très fiables, telles que la reconnaissance de la rétine ou de l'iris, elles sont coûteuses et, en général, mal acceptées par le grand public et ne peuvent donc être réservées qu'à des applications de très haute sécurité. Pour les autres applications, des techniques telles que la reconnaissance du visage ou de la voix sont très bien acceptées par les utilisateurs mais ont des performances encore trop peu satisfaisantes pour être déployées dans des conditions réelles [12].

Dans le même contexte, les systèmes biométriques sont souvent affectés par les problèmes suivants :

- **Bruit introduit par le capteur :** le bruit peut être présent dans les données biométriques acquises, ceci étant principalement dû à un capteur défaillant ou mal entretenu. Par exemple, l'accumulation de poussière sur un capteur d'empreintes digitales, un mauvais focus de caméra entraînant du flou dans des images de visage ou d'iris, etc. le taux de reconnaissance d'un système biométrique est très sensible à la qualité de l'échantillon biométrique et des données bruitées peuvent compromettre sérieusement la précision du système [13][14] ;
- **Non universalité :** si chaque individu d'une population ciblée est capable de présenter une modalité biométrique pour un système donné, alors cette modalité est dite universelle. Ce principe d'universalité constitue une des conditions nécessaires de base pour un module de reconnaissance biométrique. Cependant, toutes les modalités biométriques ne sont pas vraiment universelles. L'Institut National des Standards et Technologies (NIST) a rapporté qu'il n'était pas possible d'obtenir une bonne qualité d'empreinte digitale pour environ 2% de la population (personnes avec des handicaps liés à la main, individus effectuant de nombreux travaux manuels répétés, etc.) [15] ;
- **Manque d'individualité :** les caractéristiques extraites à partir de données biométriques d'individus différents peuvent être relativement similaires. Par exemple,

une certaine partie de la population peut avoir une apparence faciale pratiquement identique due à des facteurs génétiques (père et fils, vrais jumeaux, etc.) ;

- **Manque de représentation invariante :** les données biométriques acquises à partir d'un utilisateur lors de la phase de reconnaissance ne sont pas identiques aux données qui ont été utilisées pour générer le modèle de ce même utilisateur lors de la phase d'enrôlement. Ceci est connu sous le nom de « variation intra-classe ». ces variations peuvent être dues à une mauvaise interaction de l'utilisateur avec le capteur (par exemple, changements de pose et d'expression faciale lorsque l'utilisateur se tient devant une caméra), à l'utilisation de capteurs différents lors de l'enrôlement et de la vérification, à des changements de conditions de l'environnement ambiant (par exemple changements en éclairement pour un système de reconnaissance faciale) ;
- **Sensibilité aux attaques :** bien qu'il semble très difficile de voler les modalités biométriques d'une personne, il est toujours possible de contourner un système biométrique en utilisant des modalités biométriques usurpées. Les modalités biométriques comportementales telles que la signature et la voix sont plus sensibles à ce genre d'attaque que les modalités biométriques physiologiques.

Ainsi, à cause de tous ces problèmes pratiques, les taux d'erreur associés à des systèmes biométriques uni-modaux sont relativement élevés, ce qui les rend inacceptables pour un déploiement d'applications critiques de sécurité. Pour pallier ces inconvénients, une solution est de l'utilisation de plusieurs modalités biométriques au sein d'un même système, on parle alors de système biométrique multimodal.

Les différentes formes de multi-modalités sont les suivantes :

- **Systèmes multiples biométriques :** à titre exemple, combiner reconnaissance du visage, reconnaissance des empreintes digitales et reconnaissance du locuteur ;
- **Systèmes multiples d'acquisition :** par exemple utiliser deux scanners différents (l'un optique, l'autre thermique) pour la reconnaissance d'empreinte digitales ;
- **Mesures multiples d'une même unité biométrique :** par exemple faire la reconnaissance des deux iris ou des dix doigts d'un même individu ;
- **Instances multiples d'une même mesure :** faire une capture répétée du même attribut biométrique avec le même système d'acquisition ;
- **Algorithmes multiples :** utiliser différents algorithmes de reconnaissance sur le même signal d'entrée.

1.13 Conclusion

Dans ce chapitre, nous avons introduit le concept des systèmes biométriques, leur architecture, leurs parts de marché par technologie avec ses différentes techniques ou modalités biométriques en comparant entre les différents types et de voir les avantages et les inconvénients de chaque modalité, aussi nous sommes intéressés au domaine d'application de la biométrie.

Aussi, nous avons abordé les problèmes qui peuvent affecter les systèmes biométriques et les différentes formes de multi-modalités.

.

CHAPITRE 2 : RECONNAISSANCE AUTOMATIQUE CRIMINALISTIQUE DU LOCUTEUR

2.1 Introduction

La science criminalistique (forensique) est un ensemble de différentes méthodes d'analyse fondées sur les sciences (mathématique, chimie, physique, etc.), afin de servir au travail d'investigation de manière large. Elle applique une démarche scientifique et des méthodes techniques dans l'étude des traces qui prennent leur origine dans une activité criminelle, ou litigieuse en matière civile, réglementaire ou administrative. Elle aide la justice à se déterminer sur les causes et les circonstances de cette activité [16].

L'identification d'une personne sur la base de sa voix est une motivation ancienne. Depuis de nombreuses années, les juges, les avocats, les enquêteurs de la police ou de la gendarmerie comme les agences de sécurité nationale souhaitent utiliser des procédés d'authentification vocale criminalistique pour mener une enquête ou confirmer un verdict de culpabilité ou d'innocence [17][18][19].

En dépit du fait que les bases scientifiques de l'authentification d'une personne par sa voix aient été mises en cause par les chercheurs du domaine (Exemple, par les scientifiques dès 1970 [20]), les phonéticiens Britanniques en 1983 [21] et la communauté francophone de la communication parlée de manière soutenue depuis 1990 [22], cette tâche est perçue comme aisée par le grand public. Kersta a introduit la notion de « l'empreinte vocale ». Cette appellation, empreinte vocale, tend à faire penser qu'une représentation graphique de la voix par un spectrogramme- est de même nature que les minuties des empreintes digitales ou des empreintes génétiques (ADN) et mène à une identification fiable du locuteur.

Ce chapitre est consacré principalement aux principes de la RAL et de la RACL. Il présente une ébauche sur la production de la parole, mécanisme de l'audition, les sources de variabilité du signal de parole et les différents niveaux d'information d'un signal de parole pour comprendre comment un suspect peut être identifié par sa voix.

2.2 Production de la parole

Le processus de production de la parole est un mécanisme très complexe dans lequel sont impliqués de nombreux organes et muscles. La source de la parole provient des poumons qui émettent un flux d'air. Ce flux d'air va traverser le larynx pour faire vibrer ou non les cordes vocales, ensuite il se propage à travers un ensemble de résonateurs (cavités vocales) qui permettent d'avoir les lieux et les modes d'articulations.

L'ensemble des organes agit comme un filtre, considéré comme linéaire, dont la réponse impulsionnelle comporte des fréquences de résonance caractérisées par de pics, appelés formants, dans le spectre du signal de sortie. Le signal résultant est globalement non

stationnaire mais peut être considéré comme stationnaire sur de très courtes périodes, de l'ordre de 20 ms (signal pseudo-stationnaire). Sur un segment de parole de cette longueur la voix est habituellement et schématiquement séparée en deux classes distinctes [23] :

- Voisée lorsqu'il y a vibration des cordes vocales, le signal est alors quasi-périodique ;
- Non voisée dans le cas d'un simple soufflement, le signal est alors considéré comme aléatoire.

Dans le premier cas, la source d'excitation est modélisée par un train d'impulsions périodiques, de fréquence dite de voisement F0, qui correspond à la fréquence de vibration des cordes vocales (la fréquence fondamentale); dans le second cas, la source est modélisée par un bruit blanc (Figure 2.1).

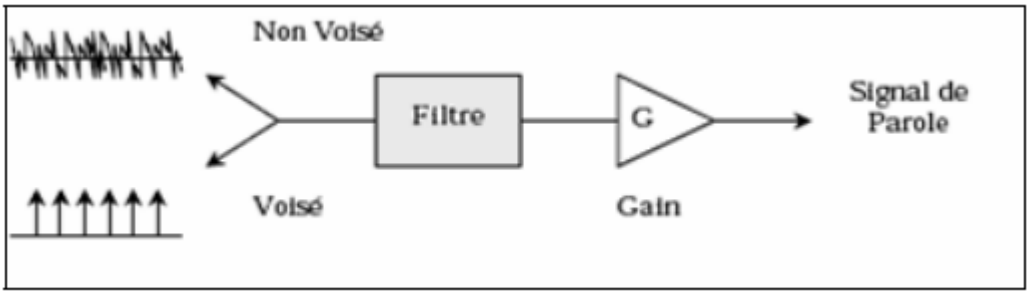


Figure 2.1 : Modèle de production de la parole

2.3 Mécanisme de l'audition

Les ondes sonores sont recueillies par l'appareil auditif, ce qui provoque les sensations auditives. Ces ondes de pression sont analysées dans l'oreille interne qui envoie au cerveau l'influx nerveux, le phénomène physique induit ainsi un phénomène psychique grâce à un mécanisme physiologique complexe [24].

L'appareil auditif comprend l'oreille externe, l'oreille moyenne et l'oreille interne (figure 2.2).

- l'oreille externe : est responsable de la transmission aérienne à travers le conduit auditif externe ;
- l'oreille moyenne : assure au moyen des trois osselets (marteau, enclume, étrier), la transmission mécanique du tympan jusqu'à la fenêtre ovale ;
- l'oreille interne : quant à elle, permet la transmission hydromécanique au niveau de la membrane basilaire ainsi que la transmission électro-chimique au niveau des cellules ciliées de l'organe de Corti.

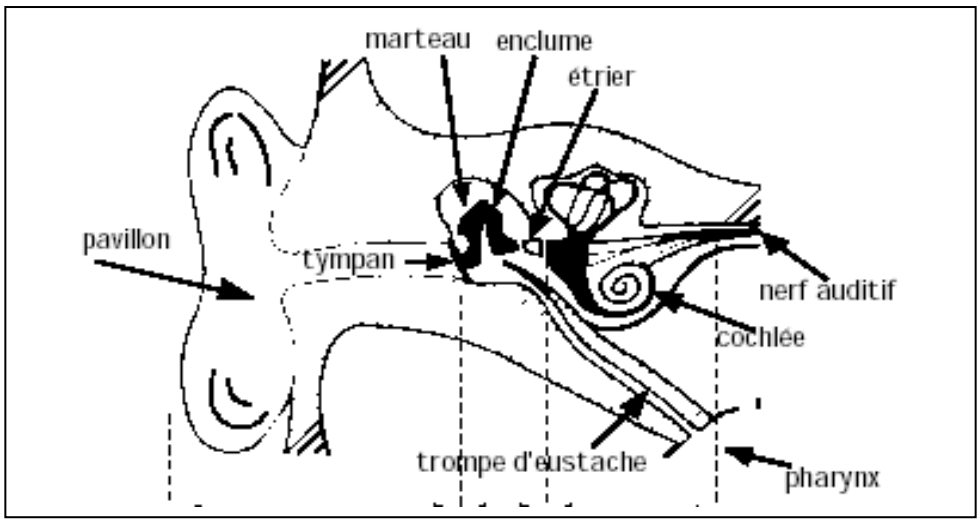


Figure 2.2 : Le système auditif humain [25]

La capacité de perception auditive de l'être humain est très limitée par rapport aux autres êtres vivants. Il lui est ainsi impossible de distinguer des sons de plus de 20 kHz, les ultrasons. De même lui est-il impossible de distinguer des sons d'une fréquence inférieure à 20 Hz, les infrasons. L'intervalle compris entre les infrasons et ultrasons s'appelle la zone d'audition. A l'intérieur de cet intervalle fréquentiel existe un sous espace délimité par les niveaux d'énergie des sons. Il existe une limite d'énergie en deçà de laquelle l'homme ne percevra pas un son d'une fréquence appartenant pourtant au spectre de l'audition. Cette limite d'énergie est appelée seuil d'audition et il est variable en fonction de la fréquence. Inversement, il existe une limite d'énergie maximale. Cette dernière ne doit pas être franchie car la cochlée, et plus particulièrement les cellules ciliées, peuvent être endommagées. Cette limite s'appelle le seuil de douleur et elle est aussi variable en fonction de la fréquence (Figure 2.3).

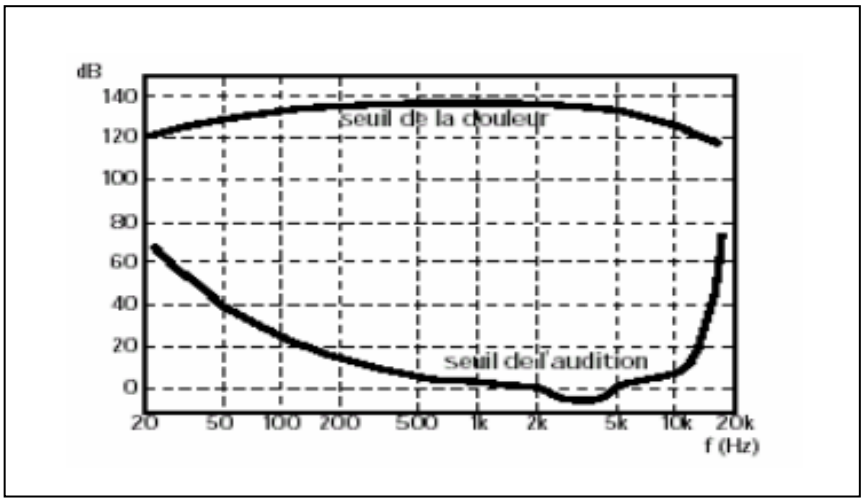


Figure 2.3 : Le champ auditif humain

2.4 Variabilités du signal de parole

Il existe deux types de variabilités.

2.4.1 Variabilité inter-locuteur

Le signal de parole permet la communication entre les individus. Il véhicule un message linguistique mais aussi des quantités d'informations extra linguistiques et des informations liées au locuteur. La personnalité, au sens large, du locuteur influence la production du signal de parole. Ceci permet notamment de discriminer les locuteurs.

2.4.2 Variabilité intra-locuteur

Il est impossible pour un même individu de reproduire exactement le même signal de parole. Les facteurs de variabilités pour un même individu sont multiples. Ils peuvent être liés à la nature physiologique de l'individu. Dans ce cas cette variabilité intra-locuteur est induite par l'évolution naturelle (volontaire ou non) de la voix d'une personne. L'état pathologique est un exemple de variation de la voix involontaire d'une personne. De plus, une altération de la voix due à l'âge est présente chez tous les individus. Cette variabilité est une difficulté majeure en RAL.

2.5 Facteurs extérieurs

La variabilité inter-session (entre sessions d'enregistrements) fait apparaître l'influence de facteurs extérieurs sur le signal de parole. A la sortie du conduit vocal humain, l'onde de parole est considérée comme idéale, car aucune déformation/distorsion de l'environnement extérieur ne l'a modifiée. L'environnement sonore lors de l'enregistrement, le matériel d'acquisition ou le canal de transmission utilisé vont ensuite déformer l'onde sonore originelle. Le canal de transmission, par exemple, agit comme un filtre en fréquence sur l'onde sonore. Ces facteurs rendent complexe la comparaison entre plusieurs échantillons d'un même individu. De nombreux travaux expérimentaux ont montré que des variations de matériel entre les phases d'apprentissage et de test sont à l'origine de graves dégradations des performances [26][27].

2.6 Différents niveaux d'information d'un signal de parole

Le signal de parole est un signal très complexe. Il véhicule non seulement un message linguistique mais aussi une quantité d'informations extra linguistiques liées au locuteur. On peut classer dans le signal de parole différents niveaux d'information. Les niveaux « bas » englobent des informations liées principalement à des traits physiques de la personne (facteurs morphologiques et physiologiques) et qui sont facilement utilisables à travers de l'analyse

numérique du signal de parole. Les informations de « haut niveaux » sont complexes à caractériser. Elles représentent des traits acquis (facteurs socio-culturels).

Il y a six niveaux d'information différents à savoir.

2.6.1 Acoustique

Les paramètres acoustiques sont relatifs à l'analyse de l'enveloppe spectrale du signal de parole et aussi aux caractéristiques physiques de l'appareil vocal. L'enveloppe du spectre caractérise principalement la morphologie du conduit vocal.

2.6.2 Prosodique

Le niveau prosodique désigne les caractéristiques d'un énoncé de parole qui sont la représentation formelle des éléments de l'expression orale tels que les accents, les tons, l'intonation employés par un locuteur et les rythmes d'élocution aux pauses et à la durée des phonèmes. Ces variations étant perçues par l'auditeur comme des changements de hauteur (mélodie).

2.6.3 Phonétique

Le niveau phonétique désigne la distinction des différents sons identifiables d'une langue. Chaque personne a une façon de prononcer ces propres phonèmes. Il est donc possible de caractériser certains phonèmes d'un locuteur afin de les distinguer des phonèmes des autres locuteurs.

2.6.4 Idiolectal

Les caractéristiques idiolectales se rapportent aux particularités langagières propres à un individu. Il s'agit en particulier des habitudes liées à l'utilisation des mots.

2.6.5 Dialogal

Le niveau dialogal définit la façon de communiquer d'un individu. Aussi, des indices de fréquence et de durée des prises de parole d'un locuteur dans une conversion peuvent servir à le caractériser.

2.6.6 Sémantique

Le niveau sémantique désigne la signification de l'énoncé de la parole, en d'autre terme, ce que l'on veut transmettre par cet énoncé.

2.7 Reconnaissance Automatique du Locuteur (RAL)

La Reconnaissance Automatique du Locuteur (RAL) s'inscrit dans le domaine du traitement de la parole [28] dont, la Figure 2.4 présente les différentes tâches du traitement de la parole. La RAL consiste à reconnaître l'identité d'une personne par l'analyse de sa voix [29][30]. Il exploite dans ce mode opératoire, les variabilités inter-locuteurs et intra-locuteurs beaucoup

plus que les informations extra-linguistiques du signal de parole. Cependant la RAL présente un certain nombre d'avantages qui le distingue des autres techniques biométriques notamment en matière de facilité de déploiement, le coût et l'acceptation par les utilisateurs.

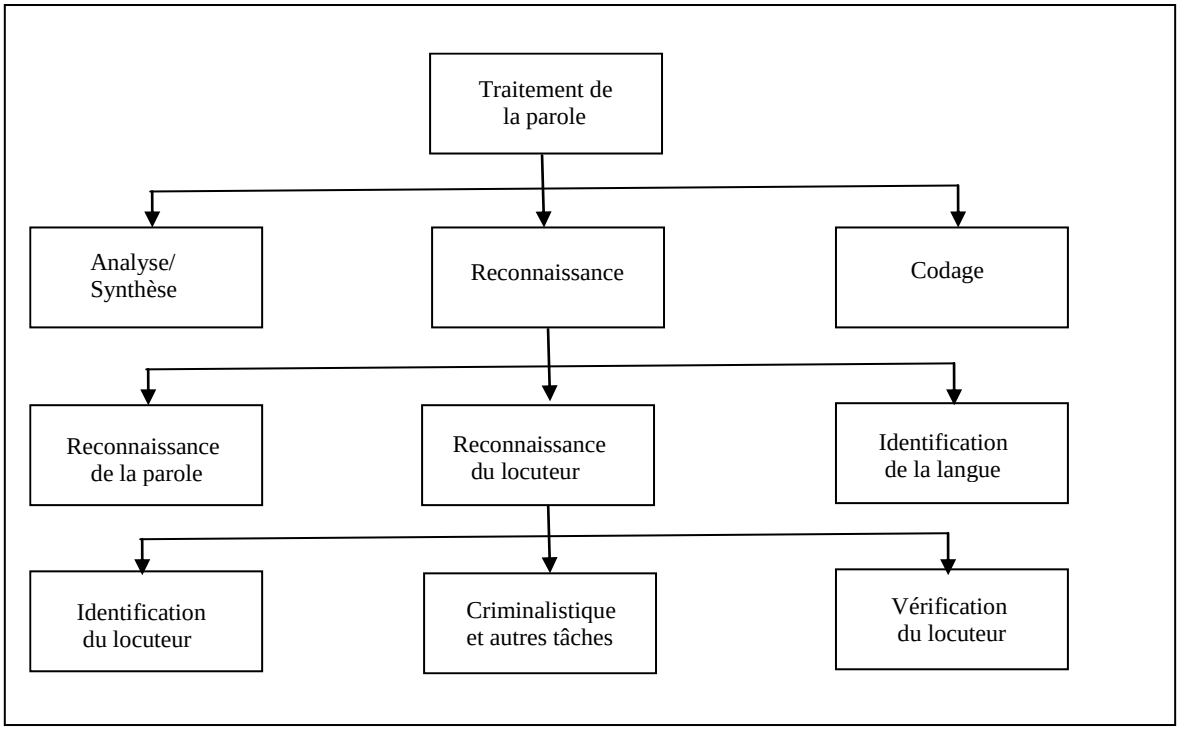


Figure 2.4 : Différentes tâches du traitement de la parole

La voix contient des informations caractérisant partiellement un locuteur. Le domaine scientifique utilisant ces informations pour vérifier l'identité d'une personne est appelé « RAL ». Les applications majeures concernent l'accès sécurisé à des locaux ou l'authentification à distance de l'utilisateur (notamment pour les services téléphoniques). Les techniques de reconnaissance du locuteur sont basées sur des mesures de ressemblance entre des enregistrements de parole. Ces mesures sont faites sur des paramètres acoustiques extraits par une analyse du signal. Elles peuvent prendre en compte les informations spécifiques au locuteur, le contenu du message vocal, les informations sur l'environnement et le matériel d'enregistrement.

Pour garantir un niveau de performance acceptable pour les applications de la reconnaissance du locuteur, plusieurs caractéristiques sont généralement nécessaires :

- les locuteurs n'essayent pas de déguiser leur voix ;
- les conditions d'enregistrement et de traitement du signal audio sont connus et/ou contrôlées ;

- des données de parole, enregistrées dans les mêmes conditions que le signal de test, sont disponibles pour référencier un locuteur dans le système ;
- la mesure de ressemblance est étalonnée au cours d'expériences réalisées dans les conditions contrôlées citées précédemment. La méthode de décision est estimée en fonction des résultats des expériences et en fonction de l'application visée.

2.7.1 Identification Automatique du Locuteur (IAL)

Il consiste à déterminer l'identité de l'individu parmi une base de N personnes connues, et ceci à partir d'un segment sonore inconnu, c'est-à-dire le système fait extraire les informations utiles de l'échantillon inconnu présenté à son entrée, puis il fait une comparaison et donne à la sortie l'identité du locuteur de la base de référence qui est le proche du segment de parole inconnu.

Il existe deux modes d'identification automatique, en milieu ouvert ou fermé :

- En milieu fermé : le locuteur appartient à la population connue. Le système RAL retourne l'identité du locuteur le plus probable parmi la population de la base ;
- En milieu ouvert : le locuteur peut ne pas être connu du système RAL. Dans ce cas le locuteur n'appartient pas forcement à cette population.

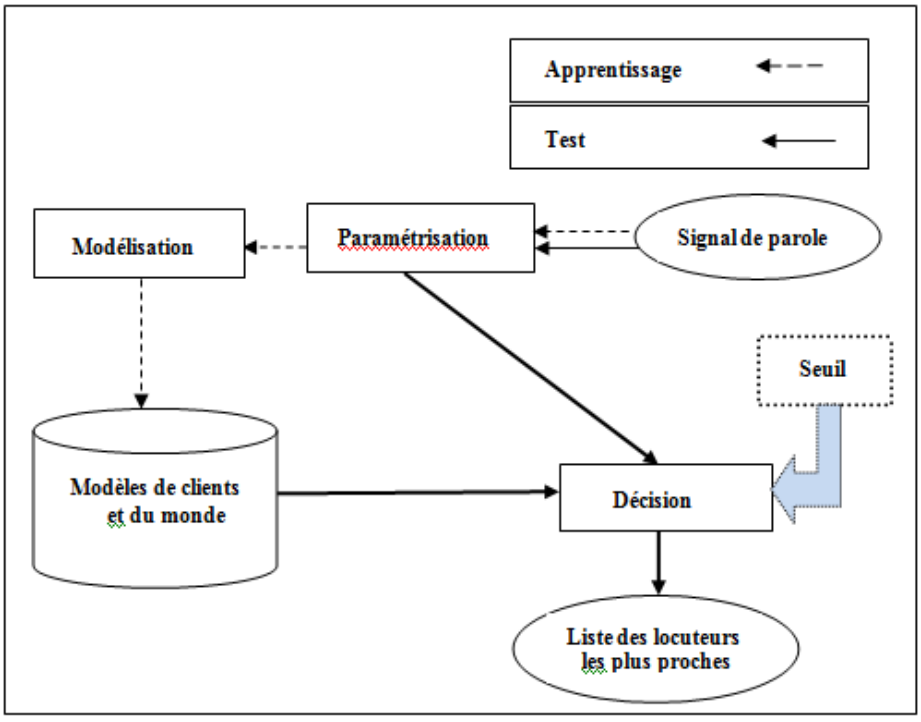


Figure 2.5 : Schéma d'un Système d'IAL

2.7.2 Vérification Automatique du Locuteur (VAL)

La VAL consiste à fournir à l'entrée du système, une identité proclamée et un segment sonore, le système calcule la ressemblance entre ledit segment et la référence du locuteur qu'il

prétend être, en fonction d'un seuil bien défini, à la sotie, on aura une réponse soit par acceptation (en cas de client) ou par rejet (en cas d'imposteur).

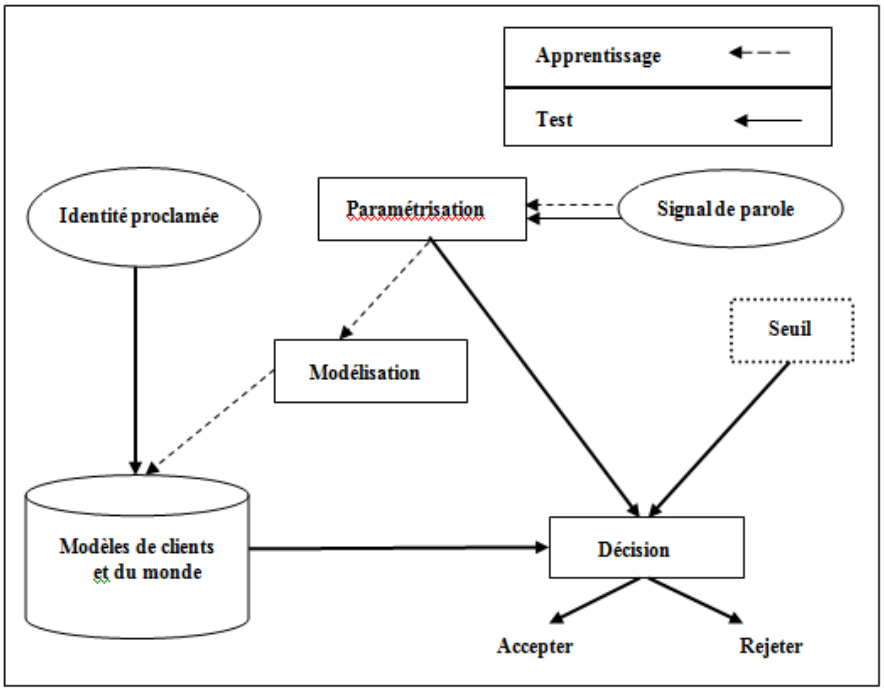


Figure 2.6 : Schéma d'un Système de VAL

2.7.3 Architecture d'un système de RAL

Un système de RAL est composé de trois étapes principales : la paramétrisation, la modélisation et la décision.

2.7.3.1 Paramétrisation

Cette étape permet de transformer un signal de parole en une suite de vecteurs. Le signal de parole, de par sa complexité (multitude d'informations et redondance) et par sa forme analogique est difficilement exploitable. La paramétrisation d'un signal de parole a pour but de proposer une représentation plus simple sous forme de vecteurs de paramètres acoustiques afin de faciliter l'extraction des informations désirées. Généralement dans cette phase qu'on précède à réduire les bruits et les redondances afin de ne conserver que les informations considérées comme utile à la tâche spécifiée.

Le signal de parole est varié lentement au cours temps, ce qui permet de le considérer comme signal quasi-stationnaire. Le calcul des paramètres acoustiques est ainsi réalisé en glissant avec une cadence régulière (ex : 10 ms) une fenêtre de pondération d'une longueur bien définie sur tout le signal. Généralement, la longueur de la fenêtre de pondération peut varier de 20 ms à 32 ms. On connaît plusieurs types de fenêtrage (Hamming, Hanning, etc). En général, le fenêtrage de Hamming est le plus utilisé en traitement du signal de parole.

Chaque fenêtre nous permet d'avoir une trame. Les trames obtenues sur tout le signal de parole sont traitées par la suite afin de produire les vecteurs acoustiques. Il existe trois grands types de paramètres :

- Paramètres issus de l'analyse spectrale : c'est la paramétrisation la plus utilisée en RAL. Elle représente les caracteristiques physiques de l'appareil phonatoire de chaque locuteur (exp : LPCC, MFCC) [31][32].
- Paramètres prosodiques : ces paramètres illustrent en général le style d'élocution, vitesse d'élocution (débit), durée, la fréquence fondamentale, l'énergie, etc [16].
- Paramètres dynamiques : le vecteur de paramètres issus des paramètrisations précédentes peut être complété par le vecteur correspondant aux dérivées du premier et second ordre de ces paramètres. Ces dérivées sont calculées à partir de plusieurs trames adjacentes. Elles permettent d'introduire une information concernant le contexte temporel d'une trame courante [33].

Les paramètres acoustiques doivent satisfaire les conditions qui influent sur les performances de tel système de RAL, ils doivent :

- être robustes aux supports de transmissions et aux bruits de l'environnement ;
- être faciles à extraire et doivent apparaître fréquemment dans le signal de parole ;
- avoir une faible intra-variabilité et une forte inter-variabilité ;
- être stables par rapport au temps.

2.7.3.2 Modélisation

Dans cette étape, on utilise les paramètres pertinents extraits dans la phase de paramétrisation pour construire un modèle pour chaque locuteur.

La phase de modélisation doit prendre en considération les contraintes suivantes :

- permettre une décision rapide lors de la phase de test ;
- la modélisation ne doit pas prendre un espace de stockage important ;
- l'algorithme d'estimation des modèles soit le moins complexe possible ;
- la modélisation permet une meilleure séparation entre les locuteurs ;
- les paramètres acoustiques issus de la phase de la paramétrisation doivent être représentés la plus complète possible ;

Il existe plusieurs méthodes de modélisation dans les systèmes RAL, à savoir :

- **Méthode vectorielle :** elle consiste à représenter un locuteur par un ensemble de vecteurs issus directement de la phase de paramétrisation. Cette approche comporte

deux techniques principales : l'alignement temporel dynamique et la quantification vectorielle ;

- **Quantification vectorielle (QV) :** s'agit de représenter l'espace acoustique par un nombre fini de vecteurs acoustiques formant ainsi un dictionnaire (codebook). Ce dernier est en général calculé de façon à ce que la distance moyenne entre un vecteur issu des données et son plus proche voisin dans le dictionnaire soit la plus petite possible. Dans les systèmes RAL. Ce dictionnaire est réalisé à partir des vecteurs spectraux provenant de l'analyse du signal de parole de locuteur.
- **Alignement Temporel Dynamique DTW (Dynamic Time Warping) :** cette méthode est souvent utilisée dans les systèmes RAL dépendant du texte, principalement pour la reconnaissance mono-locuteur à petit vocabulaire et en mots isolés [34]. Cette technique basée sur le calcul d'une distance entre deux vecteurs. Principalement, il fait la comparaison d'une séquence de vecteurs avec une autre séquence de vecteurs par le calcul de la distance accumulée entre ces deux séquences. Si les deux séquences sont identiques alors le chemin entre eux est diagonal, et par conséquent, la distance qui les sépare est minimale.

- **Méthode connexionniste :** Les Réseaux de Neurones Artificiels consistent en un ensemble d'outils et de méthodes de calcul, pouvant être appliqués dans divers domaines, tels que le traitement d'information, la classification de données, la statistique, le traitement de signal (image, parole), la prédiction de séries temporelles ou encore le contrôle, et dans de nombreuses applications industrielles. Dans cette méthode, un locuteur est représenté par un ou plusieurs réseaux de neurones appris directement par les vecteurs acoustiques obtenus lors de la phase de paramétrisation et permettant de le discriminer par rapport aux autres locuteurs [35][32]. Le réseau de neurones peut être constitué par une couche d'entrée, une couche de sortie et une ou plusieurs couches cachées. Malgré l'efficacité de discrimination des RN et leur performance de classement, ils restent incapables de résoudre leur principal problème qui est la durée d'apprentissage importante et nécessaire pour une grande population.

- **Méthode statistique :** cette catégorie correspond aux méthodes basées sur une représentation statistique du locuteur dans l'espace de paramètres acoustiques. Dans cette approche, les vecteurs acoustiques issus de la phase de la paramétrisation seront utilisés pour créer des modèles statistiques à long terme qui

tiennent compte de l'aspect temporel du signal de parole [32]. Les techniques statistiques les plus utilisées en RAL sont : les modèles de Markov cachés (HMM) en mode dépendant du texte et les modèles de mélange de gaussiennes (GMM) en mode indépendant du texte, et les Support Vector Machine (SVM) qui ont été conçus comme une fonction discriminante permettant de séparer au mieux des régions complexes dans des problèmes de classification à deux classes. Il peut y avoir un système où on combine deux méthodes (ex : le GMM/SVM) pour profiter des capacités génératives du GMM et discriminantes du SVM.

2.7.3.3 Décision

En ce qui concerne la vérification, la stratégie de décision nous permet de choisir entre les deux alternatives suivantes : l'identité de l'utilisateur correspond à l'identité proclamée ou recherchée ou elle ne correspond pas. Elle est basée sur un seuil prédéfini. L'estimation du seuil de la décision constitue la plus grande difficulté de ces techniques, et elle peut engendrer deux types d'erreurs, souvent prises comme mesures de performances pour ces techniques de vérification : Faux Rejet (FR) qui correspond à rejeter un vrai utilisateur ou une identité valable, et Fausse Acceptation (FA) qui donne l'accès à un imposteur.

2.7.4 Evaluation d'un système de RAL

En identification du locuteur, La performance du système est mesurée par : le Taux d'Identification Correcte (TIC) selon la formule suivante :

$$TCI = \frac{NTCI}{NTT} \tag{2.1}$$

Avec :
NTCI : Nombre de Tests Correctement Identifies.
NTT : Nombre Total de Tentatives.

En vérification du locuteur, le système est soumis à deux types de tests :

- Tests clients : lors desquels l'échantillon parole présenté au système correspond à l'identité proclamée ;
- Tests imposteurs : lors desquels l'échantillon parole présenté au système correspond à l'identité provient d'une personne inconnue du système.

La performance du système est mesurée par le Taux de Fausse Acceptation (TFA) et le Taux de Faux Rejets (TFR), donnés par :

$$TFA = \frac{NTIA}{NTTI} \tag{2.2}$$

Avec :
NTCA : Nombre de Tentatives des d'Imposteurs Acceptés ;
NTTI : Nombre Total de Tentatives des Imposteurs.

$$TFR = \frac{NTCA}{NTTC} \tag{2.3}$$

Avec :
NTCA : Nombre de Tentatives des Clients rejetés ;
NTTC : Nombre Total de Tentatives des Clients.

Ces deux types d'erreurs n'ont pas toujours la même incidence en termes de sécurité et de la qualité de service. La fausse acceptation peut être très pénalisante dans le cas d'une application requérant un niveau de sécurité élevé. Il n'est pas tolérable par exemple que n'importe qui puisse accéder à des informations personnelles, bancaires ou même de type secret défense. Le faux rejet peut également pénaliser des applications où l'utilisateur ne peut se permettre de perdre du temps en tenant de s'authentifier à plusieurs reprises. C'est le cas, par exemple, pour des services de secours d'urgence. Un utilisateur du système doit pouvoir être reconnu par le système dans les meilleurs délais.

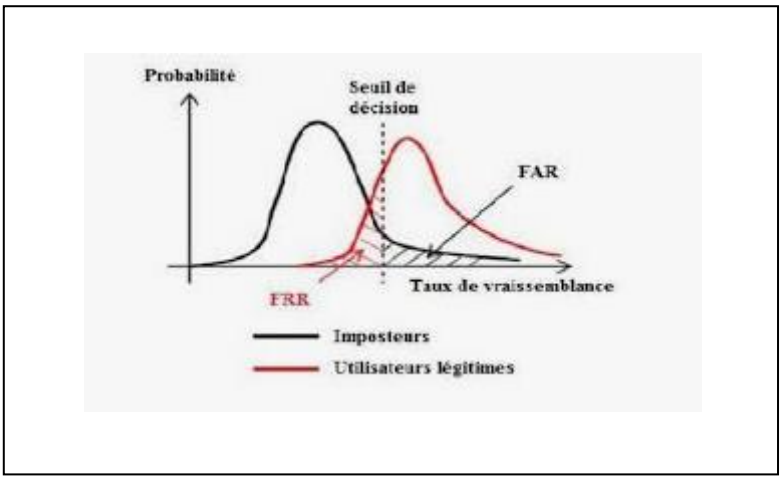


Figure 2.7: Courbe ROC (Receiver Operating Characteristic)

La courbe dite ROC (Receiver Operating Characteristic), permet de représenter graphiquement la performance d'un système de vérification pour les différentes valeurs de seuil de décision le Taux d'Erreur Egal EER (Equal Error Rate) correspond au point ou TFA=TFR (Figure 2.7).

Plus le seuil de décision est bas, plus le système acceptera les utilisations légitimes (clients) mais plus il acceptera aussi d'imposteurs. Inversement, plus le seuil de décision est élevé, plus le système rejettera d'imposteurs mais plus il rejettera aussi les utilisateurs légitimes, donc il faut choisir un compromis entre le TFA et le TFR.

2.8 Reconnaissance Automatique Criminalistique du Locuteur

Dans certaines affaires pénales, la voix enregistrée lors d'un appel téléphonique est le seul indice dont disposent les enquêteurs. Il existe donc une demande très pressante et pleinement justifiée de la part de la police judiciaire et des magistrats, d'utiliser ces enregistrements pour guider l'enquête, et pour établir la culpabilité d'un suspect ou prouver son innocence. Par conséquent, les techniques de reconnaissance du locuteur apportent une contribution précieuse au système de reconnaissance judiciaire du locuteur.

La Reconnaissance Automatique Criminalistique (Forensique) du Locuteur (RACL) est une tâche très complexe, dans les systèmes FASR, l'utilisation d'outils scientifiques est nécessaire pour répondre aux besoins d'un tribunal pour un crime ou un litige civil [36]. Les principaux domaines utilisés en sciences forensiques sont : la biologie, la chimie et la médecine [37]. Malgré la prédominance de ce dernier, il est mentionné qu'il existe d'autres disciplines utilisées telles que : la physique, l'informatique et la psychologie [37]. Par exemple, les techniques biométriques traditionnelles, telles que l'ADN et les empreintes digitales, sont souvent utilisés dans de nombreux cas criminalistiques. La nature des preuves, trouvées sur la scène du crime ou recueillis lors des opérations d'enquêtes, impose les méthodes ou disciplines scientifiques nécessaires à son étude. Dans le FASR, les experts s'intéressent aux méthodes d'identification d'une voix enregistrée. Ceci est basé sur le fait que chaque personne peut être identifiée à partir d'un échantillon de sa voix. En outre, un suspect peut laisser des enregistrements de sa voix sur le téléphone, la messagerie vocale, un répondeur ou dans un enregistreur caché, qui peuvent ensuite être utilisés comme preuve.

2.9 Techniques de la RACL

Des travaux mondialement reconnus, effectués avec la plus grande rigueur méthodologique, confirment que les techniques utilisées dans le RCL peuvent être résumées en quatre classes :

- Reconnaissance auditive ;
- Reconnaissance par spectrogrammes ;
- Reconnaissance automatique du locuteur.

2.9.1 Reconnaissance auditive

La physiologie de l'oreille humaine est adaptée à la perception de la voix humaine. Cette capacité naturelle est à la base de la reconnaissance auditive des locuteurs. Cette capacité à reconnaître les locuteurs est cependant variable suivant les individus et reste influencée par différents facteurs [19]. La familiarité entre l'auditeur et le locuteur concerné est un facteur reconnu, mais également la durée des exemples sonores, le contexte, l'intervalle temporel entre les exemples, les conditions de stress et de modifications volontaires de la voix et l'entraînement des auditeurs. De manière courante, une analyse auditive n'est pas réalisée isolement mais en addition à d'autre types de reconnaissance du locuteur.

En sciences forensiques, la reconnaissance de locuteurs par l'audition est pratiquée soit par des experts, phonéticiens ou spécialistes des sciences de la parole, sur la base de principes scientifiques, soit de manière perceptive par des profanes, principalement les victimes ou les témoins d'une infraction [36].

2.9.2 Reconnaissance par spectrogrammes

Le spectrogramme est un instrument permet de représenter les variations temporelles du spectre à court terme d'une onde de parole sous une forme graphique, appelée spectrogramme vocal ou sonagramme.

Sur le sonagramme, l'empreinte vocale est représentée sur un graphique de trois dimensions, dont le temps occupe la dimension horizontale, les fréquences la dimension verticales et la densité du trait indique l'intensité. Cette représentation permet la mise en évidence de plusieurs informations contenues dans le signal de parole comme la largeur de bande de la pente des formants des voyelles, leurs fréquences centrales, la durée des événements acoustiques, les formes caractéristiques des consonnes fricatives et l'énergie entre les formants [37]. La réalisation des stations informatiques de travail dotées d'une grande puissance de calcul offerte par les processeurs et la disponibilité de moniteurs vidéo de haute résolution permet aux experts forensiques d'élucider beaucoup d'affaires judiciaires fabuleuses au niveau des laboratoires de recherche criminelle.

En outre, le terme d'empreinte vocale est parfois employé pour faire référence à l'ensemble des méthodes relevant de la reconnaissance du locuteur, automatique ou manuelle, sans relation avec l'emploi ou non de spectrogrammes [19].

Les limites de la phonétique criminalistique incluent notamment la disponibilité limitée de phonéticiens qualifiés pour chacune des langues concernées.

2.9.3 Reconnaissance du Locuteur (RL)

En Reconnaissance du Locuteur (RL), seules les informations présentant une forte variabilité inter-locuteurs permettent de discriminer les différents individus. A l'inverse, les informations dont la variabilité intra-locuteur est élevée rendent la tâche de RL plus complexe.

Les informations les plus utilisées en RL, du fait de leur fort potentiel discriminant, sont des informations acoustiques obtenues périodiquement par une analyse fréquentielle ou temporelle [38].

D'autres paramètres tels que la prosodie ou la fréquence fondamentale, contiennent une information spécifique du locuteur [39].

D'autres informations présentes dans le signal de parole et citées précédemment peuvent s'avérer discriminantes dans le cadre de la reconnaissance du locuteur.

L'utilisation de la RL dans le domaine forensique touche les deux tâches (VAL et IAL), pour la VAL, on compare une trace prélevée d'une conversation téléphonique ou extrait d'un enregistreur de son avec le signal de parole du suspect. Par contre pour l'IAL, on cherche une personne parmi une population de suspects potentiels.

2.10 Interprétation bayésienne pour la RCL

Les travaux de recherche prouvent qu'un modèle probabiliste (le théorème de Bayes) est un outil adéquat pour aider les scientifiques à évaluer des preuves scientifiques. Il aide les juristes et les enquêteurs à interpréter ces preuves et à clarifier les rôles respectifs des scientifiques et des membres de la cour [40].

2.10.1 Calcul de la preuve

La preuve E est le résultat de l'analyse comparative des paramètres acoustiques (x) extraits de la trace (X), avec les paramètres acoustiques (y) extraits des énoncés du locuteur suspect (Y) [40].

2.10.2 Définition du rapport de vraisemblance (LR)

Le théorème de Bayes montre comment combiner de nouvelles données avec des connaissances préalables pour donner des probabilités postérieures à des problèmes juridiques. Cela permet à l'expert forensique d'évaluer la preuve en considérant deux hypothèses concurrentes :

H_0 : le locuteur suspect est la source de la voix incriminée (Trace) ;

H_1 : le locuteur à l'origine de la Trace n'est pas le locuteur suspect.

Le théorème de Bayes est représentée par la relation (2.4) :

$$\frac{p(H_0|E)}{p(H_1|E)} = \frac{p(E|H_0)}{p(E|H_1)} \times \frac{p(H_0)}{p(H_1)} \quad (2.4)$$

Avec :

$p(H_0)$: représente la probabilité que l'hypothèse « la personne mise en cause Y est effectivement auteur de l'enregistrement présenté comme indice X » soit vérifiée, avant l'analyse de x et y.

$p(H_1)$: représente la probabilité que l'hypothèse « la personne mise en cause n'est pas effectivement auteur de l'enregistrement présenté comme indice X » soit vérifiée, avant l'analyse de x et y.

$\frac{p(H_0)}{p(H_1)}$: représente le rapport de probabilité *a priori* des deux hypothèses compétitives H_0 et H1, avant l'analyse de x et y.

$p(H_0|E)$: représente l'estimation de la densité de probabilité de l'élément de preuve E, lorsque l'hypothèse que la personne mise en cause Y est la source de l'enregistrement présenté comme indice X(H_0), est vérifiée.

$p(H_1|E)$: représente l'estimation de la densité de probabilité de l'élément de preuve E, lorsque l'hypothèse que la personne mise en cause Y n'est pas la source de l'enregistrement présenté comme indice X(H_2), est vérifiée.

$\frac{p(H_0|E)}{p(H_1|E)}$: représente l'estimation du rapport de probabilité a posteriori des deux hypothèses compétitives H_0 et H_1, après l'analyse de x et y.

$\frac{p(E|H_0)}{p(E|H_1)}$: représente l'estimation du rapport de vraisemblance, Likelihood Ratio (LR), mis en évidence entre le rapport de probabilité *a priori* et le rapport de probabilité *a posteriori.*

2.11 Bases de Données d'un système FASR

Pour établir un système FASR, nous avons généralement besoin de trois Bases de Données (BD) : la BD de population potentielle (P), la BD de référence du locuteur suspect R et la BD de contrôle du locuteur suspect C, ce qui permet avec l'enregistrement incriminé (Trace) de calculer et évaluer les preuves [41], (Figure 2.8).

La BD de référence (R) contient un ensemble des enregistrements relatifs au locuteur suspect pour modéliser ses paramètres acoustiques par un modèle statistique qui est utilisé par la suite pour calculer la valeur de la preuve en comparant avec la Trace (T).

La BD de la population potentielle (P) permet de calculer la distribution scores de similarité par la comparaison de la trace (T) avec les modèles des locuteurs de la BD de la population potentielle (évaluation de la variabilité inter-locuteurs).

La BD de contrôle du locuteur suspect (C) sert à évaluer la variabilité intra-locuteur, en calculant la distribution des scores de similarité par la comparaison des vecteurs acoustiques de ladite BD avec la modèle du locuteur suspect.

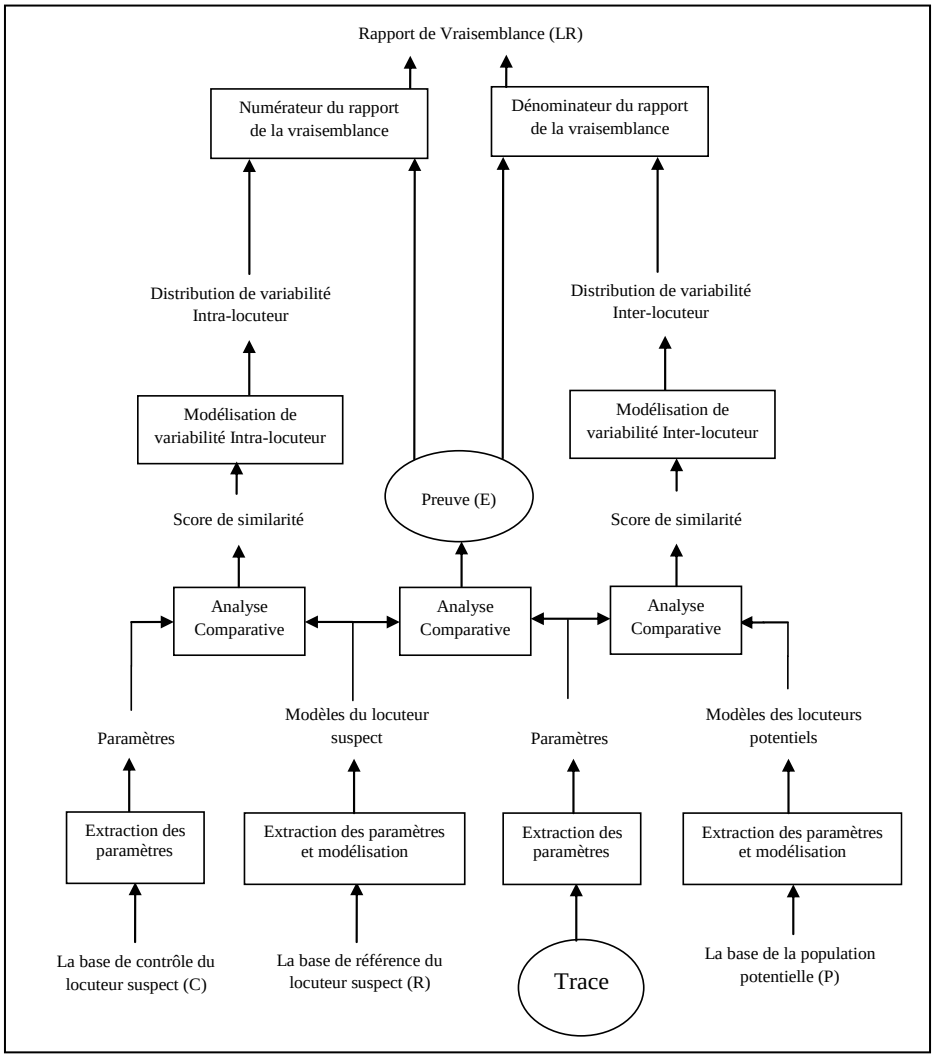


Figure 2.8 : Structure principale pour le calcul et l'interprétation des preuves [39]

2.12 Calcul de Likelihood Ratio (LR)

Il existe deux méthodes principales couramment utilisées pour calculer le rapport de vraisemblance (LR) qui dépend des modèles de locuteurs, la similarité des scores utilisé ainsi que la méthode de l'analyse de comparaison.

2.12.1 Méthode directe

Cette méthode nécessite des modèles statistiques par exemple, les GMM, HMM et les i-vectors, pour calculer le rapport de vraisemblance (LR), lorsque les vecteurs acoustiques sont comparés avec un tel modèle.

La méthode directe utilise deux BD : la base de données de la population potentielle (P) et la base de données de référence du locuteur (R), ce qui permet de créer deux modèles statistiques : le modèle statistique du locuteur suspect, et le modèle statistique de la population potentielle. Le modèle du monde (UBM) est traîné par la base de données potentielle.

Le rapport de vraisemblance est défini comme étant la probabilité relative d'observer les paramètres acoustiques extraits de la trace dans le modèle du locuteur suspect et dans le modèle du monde (Figure 2.9).

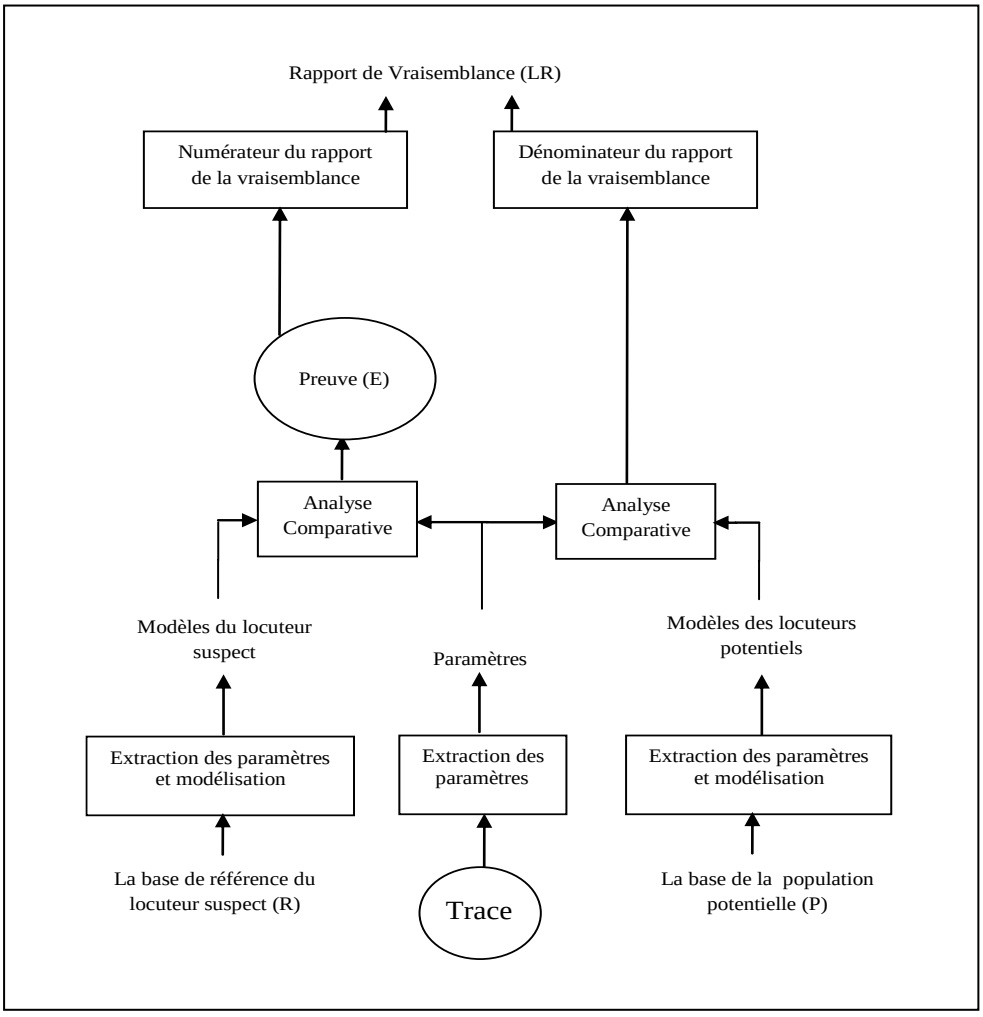


Figure 2.9 : Principe de l'approche méthodologique FASR

2.12.2 Méthode des scores

Cette méthode nécessite deux phases pour calculer le rapport de vraisemblance. La première phase consiste à calculer les scores en utilisant les paramètres acoustiques.

La deuxième phase transforme les scores de similarité en une distribution univariable basée sur la modélisation de ces scores (fonctions de la densité de la probabilité).

Le modèle statistique n'est pas utilisé seulement pour calculer la preuve en comparant la trace avec le modèle du locuteur suspect mais aussi pour modéliser la variabilité inter-locuteur de la population potentielle et la variabilité intra-locuteur du locuteur suspect. Le processus de deux phases peut se faire de plusieurs façons selon les choix que l'on fait pour formuler les hypothèses.

Si les hypothèses sont énoncées comme suit :

H_0: le locuteur suspect est la source de la voix incriminée (Trace) ;

H_1: le locuteur à l'origine de la trace n'est pas le locuteur suspect.

L'hypothèse H1 peut être représentée par la distribution inter-sources des scores de similarité qui résultent de la comparaison des paramètres acoustiques ou le modèle du locuteur interrogé avec un ou plusieurs autres locuteurs de la base de la population potentielle, et l'hypothèse H0 représentée par la distribution intra-sources des scores de similarité qui résultent de la comparaison des paramètres acoustiques ou le modèle du locuteur suspect utilisant son propre base de contrôle avec celle de la base de référence du même locuteur.

Le rapport de vraisemblance LR est défini comme étant la probabilité relative d'observer le score E dans la distribution des scores qui représente l'intra-variabilité de la voix du locuteur suspect, et la distribution des scores de l'inter-variabilité des voix de la population potentielle par rapport à la trace.

2.13 Performances métriques du système FASR

Le principe d'évaluation de la puissance d'une preuve consiste en l'estimation et la comparaison des LRs pouvant être obtenus à partir de la preuve E, d'une part lorsque l'hypothèse H_0 est vraie « le locuteur suspect est vraiment la source de la trace » dont laquelle le système fournit des valeurs de LR supérieures à 1, et d'autre part quand l'hypothèse H_1 est vraie « le locuteur suspect n'est pas vraiment la source de la trace » dont laquelle le système fournit des valeurs de LR inférieures à 1.

Pour faire l'évaluation d'un système FASR, on doit faire plusieurs expériences. Les résultats de ces expériences sont représentés graphiquement en utilisant les fonctions de la

densité de probabilité $P(LR(H_I) = LR)$, $P(LR(H_i) < LR)$ ou Tippet plots $P(LR(H_i) > LR)$ [42].

Le graphe de Tippet I, comporte deux courbes pour les deux hypothèses. Plus les deux courbes sont éloignées l'une de l'autre, plus la performance du système FASR est bonne et sa capacité de discrimination est élevée.

On peut déduire facilement du graphe (Tippet I) deux performances métriques $PMEH_0$ et $PMEH_1$ qui sont définis comme suit :

- **$PMEH_0$** (Probabiliy of Misleading Evidence in favour of hypothesis H_1): la probabilité de tous les LRs inférieurs à 1, sachant que l'hypothèse H_0 est vraie $PMEH_0 = P(LR(H_0) < 1)$.
- **$PMEH_1$** (Probabiliy of Misleading Evidence in favour of hypothesis H_0): la probabilité de tous les LRs supérieurs à 1, sachant que l'hypothèse H1 est vraie $PMEH_1 = P(LR(H_1) > 1)$.

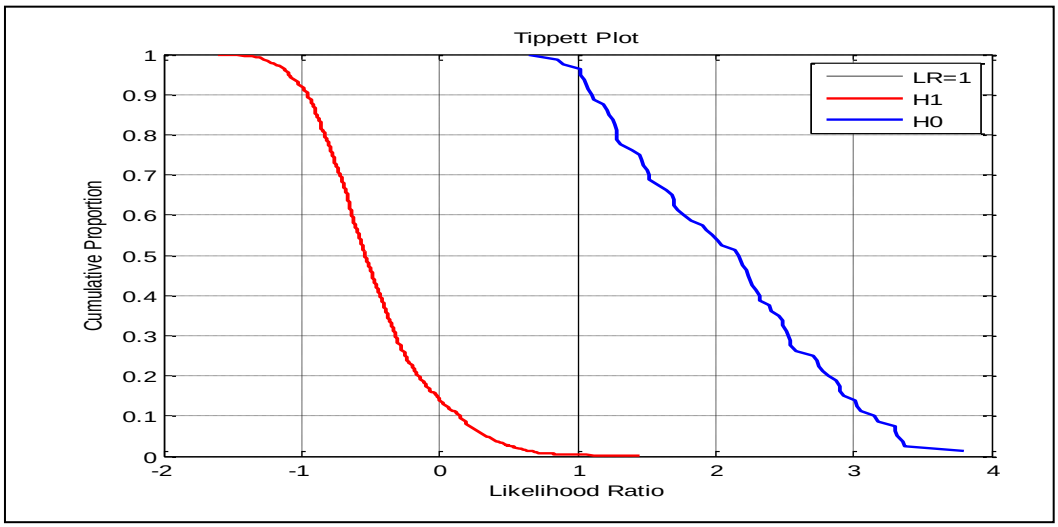


Figure 2.10 : Tippett plots I

La courbe (Tippet II) donne la représentation graphique de CDF pour l'hypothèse H_1 $P(LR(H_1) < LR)$ et l'inverse pour l'hypothèse H0 $P(LR(H_0) > LR)$. L'intersection de deux courbes résulte une autre performance métrique appelée Equal Proportion Probability (EPP) [42].

Plus l'EPP est plus proche de 0, plus les performances du système FASR sont bonnes.

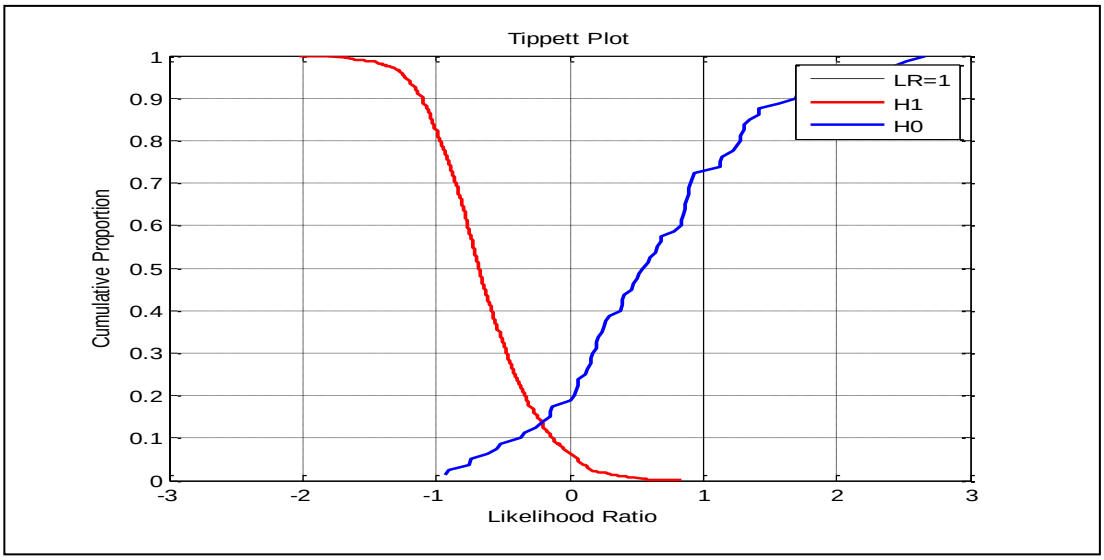


Figure 2.11 : Tippett plots II

2.14 Conclusion

Dans ce chapitre, nous avons détaillé l'architecture d'un système RAL et les différentes techniques du RCL en s'appuyant sur l'approche Bayésienne appliquée sur n'importe qu'elle discipline forensique. Ensuite nous avons vu les principales méthodes utilisées pour calculer le rapport de vraisemblance pour faire la reconnaissance forensique. Finalement, nous avons abordé les performances métriques du système FASR.

CHAPITRE 3 : TECHNIQUES DE REHAUSSEMENT DE LA PAROLE

3.1 Introduction

Le rehaussement de la parole (débruitage) est un processus très complexe qui sert à éliminer le bruit d'un signal utile. Il consiste aussi à réduire une grande quantité du bruit, tout en préservant l'intégrité du signal. Souvent, les systèmes de traitement de la parole introduisent des dégradations et modification des signaux vocaux, par exemple, bruit de quantification dans un codeur de parole ou bruit résiduel et distorsion de la parole dans un système de réduction du bruit.

La complexité dudit processus se situe au type du signal et la nature du bruit et aussi à la stationnarité ou à la non stationnarité du signal et du bruit.

Durant une longue période, plusieurs méthodes de débruitage ont été développées pour résoudre la réduction du bruit et l'amélioration de la parole dans le domaine du traitement de la parole, tels que le codage et la reconnaissance de la parole. Nous pouvons diviser ces méthodes principalement en trois catégories en fonction du domaine d'application, à savoir le domaine temporel, fréquentiel et temps-fréquence [43]. Le choix de la méthode de débruitage est principalement basé sur l'efficacité et la simplicité d'application.

Parmi ces méthodes ; la méthode de Soustraction Spectrale (SS) [44], Soustraction Spectrale avec Modèle de Sur-soustraction (SSOM) [45], Soustraction Spectrale Non linéaire (NSS) [46], Annulation Adaptative du Bruit (ANC) [47], Erreur Quadratique Moyenne Minimale (MMSE) [1], filtrage de Wiener [48][49] et la Transformée en ondelettes [50][51].

3.2 Définition et caractéristique du bruit

Le bruit est défini comme un signal nuisible qui se superpose au signal utile. Il constitue donc une gêne pour la transmission ou l'interprétation d'un signal utile, qui dans notre cas, la parole. En traitement du signal, bien que le bruit soit, par nature, aléatoire, il possède certaines caractéristiques statistiques, spectrales ou spatiales. Le Tableau 3.1, représente les différentes classes auxquelles un bruit peut appartenir [52].

Comme notre but est essentiellement le débruitage de la parole, nous nous limitons dans notre travail aux bruits additifs, stationnaires ou non stationnaires, et décorrélé avec la parole (indépendance au sens statistique), tels que le bruit de bavardage, de l'usine et le bruit blanc.

Tableau 3.1 : Différentes classes de bruit

Propriétés	Types
Structure	Continu/impulsif/périodique
Type d'interaction	Additif/convolutif
Comportement temporel	Stationnaire/non-stationnaire
Bande de fréquence	Etroit/large
Dépendance	Corrélé/décorrélé
Propriétés statistiques	Dépendant/indépendant
Propriétés spatiales	Cohérent/incohérent

3.3 Différents types de bruit

Les différents bruits pouvant influés sur un message peuvent être divisés en deux grandes catégories : les bruits additifs et les bruits convolutionnels. La distinction entre les deux peut être faite par le nombre d'agents agresseurs extérieurs à la transmission du message. Les bruits additifs sont causés par des agents extérieurs au trinôme source-voie-destinataire alors que les bruits convolutionnels sont causés par la moindre qualité de la voie de communication, celle-ci ayant alors un rôle ambigu, du point de vue du message, de médium et d'agresseur.

3.3.1 Bruits additifs

Les bruits additifs sont dûs à la multiplicité des systèmes de communication dans un même environnement. Plusieurs émetteurs et plusieurs receveurs pouvant être confinés dans un même espace, les messages de tous les émetteurs peuvent donc se trouver en concurrence sur une même voie sans que les récepteurs possèdent un mécanisme infaillible pour isoler le message qui leur est destiné. L'émetteur et le récepteur peuvent aussi se trouver en présence d'un ou de plusieurs équipements générant un bruit de fond de force variable.

Les bruits additifs peuvent être subdivisés en trois catégories en fonction des lieux où ils peuvent être rencontrés :

- Bruits des systèmes industriels : ils peuvent être très intenses et sont, par nature, non stationnaires. Ils correspondent aux bruits émis par des machines possédant une faible isolation phonique ;
- Bruits des moyens de transport : ils correspondent aux bruits qui peuvent être observés dans divers véhicules tels que les voitures, les trains ou les avions ;

- Bruits des milieux administratifs et urbain : ce sont les bruits présents dans les bureaux, les domiciles ou dans les concentrations urbaines. Ces bruits peuvent être très variés (climatisation ou bruit de parole) mais sont peu intenses [53].

3.3.2 Bruits convolutionnels

La parole est l'information la plus véhiculée dans les systèmes de télécommunications et un grand effort de recherche en Traitement Automatique de la Parole a ciblé les réseaux téléphoniques [54].

Les bruits convolutionnels sont dûs à la distorsion induite par la voie de communication. Ils résultent de la mauvaise qualité d'un ou de plusieurs éléments de support de message ou, tout simplement, de son étroitesse en bande passante. De nos jours les moyens de communication les plus utilisés sont les téléphones. La parole, lorsqu'elle est transmise par un tel moyen, est forcément dégradée tout en gardant une grande intelligibilité.

3.4 Estimation du bruit

L'estimation du bruit peut avoir une influence importante sur la qualité du signal amélioré. Si l'estimation de bruit est trop faible, un bruit résiduel gênant sera audible, et si l'estimation de bruit est trop élevée, la parole sera déformée, ce qui peut engendrer une perte d'intelligibilité.
Le but de cette section est d'estimer le bruit pour appliquer à la sortie un algorithme de réduction de bruit. Le calcul du bruit se fait généralement sur la Densité Spectrale de Puissance (DSP) qui est obtenue en calculant l'énergie (Théorème de Parseval).

La plupart des algorithmes de débruitage sont basés sur les hypothèses suivantes :

- Comme mentionné dans la sous section 3.3.1, le signal de parole est dégradé par un bruit additif statiquement indépendant ;
- La parole n'est pas toujours présente. Ainsi, on peut toujours trouver des segments d'analyse, formé par quelques trames consécutives, qui contiennent une pose de parole ou du bruit uniquement ;
- Le bruit est plus stationnaire que le signal de parole propre, donc on peut supposer que le bruit reste stationnaire dans un segment d'analyse donné.

3.5 Réduction du bruit musical

Le bruit musical est un bruit résiduel constamment gênant suite à un débruitage d'un signal de parole par un algorithme. Par exemple, les deux algorithmes cités précédemment, la soustraction spectrale et le filtrage de Wiener sont des réducteurs de bruit à court terme. Le

spectre de ce bruit est tonal d'où le caractère musical. Le bruit musical est plus gênant que le bruit de base [55]. Les principales causes de l'apparition de ce type du bruit sont :

- L'évaluation non précise de la densité spectrale du bruit ;
- L'estimation basée sur les spectrogrammes.

3.6 Atténuation spectrale à court terme

Le but consiste à estimer le signal de parole utile x(n), perturbé par un bruit additif d(n) supposé indépendant du signal de parole, à partir du signal observé y(n), (Figure 3.1).

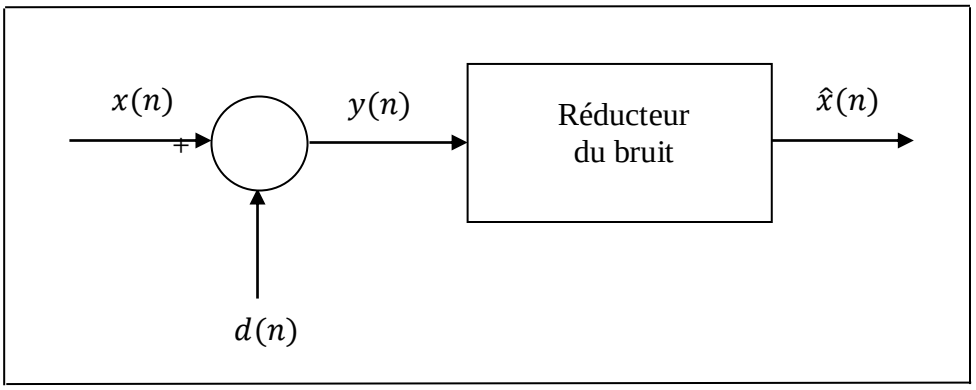


Figure 3.1 : Modèle de rehaussement de parole monovoie

$$y(n) = x(n) + d(n) \tag{3.1}$$

En appliquant la Transformée de Fourier à Court Terme (TFCT), nous obtenons :

$$Y(w_k) = X(w_k) + D(w_k) \tag{3.2}$$

L'Equation 3.2 peut être exprimée sous forme polaire comme suit :

$$Y_k e^{j\theta_y(k)} = X_k e^{j\theta_x(k)} + D_k e^{j\theta_d(k)} \tag{3.3}$$

Où, $\{Y_k, X_k, D_k\}$ sont respectivement les amplitudes et $\{\theta_y(k), \theta_x(k), \theta_d(k)\}$ sont les phases à la fréquence discrète $f_{k.}$ du signal de la parole bruité, propre et du bruit.

3.7 Définition des rapports Signal sur Bruit (SNR)

Un gain spectral $G(p,k)$ qui dépend du SNR est obtenu puis est appliqué au spectre bruité $Y(p,k)$:

$$\hat{X}(p,k) = G(p,k)Y(p,k) \tag{3.4}$$

Où, p désigne l'indice temporel de trame.

Pour exprimer le gain spectral, deux types de SNR sont utilisés, le SNR a posteriori et le SNR a priori [55].

$$SNR_{post}(p,k) = \frac{|Y(p,k)|^2}{E[|D(p,k)|^2]} = \frac{|Y(p,k)|^2}{\gamma_d(k)} \tag{3.5}$$

$$SNR_{prio}(k) = \frac{E[|X(p,k)|^2]}{E[|D(p,k)|^2]} = \frac{\gamma_x(k)}{\gamma_d(k)} \tag{3.6}$$

3.8 Mise en œuvre de l'atténuation spectrale à court terme

Les techniques de réduction de bruit par atténuation spectrale à court terme sont effectuées suivant les étapes suivantes :

- Le signal de parole bruité $y(n)$, est découpé en trames, puis chaque trame est fenêtrée. Les fenêtres peuvent se chevaucher. Chaque trame d'analyse est transformée dans le domaine fréquentiel où l'étape de réduction de bruit est réalisable, en utilisant la Transformée de Fourier Discrète (TFD). Cette étape constitue la Transformée de Fourier à Court Terme (TFCT) ;
- Une estimation de la DSP du bruit à long terme est réalisée, pendant la présence du bruit seul, ce qui nécessite une Détection d'Activité Vocale (VAD), ou de façon continue pendant l'activité vocale ;
- Une atténuation spectrale à court terme est appliquée au module du signal de parole bruité, en calculant le gain spectral. Ce dernier requiert l'estimation de $SNR_{post}(p,k)$ et $SNR_{prio}(k)$;
- Le module du signal de parole estimé $|\hat{x}(p,n)|$ et la phase du signal bruité sont alors utilisés pour revenir dans le domaine temporel en utilisant une TFD Inverse (TFDI). Le signal de sortie est synthétisé à partir d'une technique de type OLS (OverLap and Save) ou OLA (OverLap and Add).

Le principe général du système d'amélioration du signal de parole est représenté dans la Figure 3.2.

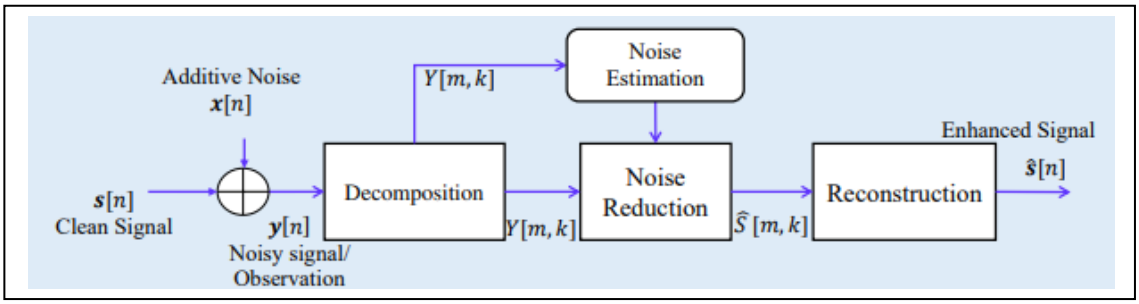


Figure 3.2 : Principe général du système d'amélioration du signal de parole [56]

3.9 Principales méthodes d'atténuation spectrale à court terme

Les techniques de réduction de bruit par atténuation spectrale à court terme peuvent être classées de différentes méthodes. Nous allons différencier les approches ne nécessitant pas de modèle statistique pour les signaux traités de celles qui en requièrent.

3.9.1 Approches ne nécessitant pas de modèle statistique

Les approches qui ne nécessitent pas de modèle statistique seront citées par la suite.

3.9.1.1 Soustraction Spectrale

La Soustraction Spectrale est un algorithme de traitement du signal le plus ancien, destinée pour réduire les effets du bruit additif par une opération de soustraction spectrale. Bien que cette méthode soit efficace pour plusieurs applications de rehaussement de la parole, il existe des lacunes inhérentes à sa capacité à éliminer efficacement le bruit, y compris la production du bruit musical [57]. L'estimation du bruit se fait sur plusieurs trames d'acquisition [55]. La Soustraction Spectrale d'Amplitude (SSA) à court terme est estimée comme ceci :

$$\left|\hat{X}(p,k)\right| = |Y(p,k)| - \sqrt{E[|D(p,k)|^2]} \tag{3.7}$$

Le module du spectre du signal estimé doit rester positif ou nul, une valeur négative n'ayant pas de signification physique, cette contrainte est satisfaite par un simple seuillage :

$$\left|\hat{X}(p,k)\right| = \begin{cases} |Y(p,k)| - \sqrt{E[|D(p,k)|^2]} & si\ |Y(p,k)| \geq \sqrt{E[|D(p,k)|^2]} \\ 0 & si\ non \end{cases} \tag{3.8}$$

Plusieurs travaux ont été initiés, pour améliorer la soustraction spectrale par Boll [58], Berouti [59] et Virag [52].

3.9.1.2 Filtrage de Wiener

Le filtre de Wiener est le filtre optimal au sens de l'Erreur Quadratique Moyenne Minimum (EQMM), il adapte le rapport signal sur bruit pour chaque trame traitée. Il minimise la fonction d'erreur suivante :

$$E[(X(p,k) - G_w(p,k)Y(p,k))^2] \tag{3.9}$$

En s'appuyant sur les équations de la partie 3.7, le filtre de Wiener peut s'exprimer ainsi :

$$G_w(p,k) = \frac{E[|X(p,k)|^2]}{E[|Y(p,k)|^2]} = \frac{E[|X(p,k)|^2]}{E[|X(p,k)|^2] + E[|D(p,k)|^2]} \tag{3.10}$$

Les quantités intervenant dans l'expression du filtre de Wiener étant calculées à long terme, on peut l'exprimer en fonction du SNR a priori :

$$G_w(p,k) = \frac{SNR_{prio}(p,k)}{1 + SNR_{prio}(p,k)} \tag{3.11}$$

3.9.2 Approches nécessitant des modèles statistiques

Dans cette section, nous allons aborder les différents estimateurs qui utilisent les modèles statistiques et critères d'optimisation. Ces estimateurs non linéaires utilisés pour le rehaussement de la parole, prennent en compte la Fonction de la Densité de Probabilité (PDF) de bruit, les coefficients (DFT) de la parole et la probabilité de présence de parole.

3.9.2.1 Estimateur de l'Erreur Quadratique Moyenne Minimum (MMSE) du spectre à court terme du signal bruité

La méthode de soustraction spectrale basée sur MMSE et les Statistiques de bruit Minimum (MS) [60] a été utilisée pour améliorer le signal de parole endommagé par le bruit additif. L'amplitude du signal bruité a été multipliée par un certain facteur de gain. La soustraction spectrale introduite par Boll [58], est la méthode la plus ancienne pour supprimer le bruit. Il fonctionne dans le domaine fréquentiel, et son principe est de soustraire une estimation de bruit du signal observé. Le bruit est supposé additif, stationnaire ou légèrement variable, ce qui permet de l'estimer pendant les périodes de silences.

L'estimateur MMSE du spectre de puissance à court terme (MMSE) est donné par [61] comme ceci :

$$\begin{aligned}\hat{X}_k^2 &= E\{X_k^2/Y(w_k)\} \\ &= \int_0^\infty X_k^2 f_{X_k}(X_k/Y(w_k))dX_k \\ &= \frac{\xi_k}{1+\xi_k}\left(\frac{1}{\gamma_k}+\frac{\xi_k}{1+\xi_k}\right)Y_k^2\end{aligned} \tag{3.12}$$

et,

$$\xi_k = \frac{\sigma_x^2(k)}{\sigma_d^2(k)}, \quad \gamma_k = \frac{Y_k^2}{\sigma_d^2(k)} \tag{3.13}$$

$$\sigma_x^2(k) = E\{X_k^2\}, \quad \sigma_d^2(k) = E\{D_k^2\} \tag{3.14}$$

Où, ξ_k et γ_k désignent respectivement, a priori et a posteriori les SNRs.

Les dérivations de l'estimateur MMSE, ci-dessus étaient basées sur la densité a posteriori $f_{X_k}(X_k/Y(w_k))$:

$$f_{X_k}(X_k/Y(w_k)) = \frac{X_k}{\sigma_k^2}\exp\left(-\frac{X_k^2+s_k^2}{2\sigma_k^2}\right)I_0\left(\frac{X_k s_k}{\sigma_k^2}\right) \tag{3.15}$$

où,

$$\frac{1}{\lambda'(k)} = \frac{1}{\sigma_x^2(k)} + \frac{1}{\sigma_d^2(k)} \tag{3.16}$$

$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k \tag{3.17}$$

$$\sigma_k^2 = \frac{\lambda'(k)}{2}, \quad s_k^2 = v_k\lambda'(k) \tag{3.18}$$

$I_0(.)$ est la première fonction de Bessel modifiée d'ordre zéro.

Cependant, l'analyse des courbes de suppression a révélé que la règle de suppression de puissance spectrale MMSE de l'équation 3.12 fournit moins de suppression dans les régions de faible SNR a priori [61]. Yang Lu et al [1] ont proposé l'estimateur MMSE amélioré du spectre de puissance à court terme, pour remédier au problème de la faiblesse de la phase de réduction de bruit dans les régions à faible SNR a priori.

Le spectre de puissance du signal corrompu par le bruit est supposé être la somme des spectres de puissance de la parole propre et du bruit, écrit comme suit :

$$P_y(w) = P_x(w) + P_d(w) \tag{3.19}$$

De plus, une hypothèse est utilisée dans la dérivation de ces estimateurs sur la base de l'équation 3.19 en approximant le spectre de puissance en utilisant le spectre d'amplitude au

carré, qui est l'estimation d'échantillon de la moyenne d'ensemble. Par conséquent, l'équation 3.19 peut s'écrire comme suit :

$$Y_k^2 \approx X_k^2 + D_k^2 \tag{3.20}$$

Aussi, en supposant que les parties réelle et imaginaire des coefficients de la Transformée de Fourier Discrète (TFD) sont modélisées comme des variables aléatoires gaussiennes indépendantes [62], la densité de probabilité de $\mathrm{X_k^2}$ peut s'écrire comme suit :

$$f_{X_k^2}(X_k^2) = \frac{1}{\sigma_{\mathrm{x}}^2(\mathrm{k})} \mathrm{e}^{-\frac{\mathrm{X_k^2}}{\sigma_{\mathrm{x}}^2(\mathrm{k})}} \tag{3.21}$$

De même, la densité de $\mathrm{D_k^2}$ est donnée par l'équation 3.22 :

$$f_{D_k^2}(D_k^2) = \frac{1}{\sigma_{\mathrm{d}}^2(\mathrm{k})} \mathrm{e}^{-\frac{\mathrm{D_k^2}}{\sigma_{\mathrm{d}}^2(\mathrm{k})}} \tag{3.22}$$

Où, $\sigma_{\mathrm{x}}^2(\mathrm{k})$ et $\sigma_{\mathrm{d}}^2(\mathrm{k})$ sont donnés par l'équation 3.14.

Lu et Loizou [1] proposent un estimateur MMSE du spectre d'amplitude au carré. Dans ce cas, la densité de probabilité a postériori du spectre de l'amplitude au carré du signal de parole propre est obtenue en utilisant la règle de Bayes comme suit :

$$f_{X_k^2}(X_k^2/Y_k^2) = \frac{f_{Y_k^2}(Y_k^2/X_k^2) f_{X_k^2}(X_k^2)}{f_{Y_k^2}(Y_k^2)}$$

$$= \begin{cases} \psi_k e^{-\frac{X_k^2}{\lambda(k)}}, & \text{si } \sigma_x^2(k) \neq \sigma_d^2(k) \\ \dfrac{1}{Y_k^2}, & \text{si } \sigma_x^2(k) = \sigma_d^2(k) \end{cases} \tag{3.23}$$

$\lambda(k)$ est défini comme :

$$\frac{1}{\lambda(k)} = \frac{1}{\sigma_x^2(k)} - \frac{1}{\sigma_d^2(k)}, \qquad \text{si } \sigma_x^2(k) \neq \sigma_d^2(k) \tag{3.24}$$

et

$$\psi_k = \frac{1}{\lambda(k)\left\{1 - exp\left[-\frac{Y_k^2}{\lambda(k)}\right]\right\}}. \tag{3.25}$$

En utilisant les équations 3.20 et 3.23, l'estimateur MMSE est obtenu en calculant la moyenne de la densité a posteriori donnée dans l'équation 3.23 :

$$\begin{aligned} \hat{X}_k^2 &= E\{X_k^2/Y_k^2\} \\ &= \int_0^{Y_k^2} X_k^2 f_{X_k^2}(X_k^2/Y_k^2) dX_k^2 \\ &= \begin{cases} \left(\frac{1}{\upsilon_k} - \frac{1}{e^{\upsilon_k - 1}}\right) Y_k^2, & \text{si } \sigma_x^2(k) \neq \sigma_d^2(k) \\ \frac{1}{2} Y_k^2, & \text{si } \sigma_x^2(k) = \sigma_d^2(k) \end{cases} \end{aligned} \tag{3.26}$$

Où, υ_k est défini comme :

$$\upsilon_k = \frac{1 - \xi_k}{\xi_k} \gamma_k \tag{3.27}$$

3.9.2.2 Estimateur Maximum Likelihood (ML)

Cette approche est utilisée par McAulay et Malpass [63][64][65] pour l'amélioration de la parole. Les spectres d'amplitude et de phase X_k et $\theta_x(k)$ du signal propre, sont supposés inconnus mais déterministes. La PDF des coefficients de Transformée de Fourier du bruit est supposé être gaussien complexe et de moyenne nulle. Les parties réelle et imaginaire de $D(w_k)$ sont supposées avoir des variances $\frac{\lambda_d(k)}{2}$. Sur la base de ces deux hypothèses, nous pouvons former la densité de probabilité des coefficients DFT de la parole bruyante observés $Y(w_k)$.

La densité de probabilité de $Y(w_k)$ est également gaussienne avec la variance $\lambda_d(k)$ et la moyenne $X_k e^{j\theta_x(k)}$.

$$\begin{aligned} p\left(Y(w_k; X_k, \theta_x(k))\right) &= \frac{1}{\pi\lambda_d(k)} exp\left[-\frac{\left|Y(w_k) - X_k e^{j\theta_x(k)}\right|^2}{\lambda_d(k)}\right] \\ &= \frac{1}{\pi\lambda_d(k)} exp\left[-\frac{Y_k^2 - 2X_k Re\left\{e^{-j\theta_x(k)} Y(w_k)\right\} + X_k^2}{\lambda_d(k)}\right] \end{aligned} \tag{3.28}$$

Pour obtenir l'estimation ML de X_k, nous devons calculer le maximum de $p\left(Y(w_k; X_k, \theta_x(k))\right)$ par rapport à X_k, ce n'est pas simple, cependant, car $p\left(Y(w_k; X_k, \theta_x(k))\right)$ est une fonction de deux paramètres inconnus : l'amplitude et la phase. On peut éliminer le paramètre de phase en maximisant à la place la moyenne de la fonction de vraisemblance suivante :

$$p_L(Y(w_k); X_k) = \int_0^{2\pi} p(Y(w_k); X_k, \theta_x)\, p(\theta_x) d\theta_x \tag{3.29}$$

En supposant une distribution uniforme sur $(0,2\pi)$ pour la phase θ_x, c'est-à-dire, on suppose que $p(\theta_x) = \frac{1}{2\pi}$ pour $\theta_x \in [0, 2\pi]$, la fonction de vraisemblance devient :

$$p_L(Y(w_k); X_k) = \frac{1}{\pi\lambda_d(k)} exp\left[-\frac{Y_k^2 + X_k^2}{\lambda_d(k)}\right] \frac{1}{2\pi} \int_0^{2\pi} exp\left[\frac{2X_k Re\left(e^{-j\theta_x} Y(w_k)\right)}{\lambda_d(k)}\right] d\theta_x \tag{3.30}$$

L'intégrale de l'équation précédente est connue sous le nom de fonction de Bessel modifiée de premier type et donné par :

$$I_0(|x|) = \frac{1}{2\pi} \int_0^{2\pi} exp\left[Re\left(xe^{-j\theta_x}\right)\right] d\theta_x \tag{3.31}$$

La fonction de Bessel précédente peut être approximée comme suit :

$$I_0(|x|) \approx \frac{1}{\sqrt{2\pi|x|}} exp(|x|) \tag{3.32}$$

et la fonction de vraisemblance de l'équation 3.30 se simplifie en :

$$p(Y(w_k); X_k) = \frac{1}{\pi\lambda_d(k)} \frac{1}{\sqrt{2\pi \frac{2X_k Y_k}{\lambda_d(k)}}} exp\left[-\frac{Y_k^2 + X_k^2 - 2Y_k X_k}{\lambda_d(k)}\right] \tag{3.33}$$

Après avoir différencié la fonction log-vraisemblance $\log p(Y(w_k); X_k)$ par rapport au paramètre inconnu X_k et en réglant la dérivée à zéro, nous obtenons l'estimation ML du spectre d'amplitude :

$$\hat{X}_k = \frac{1}{2}\left[Y_k + \sqrt{Y_k - \lambda_d(k)}\right] \tag{3.34}$$

3.9.2.3 Estimateur Maximum A Posteriori (MAP)

Les algorithmes MAP sont souvent utilisés comme alternative aux algorithmes MMSE dans des circonstances, dont il est extrêmement difficile de calculer la moyenne de la PDF a posteriori [66][67]. Dans certains cas, il est plus facile de maximiser la PDF a posteriori $p(x_k, \theta_x / Y(w_k))$ plutôt que pour évaluer la moyenne de $p(x_k, \theta_x / Y(w_k))$.

Les estimateurs MAP d'amplitude et de la phase peuvent être dérivés comme suit :

$$(\hat{x}_k, \hat{\theta}_x) = argmax_{x_k,\theta_x} p(x_k, \theta_x / Y(w_k)) \tag{3.35}$$

En utilisant la règle de Bayes, nous pouvons exprimer $p(x_k, \theta_x / Y(w_k))$ comme suit :

$$p(x_k, \theta_x / Y(w_k)) = \frac{p(Y(w_k)/ x_k, \theta_x) p(x_k, \theta_x)}{p(Y(w_k))} \tag{3.36}$$

Puisque $p(Y(w_k))$ n'est pas une fonction de x_k ou θ_x, nous pouvons maximiser $p(Y(w_k)/ x_k, \theta_x) p(x_k, \theta_x)$. les estimateurs MAP de x_k et θ_x peuvent alors être obtenus comme solution de :

$$(\hat{x}_k, \hat{\theta}_x) = argmax_{x_k,\theta_x} p(x_k, \theta_x / Y(w_k)) \tag{3.37}$$

En supposant le modèle statistique gaussien et en utilisant les équations (7.38) et (7.39) données dans [65], nous avons :

$$p(Y(w_k)/ x_k, \theta_x) = \frac{1}{\pi \lambda_d(k)} exp\left\{\frac{-1}{\lambda_d(k)} |Y(w_k) - X(w_k)|^2\right\} \tag{3.38}$$

$$p(x_k, \theta_x) = \frac{x_k}{\pi \lambda_d(k)} exp\left\{\frac{-x_k^2}{\lambda_x(k)}\right\} \tag{3.39}$$

nous avons :

$$p(Y(w_k)/ x_k, \theta_x) p(x_k, \theta_x) = \frac{x_k}{\pi^2 \lambda_x(k) \lambda_d(k)} \exp\left(-\frac{\left[Y(w_k) - x_k e^{j\theta_x}\right]^2}{\lambda_d(k)} - \frac{{x_k}^2}{\lambda_x(k)}\right) \tag{3.40}$$

Comme la fonction log est une fonction à croissance monotone, nous pouvons alternativement, maximiser le logarithme de l'équation précédente, c'est-à-dire :

$$J_1 = ln[p(Y(w_k)/ x_k, \theta_x) p(x_k, \theta_x)] = -\frac{\left[Y(w_k) - x_k e^{j\theta_x}\right]^2}{\lambda_d(k)} - \frac{{x_k}^2}{\lambda_x(k)} + lnx_k + cste \tag{3.41}$$

Après la différenciation de J_1 par rapport à la phase θ_x et la mise à zéro de la dérivée, nous obtenons :

$$\frac{\partial J_1}{\partial \theta_x} = 2jsin(\theta_y - \theta_x) = 0 \tag{3.42}$$

Donc :

$$\hat{\theta}_x = \theta_y \tag{3.43}$$

Maintenant, différencier J_1 par rapport à l'amplitude x_k et définir la dérivée égale à zéro, nous obtenons l'estimateur d'amplitude MAP.

$$\hat{X}_k = \frac{\xi_k + \sqrt{\xi_k + 2(1+\xi_k)/\gamma_k}}{2(1+\xi_k)} Y_k \tag{3.44}$$

3.9.2.4 Estimateur Log MMSE

Il a été proposé qu'une méthode basée sur l'erreur quadratique des spectres d'amplitude logarithmique peut être plus appropriée pour le traitement de la parole [62][68]. Ensuite, nous dérivons un estimateur qui minimise l'erreur quadratique moyenne des spectres de log-amplitude [62].

$$E\left\{\left(logX_k - log\hat{X}_k\right)^2\right\} \tag{3.45}$$

L'estimateur log-MMSE peut être obtenu en évaluant la moyenne conditionnelle de $\log X_k$, c'est- à-dire :

$$log\hat{X}_k = E\{(logX_k/Y(w_k))\} \tag{3.46}$$

Nous pouvons obtenir $\hat{X}_k$,

$$\hat{X}_k = \exp(E\{logX_k/Y(w_k)\}) \tag{3.47}$$

L'évaluation de $E\{\log X_k/Y(w_k)\}$ n'est pas simple mais peut être simplifiée, si nous utilisons la fonction génératrice de moment de X_k conditionnée sur $Y(w_k)$.

Soit $Z_k = \log X_k$, alors la fonction génératrice de moment de Z_k conditionnée sur $Y(w_k)$ est donné par :

$$\Phi_{Z_k/Y(w_k)}(\mu) = E\{exp[\mu Z_k]/Y(w_k)\} = E\{X_k^{\mu}/Y(w_k)\} \quad (3.48)$$

La moyenne conditionnelle de $\log X_k$ peut alors être obtenue à partir de la génération de moment fonction en évaluant la dérivée de $\Phi_{Z_k|Y(w_k)}(\mu)$ à $\mu = 0$, donc,

$$E\{logX_k/Y(w_k)\} = \frac{d}{d\mu}\phi_{Z_k/Y(w_k)}(\mu)|\mu = 0 \quad (3.49)$$

Aussi, dans la proche MSE (Erreur Quadratique Moyenne), l'estimateur $\hat{X}_k$ est donné par l'Equation suivante :

$$\hat{X}_k = E[X_k/Y(w_k)] = \int_0^{\infty} x_k p(x_k/Y(w_k))dx_k = \frac{\int_0^{\infty} x_k p(Y(w_k)/x_k)p(x_k)dx_k}{\int_0^{\infty} p(Y(w_k)|x_k)p(x_k)dx_k} \quad (3.50)$$

Il reste alors la tâche d'évaluer la fonction génératrice de moment $\phi_{Z_k/Y(w_k)}(\mu)$. A partir de l'équation 3.48, nous voyons que nous devons évaluer le terme $E\{X_k^{\mu}/Y(w_k)\}$ qui est très similaire (sauf pour la puissance μ) à l'équation 3.50, donc,

$$\Phi_{Z_k/Y(w_k)}(\mu) = E\{X_k^{\mu}/Y(w_k)\} = \frac{\int_0^{\infty}\int_0^{2\pi} x_k^{\mu} p(Y(w_k)/x_k,\theta_x)p(x_k,\theta_x)d\theta_x dx_k}{\int_0^{\infty}\int_0^{2\pi} p(Y(w_k)/x_k,\theta_x)p(x_k,\theta_x)d\theta_x dx_k} \quad (3.51)$$

En utilisant le même modèle statistique que le calcul de l'estimateur MMSE, et après en remplaçant les équations 3.38 et 3.39 dans l'équation 3.51, nous obtenons :

$$\Phi_{Z_k/Y(w_k)}(\mu) = \lambda_k^{\mu/2}\Gamma\left(\frac{\mu}{2}+1\right)\phi\left(-\frac{\mu}{2},1;-v_k\right) \quad (3.52)$$

Où

$\Gamma(.)$ est la fonction gamma ;

$\phi(a, b; x)$ est la fonction hypergéométrique confluente [61] ;

υ_k est défini dans l'équation 3.17 ;

λ_k est donné par l'équation suivante :

$$\lambda_k = \frac{\sigma_x^2(k)}{1+\xi_k} \tag{3.53}$$

Après avoir pris la dérivée de $\Phi_{Z_k/Y(w_k)}(\mu)$ par rapport à μ en l'évaluant à $\mu = 0$, on obtient la moyenne conditionnelle de $\log X_k$:

$$E\{logX_k/Y(w_k)\} = \frac{1}{2}log\lambda_k + \frac{1}{2}logv_k + \frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt \tag{3.54}$$

Finalement, en remplaçant l'équation précédente dans l'équation 3.47, nous obtenons la valeur optimale estimateur log-MMSE :

$$\begin{aligned}\hat{X}_k &= \frac{\xi_k}{\xi_k+1}exp\left\{\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right\}Y_k \\ &= G_{LSA}(\xi_k, v_k)Y_k\end{aligned} \tag{3.55}$$

Où

ξ_k est le SNR a priori ;

$G_{LSA}(\xi_k, v_k)Y_k$ est la fonction de gain de l'estimateur log-MMSE.

L'intégrale de l'équation précédente peut être approximée comme suit :

$$Ei(x) = \int_x^{\infty}\frac{e^{-x}}{x}dx = \frac{e^x}{x}\sum_k\frac{k!}{x^k} \tag{3.56}$$

3.9.2.5 Estimateur Incorporating speech presence probability in MMSE (MMSE-ISP)

Le modèle à deux états pour les événements vocaux peut être exprimé à l'aide d'un modèle d'hypothèse binaire [65] :

$$\begin{aligned}&H_0^k\text{: parole absente : } |Y(w_k)| = |D(w_k)| \\ &H_1^k\text{: parole presente : } |Y(w_k)| = |X(w_k)| + |D(w_k)| \\ &\qquad\qquad\qquad\qquad = \left|X_k e^{j\theta_x} + |D(w_k)|\right|\end{aligned} \tag{3.57}$$

H_0^k désigne l'hypothèse nulle selon laquelle la parole est absente dans le cas de fréquence discrète fk :

H_1^k désigne l'hypothèse que le parole est présente.

L'estimateur MMSE-ISP est donné par :

$$\hat{X}_k = E\left(X_k/Y(w_k).H_1^k\right)P\left(H_1^k/Y(w_k)\right) \tag{3.58}$$

Pour calculer $P\left(H_1^k/Y(w_k)\right)$, on utilise la règle de Bayes :

$$P\left(H_1^k/Y(w_k)\right) = \frac{p\left(Y(w_k)/H_1^k\right)P(H_1)}{p\left(Y(w_k)/H_1^k\right)P(H_1) + p\left(Y(w_k)/H_0^k\right)P(H_0)} = \frac{\Lambda(Y(w_k), q_k)}{1 + \Lambda(Y(w_k),\ q_k)} \tag{3.59}$$

Où $\Lambda(Y(w_k), q_k)$ est le rapport de vraisemblance généralisé, défini par :

$$\Lambda(Y(w_k), q_k) = \frac{1 - q_k}{q_k}\frac{p(Y(w_k)/H_1)}{p(Y(w_k)/H_0)} \tag{3.60}$$

Où $q_k = P\left(H_0^k\right)$ désigne la probabilité a priori d'absence de parole pour la fréquence discrète fk. La probabilité a priori de présence de la parole, c'est-à-dire $P\left(H_1^k\right)$, est donnée par $(1 - q_k)$.

Sous l'hypothèse H_0, $(Y(w_k) = D(w_k)$, et comme la PDF des coefficients de la transformée de Fourier du bruit, $D(w_k)$, est gausssien complexe avec une moyenne et une variance nulles $\lambda_d(k)$, il suit que $p\left(Y(w_k)/H_0^k\right)$ aura aussi une distribution gaussienne avec la même variance, alors :

$$p\left(Y(w_k)/H_0^k\right) = \frac{1}{\pi\lambda_d(k)} exp\left(-\frac{Y_k^2}{\lambda_d(k)}\right) \tag{3.61}$$

Sous l'hypothèse H_1, $Y(w_k) = X(w_k) + D(w_k)$ et parce que les PDF de $X(w_k)$ et $D(w_k)$ sont gaussiennes complexes avec une moyenne et des variances nulles $\lambda_x(k)$ et $\lambda_d(k)$, respectivement, ensuite $Y(w_k)$ aura aussi une distribution gaussienne de variance $\lambda_d(k) + \lambda_x(k)$ (puisque $D(w_k)$ et $X(w_k)$ ne sont pas corrélés.

$$p\big(Y(w_k)/H_1^k\big) = \frac{1}{\pi[\lambda_k(k) + \lambda_x(k)]} exp\left(-\frac{Y_k^2}{\lambda_d(k) + \lambda_x(k)}\right) \quad (3.62)$$

En substituant les équations 3.61 et 3.62 à l'équation 3.60, nous obtenons une expression pour le rapport de vraisemblance :

$$\Lambda(Y(w_k), q_k, \xi_k') = \frac{1 - q_k}{q_k} \frac{exp[(\xi_k'/(1 + \xi_k')\gamma_k]}{1 + \xi_k'} \quad (3.63)$$

Où ξ_k' indique le SNR conditionnel a priori :

$$\xi_k' = \frac{E[X_k^2/H_1^k]}{\lambda_d(k)} \quad (3.64)$$

Le SNR peut être exprimé en termes de SNR inconditionnel ξ_k comme suit :

$$\begin{aligned} \xi_k &= \frac{E[X_k^2]}{\lambda_d(k)} \\ &= P\big(H_1^k\big) \frac{E[X_k^2/H_1^k]}{\lambda_d(k)} \\ &= (1 - q_k)\xi_k' \end{aligned} \quad (3.65)$$

Par conséquent, le SNR conditionnel ξ_k' est lié au SNR inconditionnel ξ_k par :

$$\xi_k' = \frac{\xi_k}{1 - q_k} \quad (3.66)$$

En remplaçant l'équation 3.63 dans l'équation 3.59 et après quelques manipulations algébriques, nous exprimons la probabilité a posteriori de la présence de la parole comme :

$$P\left(H_1^k/Y(w_k)\right) = \frac{1 - q_k}{1 - q_k + q_k(1 + \xi_k')exp(-v_k')} \quad (3.67)$$

$$v_k' = \frac{\xi_k'}{\xi_k' + 1}\gamma_k \quad (3.68)$$

Il est important de signaler que lorsque ξ_k' est grand, ce qui suggère que la parole est sûrement présente, $P\left(H_1^k/Y(w_k)\right) \approx 1$, comme prévu. En revanche, lorsque ξ_k' est

extrêmement petit, $\left(\mathrm{H}_1^{\mathrm{k}}/\mathrm{Y}(\mathrm{w}_{\mathrm{k}})\right) \approx 1 - \mathrm{q}_{\mathrm{k}}$, c'est-à-dire qu'elle est égale à la probabilité a priori de présence de la parole, $\mathrm{P}(\mathrm{H}_1^{\mathrm{k}})$.

L'estimateur MMSE final qui incorpore l'incertitude de présence de signal à la forme :

$$\begin{aligned}\hat{X}_k &= P\left(H_1^k/Y(w_k)\right)G(\xi_k,\gamma_k)|_{\xi_k=\xi_k'}Y_k \\ &= \frac{1-q_k}{1-q_k+q_k(1+\xi_k')exp(-v_k')}G(\xi_k',\gamma_k)Y_k \end{aligned} \quad (3.69)$$

3.9.2.6 Estimateur Incorporating speech presence probability in Log MMSE (Log MMSE-ISP)

Nous pouvons dériver l'estimateur log-MMSE qui prend en compte l'incertitude de présence du signal, suivant l'équation 3.58, nous avons :

$$log\hat{X}_k = E[logX_k/Y(w_k), H_1^k]P\left(H_1^k/Y(w_k)\right) \quad (3.70)$$

Et après avoir résolu pour $\widehat{\mathrm{X}}_{\mathrm{k}}$, on obtient :

$$\begin{aligned}\hat{X}_k &= e^{E[logX_k/Y(w_k),H_1^k]P\left(H_1^k/Y(w_k)\right)} \\ &= (e^{E[logX_k/Y(w_k),H_1^k]})^{P\left(H_1^k/Y(w_k)\right)}\end{aligned} \quad (3.71)$$

Le terme exponentiel entre parenthèses est l'estimateur log-MMSE et peut également être exprimé en utilisant l'équation 3.55 comme suit :

$$\hat{X}_k = [G_{LSA}(\xi_k, v_k)Y_k]^{P\left(H_1^k/Y(w_k)\right)} \quad (3.72)$$

Le terme de probabilité a posteriori $\mathrm{P}\left(\mathrm{H}_1^{\mathrm{k}}/\mathrm{Y}(\mathrm{w}_{\mathrm{k}})\right)$ n'est plus multiplicatif comme il l'était dans l'équation 3.69. Les résultats de la simulation [62] ont montré que l'estimateur précédent, équation 3.72 n'a entraîné aucune amélioration significative sur l'estimateur originel log-MMSE. Pour cette raison, l'estimateur modifié par multiplication suivante a été suggéré [69].

$$\hat{X}_k = [G_{LSA}(\xi_k', v_k')]P\left(H_1^k/Y(w_k)\right)Y_k \quad (3.73)$$

Où $P\left(H_1^k/Y(w_k)\right)$ est défini dans l'équation 3.67, et $G_{LSA}(\xi'_k, v'_k)$ est donné par l'équation 3.55 :

$$G_{LSA}(\xi'_k, v'_k) = \frac{\xi'_k}{\xi'_k + 1} exp\left\{\frac{1}{2}\int_{v'_k}^{\infty} \frac{e^{-t}}{t} dt\right\} \quad (3.74)$$

et ξ'_k, v'_k sont donnés respectivement par les équations 3.66 et 3.68.

L'approche précédente était sous-optimale car le terme de probabilité $P\left(H_1^k/Y(w_k)\right)$ a été forcé d'être multiplicatif. Un estimateur modifié de manière optimale a été proposé dans [70][71]. A partir du modèle de parole binaire originel donné dans Equation suivante [65] :

$$\hat{X}_k = E\left(X_k/Y_k, H_1^k\right)P\left(H_1^k/Y_k\right) + E\left(X_k/Y_k, H_0^k\right)P\left(H_0^k/Y_k\right) \quad (3.75)$$

$P\left(H_1^k/Y_k\right)$ désigne la probabilité conditionnelle, dont la parole présenté dans la fréquence discrète fk , étant donnée l'amplitude du signal de parole bruité Y_k, de même $P\left(H_0^k/Y_k\right)$ désigne la probabilité conditionnelle, où la parole est absente dans la fréquence fk, étant donnée l'amplitude du signal de parole bruité Y_k.

Nous avons :

$$\begin{aligned} log\hat{X}_k = {} & E[logX_k/Y(w_k), H_1^k]P(H_1^k/Y(w_k)) \\ & + E[logX_k/Y(w_k), H_0^k]P\left(H_0^k/Y(w_k)\right) \end{aligned} \quad (3.76)$$

Où $P\left(H_0^k/Y(w_k)\right) = 1 - P\left(H_1^k/Y(w_k)\right)$ désigne la probabilité a posteriori de l'absence de parole. Le deuxième terme $E\left[logX_k/Y(w_k), H_0^k\right]$ était précédemment supposé être nul sous l'hypothèse H_0^k. Si nous supposons maintenant que ce terme n'est pas nul mais très petit [70][72], alors nous obtenons :

$$\begin{aligned} \hat{X}_k &= e^{E[logX_k/Y(w_k), H_1^k]P\left(H_1^k/Y(w_k)\right)} e^{E[logX_k/Y(w_k), H_0^k]P\left(H_0^k/Y(w_k)\right)} \\ &= \left(e^{E[logX_k/Y(w_k), H_1^k]}\right)^{P\left(H_1^k/Y(w_k)\right)} \left(e^{E[logX_k/Y(w_k), H_0^k]}\right)^{P\left(H_0^k/Y(w_k)\right)} \end{aligned} \quad (3.77)$$

La première exponentielle entre parenthèses est l'estimateur log-MMSE d'origine et peut être exprimé comme $G_{LSA}(\xi_k, v_k)Y_k$ (voir l'équation 3.55), et la deuxième exponentielle entre parenthèses est supposée être petite et fixée à $G_{min}Y_k$, où G_{min} est une petite valeur.

L'estimateur précédent devient alors :

$$\hat{X}_k = [G_{LSA}(\xi_k, v_k)Y_k]^{P\left(H_1^k/Y(w_k)\right)}[G_{min}Y_k]^{P\left(H_0^k/Y(w_k)\right)}$$

$$= \left[G_{LSA}(\xi_k, v_k)^{P\left(H_1^k/Y(w_k)\right)} {G_{min}}^{1-P\left(H_1^k\middle|Y(w_k)\right)}\right] {Y_k}^{P\left(H_1^k\middle|Y(w_k)\right)} {Y_k}^{1-P\left(H_1^k\middle|Y(w_k)\right)}$$

$$= \left[G_{LSA}(\xi_k, v_k)^{P\left(H_1^k/Y(w_k)\right)} {G_{min}}^{1-P\left(H_1^k/Y(w_k)\right)}\right] Y_k$$

$$= G_{OLSA}(\xi_k, v_k)Y_k \quad (3.78)$$

La fonction de gain $G_{OLSA}(\xi_k, v_k)$ est multiplicative, l'estimateur logMMSE-ISP donne de meilleures performances en termes de SNR segmental par rapport à l'estimateur logMMSE, surtout à de faibles niveaux de SNR.

3.10 Techniques d'évaluation des méthodes d'amélioration du signal de parole

L'intelligibilité de la parole qui s'appuie sur l'idée de compréhension de l'auditeur [73]. L'intelligibilité peut être définie comme un aspect du signal de parole qui permet aux auditeurs de comprendre ce qu'un locuteur est en train de dire. L'intelligibilité est donc une mesure objective définie par le nombre de mots énoncés par le locuteur et qui sont correctement identifiés par l'auditeur [74].

3.10.1 Mesures subjectives

Les mesures subjectives sont les plus fiables pour évaluer la qualité perçue de la parole. Les tests d'écoute subjectives ont été conçus selon l'UIT-T recommandation P.835 et ont été menées par Dynastre [75][76]. La présente recommandation décrit une méthodologie d'évaluation subjective de la qualité de la parole en présence de bruit et qui permet tout particulièrement l'évaluation des algorithmes de suppression de bruit. Cette méthodologie utilise des échelles de notation distinctives pour la qualité du signal vocal seul, du bruit de fond seul et de la qualité globale :

- Le signal de parole seul utilisant une échelle de signal à cinq points de distorsion (SIG) ;

- Le bruit de fond seul en utilisant une échelle de cinq points d'intrusion de fond (BAK) ;
- L'effet global en utilisant l'échelle du score moyen d'opinion (OVRL) : [1= Mauvais, 2 = Médiocre, 3 = Passable, 4 = Bon, 5 = Excellent].

3.10.2 Mesures objectives

Les mesures objectives se présentent comme une alternative aux mesures subjectives et permettent d'automatiser l'évaluation de la qualité de la parole. Ces mesures peuvent être classées selon le domaine dans lequel, ils opèrent. En effet, il est possible de mesurer la ressemblance des signaux dans le domaine temporel, spectral ou perceptuel. Le point commun de tous ces critères est l'évaluation globale de la qualité du signal de parole par une seule quantité qui englobe tous les types de dégradation et donne une idée globale sur la qualité de la parole.

3.10.2.1 Mesures dans le domaine temporel

Les mesures du domaine temporel sont généralement applicables aux systèmes de codage analogique ou de forme d'onde dans lesquels la cible est de reproduire la forme d'onde. Le Rapport Signal / Bruit (SNR) et le SNR segmentaire (SNRseg) sont des mesures du domaine temporel [77][78] :

- SNR segmental (segSNR) : Le SNR est un paramètre métrique très utilisé dans la mesure objective, il est calculé par l'équation suivante :

$$SNR = 10log_{10}\frac{\sum_{n=1}^{N} x^2(n)}{\sum_{n=1}^{N}\big(x(n) - d(n)\big)^2} \tag{3.79}$$

Où x(n) représente le signal de parole d'origine, d(n) représente le signal débruité, et N le nombre d'échantillons.

Le SNR classique est un mauvais critère de mesure objective pour une large gamme de distorsion de la parole [77], parce que ce dernier quasi stationnaire est principalement traité dans des trames courtes, généralement avec environ 30 ms de longueur. En général le rapport SNRseg représente l'une des classes les plus populaires des mesures du domaine temporel. La mesure est définie comme une moyenne des valeurs de SNR de courts segments, et peut être calculé comme suit :

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1} log_{10}\left(\frac{\sum_{n=N_m}^{N_m+N-1} x(n)^2}{\sum_{n=N_m}^{N_m+N-1}[d(n)-x(n)]^2}\right) \quad (3.80)$$

M est le nombre total des trames et N est la longueur de la trame.

3.10.2.2 Mesures dans le domaine fréquentiel

Les mesures du domaine spectral sont plus crédibles que les mesures du domaine temporel car elles sont moins sensibles à l'apparition du désalignement temporels et de déphasage entre les signaux d'origine et les signaux déformés. Dans le domaine spectral, les mesures sont liées à la conception des codeurs vocaux et utilisent les paramètres des modèles de production vocale.

- **Log Likelihood Ratio (LLR) :** le rapport de vraisemblance logarithmique (LLR) est une mesure de la distance d'un signal amélioré en comparant les vecteurs de Codage de Prédiction Linéaire (LPC) du signal propre avec ceux du signal débruité [79].

Le LLR est calculé par la formule suivante :

$$LLR = log\left(\frac{a_d R_c a_d^T}{a_c R_c a_c^T}\right) \quad (3.81)$$

Où a_c est le vecteur LPC du signal propre, a_d est le vecteur LPC du signal débruité, et R_c est la matrice d'autocorrélation du signal propre.

- **Weighted Spectral Slope (WSS) :** la pente spectrale pondérée (WSS) [80] mesure les différences pondérées des pentes spectrales des différentes bandes de fréquences dans chaque trame de la distorsion du signal amélioré par rapport à celui du signal propre.
  La distance WSS peut être calculée par la formule suivante :

$$WSS\ Distance = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{k} W(j,m)\left(S_c(j,m)-S_d(j,m)\right)^2}{\sum_{j=1}^{k} W(j,m)} \quad (3.82)$$

Où K est le nombre de bandes de fréquence, M est le nombre total des trames, Sc est la pente spectrale du signal propre, Sd est la pente spectrale du signal bruité et W (j,m) sont des poids qui peuvent être calculés comme indiqué par Klatts [80].

- **Cepstral Distance (CD) :** la Distance Cepstrale fournit une estimation du log spectral distance entre deux spectres. Les coefficients cepstrales peuvent être obtenus à partir des coefficients LPC en utilisant l'équation suivante :

$$CD(\vec{c}_c, \vec{c}_p) = \frac{10}{log10} \sqrt{2.\sum_{k=1}^{p} [c_c(k) - c_p(k)]^2} \quad (3.83)$$

Où $\vec{c}_c$ et $\vec{c}_p$ sont respectivement les vecteurs de coefficients cepstrales de signal propre et débruité et p est l'ordre des coefficients LPC. La Distance Cepstrale est limitée dans l'intervalle de [0, 10].

3.10.2.3 Mesures dans le domaine perceptuel

Comme la plupart des mesures du domaine spectral utilisent les paramètres des modèles de production vocale utilisés dans les codecs, leurs performances sont généralement limitées par les contraintes de ces modèles. Contrairement aux mesures du domaine spectral, les mesures du domaine perceptuel sont basées sur des modèles de perception auditive humaine, et par conséquent, ont le meilleur potentiel de prédiction de la qualité subjective de la parole. Dans ces mesures, les signaux vocaux sont transformés en un domaine basé sur la perception en utilisant des concepts de psychophysique de l'audition, tels que la résolution spectrale de la bande critique, la sélectivité en fréquence pour obtenir une estimation du spectre auditif [81]. Les informations perceptuelles sont nécessaires et suffisantes pour une évaluation précise de la qualité de la parole.

- **Bark Spectral Distorsion (BSD) et Modified Bark Spectral Distorsion (MBSD)**

La mesure de BSD a été développée par Wang [82] comme une méthode de calcul d'une mesure objective de la distorsion du signal basée sur les propriétés quantifiables de la perception auditive. La mesure BSD globale représente la moyenne euclidienne entre le signal de parole de référence et celui codé, dans le domaine de Bark.

Le MBSD suppose que la qualité de la parole dépend directement de l'intensité du signal de parole [83], dont le seuil de masquage de bruit est incorporé pour calculer la distorsion dans le BSD, en tenant compte uniquement de la distorsion audible. Tout ce qui au-dessous du seuil de masquage de bruit, n'est pas perceptible à l'oreille humaine.

- **Evaluation Perceptive de la qualité de la parole (PESQ)**

Le PESQ est l'évaluation de la qualité vocale perçue, recommandée par la norme P.862 en 2001 [84]. Le PESQ a été développé à partir d'un grand nombre d'enregistrements contenant des phrases prononcées par une variété de locuteurs dans une variété de langues. Les enregistrements ont été réalisés à l'aide de plusieurs codeurs vocaux différents avec différents niveaux de qualité et avec des perturbations typiques de la transmission du réseau. Dans une série de tests d'écoute, un nombre adéquat d'auditeurs de test a classé ces exemples sur une échelle de qualité vocale allant de 1 (Médiocre) à 5 (Excellent).

Le PESQ est une méthode qui détermine une mesure objective qui permet de calculer la distance perceptuelle entre le signal vocal d'origine non dégradé (le signal de référence) avec le signal dégradé (signal mesuré). D'autres facteurs supplémentaires sont pris en considération pour mieux simuler les conditions réelles, à savoir le temps de propagation, les distorsions dues aux erreurs de transmission, les pertes de paquets, etc.

3.11 Conclusion

Dans ce chapitre, nous avons abordé quelques estimateurs de rehaussement de parole, telles que Maximum-Likelihood (ML), MMSE, log MMSE, maximum a posteriori (MAP), incorporating speech presence probability in MMSE (MMSE-ISP), incorporating speech presence probability in log MMSE (log MMSE-ISP), ainsi que l'estimateur Wiener. Ces méthodes seront comparées dans le dernier chapitre avec celle proposée dans le cadre de notre travail.

Finalement, nous avons donné les techniques d'évaluation des méthodes d'amélioration du signal parole, en ce qui concerne les mesures subjectives et objectives.

CHAPITRE 4 : APPLICATION DE L'ESTIMATEUR MMSE-MODGD AU SYSTEME FASR

4.1 Introduction

Cette étude propose une modification de l'estimateur MMSE, en remplaçant le spectre de l'amplitude estimé à l'aide d'une Transformée de Fourier (TF), par le spectre MODGD [85]. En d'autres termes, les variables aléatoires gaussiennes indépendantes sont dérivées du spectre MODGD, au lieu de leur estimation directe à partir de la Transformée de Fourier Discrète, pour améliorer l'algorithme MMSE en exploitant les informations contenues dans les spectres de phase.

Le conduit vocal d'un locuteur est un système à minimum de phase [86], ce qui permet d'extraire les informations du signal de parole à partir du spectre de phase ou d'amplitude. De plus, Parthasarathi et al [87], ont indiqué que le spectre de retard de groupe conserve la plupart des informations sur les formants, même à des SNR faibles du bruit ambiant. Le spectre MODGD est moins affecté par le bruit que le spectre d'amplitude.

Dans ce chapitre, nous avons expliqué l'estimateur MMSE-MODGD proposé pour améliorer le système FASR, et puis nous allons aborder les différentes techniques de traitement du signal vocal, le modèle universel (UBM), l'adaptation MAP, protocole expérimentale de l'amélioration de la parole et configuration du système FASR. Finalement, nous allons procéder à la mise en service dudit système dans les conditions propres, bruitées et dans le cas d'utilisation de l'estimateur MMSE et MMSE-MODGD. Les résultats et leurs interprétations sont discutés à la fin de chapitre.

Le rehaussement de la parole (débruitage) est un processus très complexe qui sert à éliminer le bruit d'un signal utile. Il consiste aussi à réduire une grande quantité du bruit, tout en préservant l'intégrité du signal. Souvent, les systèmes de traitement de la parole introduisent des dégradations et modification des signaux vocaux, par exemple, bruit de quantification dans un codeur de parole ou bruit résiduel et distorsion de la parole dans un système de réduction du bruit.

4.2 Techniques de traitement du signal vocal

Le signal de la parole est un signal complexe. Il contient une quantité importante d'informations. L'objectif de traitement du signal vocal est de fournir une représentation optimale moins redondante de la parole, tout en permettant une extraction précise des paramètres pertinents pour préserver au maximum l'information présente dans le signal d'origine.

Les principales classifications des méthodes de traitement du signal vocal sont :

- Les transformées classiques comme la Transformée de Fourier Discrète qui ne se réfère pas à un modèle de production ni de perception ;
- Les méthodes fondées sur la déconvolution « source-conduit vocal » qui s'appuient sur le modèle de production de la parole [88].

4.2.1 Méthodes non paramétriques

Le signal de la parole peut être analysé dans les domaines temporel ou spectral par des méthodes non paramétriques, sans faire appel à un modèle pour rendre compte du signal observé.

4.2.1.1 Processus de prétraitement

Le calcul de la représentation du signal est réalisé par un processus numérique selon les étapes suivantes :

- **Echantillonnage :** transforme le signal à temps continu s(t) en un signal à temps discret s(nTe) défini aux instants d'échantillonnage Te, en respectant le théorème de Shannon (la fréquence d'échantillonnage doit être supérieure ou égale à deux fois leur plus haute composante fréquentielle) ;
- **Préaccentuation :** le but de cette étape est d'augmenter la quantité d'énergie dans les hautes fréquences et d'avoir une compensation de filtrage des effets de l'acquisition du signal, car l'information pertinente du signal vocal se trouve d'une façon générale dans les régions de hautes fréquences, pour cela, on applique un filtre sur toutes les trames pour amplifier l'amplitude dans les hautes fréquences, sa transformée en Z est donnée par :

$$H(Z) = 1 - 0.95\, Z^{-1} \tag{4.1}$$

- **Elimination de silence :** le signal de parole contient des zones de silences, qui ne sont pas porteuses d'informations utiles et par conséquent affecteront les performances du système. D'où, il est nécessaire de localiser les segments de parole dépourvus de segments de silence. Pour effectuer cette tâche, nous avons utilisé l'algorithme VAD (Voice Activity Detection).
- **Fenêtrage :** à cause de la non stationnarité du signal vocal, nous utilisons une fenêtre glissante, chaque trame couvrant une durée de 20 à 30 ms sur laquelle le signal est supposé quasi-stationnaire. Le pas d'analyse entre deux trames successives est de l'ordre de quelques dizaines de ms. Le découpage du signal vocal en trames produit

des discontinuités aux frontières des trames, qui se manifestent par des lobes secondaires dans le spectre, pour remédier à ces déformations, nous multiplions chaque tranche d'analyse par une fenêtre de pondération appelée fenêtre de Hamming [89].

4.2.1.2 Analyse temporelle

Dans l'analyse temporelle du signal vocal, nous pouvons extraire des paramètres tels que l'énergie et la fréquence fondamentale :

- **Energie :** l'énergie E_0 est calculée directement dans le domaine temporel par trame de signal S_n, $0 \leq n \leq N-1$ par :

$$E_0 = \sum_{n=0}^{N-1} S_n^2 \tag{4.2}$$

- **Fréquence fondamentale :** La fréquence fondamentale est également appelée F_0. Elle représente le nombre de vibrations par seconde des cordes vocales. La F_0 n'est calculée que sur des parties voisées de la parole, c'est-à-dire principalement les voyelles, les semi-voyelles et aussi les consonnes voisées.

4.2.1.3 Analyse spectrale

Parmi les techniques utilisées, la FFT (Fast Fourier Transform) exprime la répartition fréquentielle de l'amplitude, de la phase et de l'énergie des signaux.

Soit s(t) un signal déterministe. Sa transformée de Fourier est une fonction, généralement complexe, de la variable f et définie par :

$$S(f) = TF[s(t)] = \int_{-\infty}^{+\infty} s(t) e^{-2\pi j f t}\, dt \tag{4.3}$$

4.2.2 Méthodes paramétriques

Les méthodes paramétriques sont fondées sur une connaissance des mécanismes de production de la parole. Les plus utilisées sont celles basées sur :

- L'analyse prédictive Linéaire ;
- L'analyse cepstrale.

4.2.2.1 Codage Prédictif Linéaire (LPC)

Le codage prédictif linéaire (LPC) est une technique d'analyse utilisée comme un modèle estimateur des paramètres de la parole dans la synthèse, le codage et la reconnaissance de

parole. Elle s'appuie sur l'idée que le système phonatoire peut être modélisé par un filtre linéaire. Ce filtre est excité par un train d'impulsions pour les sons voisés et un bruit blanc pour les sons non voisés. Il s'agit donc à prédire le signal à un instant *n* à partir des *p* échantillons

précédents :

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{4.4}$$

Où :

p est l'ordre de filtre ;

a_k sont les coefficients de prédiction linéaire ;

G est le gain de l'excitation et u(n) est le signal d'excitation.

$$H(z) = \frac{S(z)}{G.U(z)} = \frac{1}{1 - \sum_{k=1}^{p} Z^{-k} a_k} = \frac{1}{A(z)} \tag{4.5}$$

Notons que H(z) ne contient alors que des pôles et c'est pour cette raison que ce modèle est aussi appelé modèle tous pôles.

Ces coefficients sont calculés par la minimisation de l'erreur quadratique moyenne entre le signal original *s(n)* et le signal prédit $\tilde{s}(n)$ sur une fenêtre donnée, selon l'équation suivante :

$$\sum_n e^2(n) = \sum_n (s(n) - \tilde{s}(n))^2 = \sum_n \left(s(n) - \sum_{k=1}^{p} a_k s(n-k) \right)^2 \tag{4.6}$$

4.2.2.2 Cepstre

La parole est une combinaison entre une excitation des poumons et des interventions des différents organes vocaux qui caractérisent le locuteur et le différencie des autres. Ce qui est exprimé par une convolution :

$$s(t) = e(t) * h(t) \tag{4.7}$$

avec ;

$s(t)$: signal de parole ;
$e(t)$: signal d'excitation (source) ;
$h(t)$: la réponse impulsionnelle du conduit vocal.

En passant au domaine fréquentiel, la convolution devient une multiplication, lorsqu'on applique la Transformée de Fourier Rapide (FFT) sur les deux membres de l'équation 4.3 :

$$S(f) = E(f).H(f) \tag{4.8}$$

On applique le log sur les deux membres de l'équation 4.19 :

$$Log\big(S(f)\big) = Log\big(E(f)\big) + Log(H(f)) \tag{4.9}$$

En appliquant la Transformée de Fourier Rapide Inverse (IFFT), les coefficients cepstraux sont donnés comme suit :

$$c(n) = IFFT(\log S(f)) \tag{4.10}$$

La dimension du nouveau domaine est homogène à un temps et s'appelle l'espace quéfrentiel. Il est possible, par un filtrage temporel (liftrage), de séparer la contribution de la source (la fréquence fondamentale) de celle du conduit vocal (les formants) dans le signal de parole [90].

Pour estimer la contribution du conduit vocal dans le signal de parole, nous ne conservons que les premiers échantillons du cepstre c(n) qui correspondent en particulier aux informations sur les formants.

En ce qui concerne, les échantillons du cepstre d'ordre plus élevé correspondent, en général, aux caractéristiques de la fréquence fondamentale des cordes vocales.

- Coefficients MFCC (Mel Frequency Cepstral Coefficients)

Les coefficients cepstraux les plus répandus sont les MFCC (Mel Frequency Cepstral Coefficients). Ils présentent l'avantage d'être faiblement corrélés entre eux, et qu'on peut donc approximer leur matrice de covariance par une matrice diagonale. Pour simuler le fonctionnement du système auditif humain, les fréquences centrales du banc de filtres sont réparties uniformément sur une échelle perceptive. Plus la fréquence centrale d'un filtre est élevée, plus sa bande passante est large. Cela permet d'augmenter la résolution dans les basses fréquences, zone qui contient le plus d'informations utiles dans le signal de parole. Nous avons utilisé une échelle perceptive la plus utilisée :

Echelle Mel : Plusieurs études ont montré que la perception humaine des sons ne suit pas une échelle linéaire [88], en d'autre terme la sélectivité de l'oreille humaine diminue avec l'accroissement des fréquences, d'où l'idée de définir pour chaque valeur de fréquence f, une hauteur subjective qui est mesurée sur une échelle « Mel » (en Mels).

L'échelle Mel est logarithmique en hautes fréquences et donnée par l'Equation suivante :

$$f_{mel} = 2595.\log(1 + \frac{f_{Hz}}{700}) \tag{4.11}$$

Le spectre Mel est simulé en utilisant un banc de K filtres triangulaires positionné uniformément sur l'échelle Mel (Figure 4.1).

Le spectre Mel est simulé en utilisant un banc de K filtres triangulaires positionné uniformément sur l'échelle Mel (Figure 4.1).

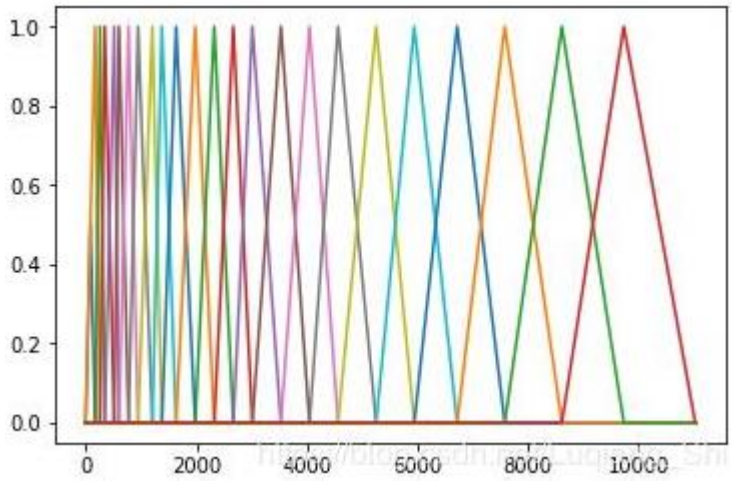


Figure 4.1: Banc de filtres utilisé dans le calcul des MFCC

Les coefficients MFCC d'une trame de parole sont calculés suivant les étapes suivantes (Figure 4.2) :

- Après la phase de fenêtrage et préaccentuation, on applique la FFT sur chaque trame d'une façon uniforme le long du signal de parole, Ensuite on calcule l'énergie ;
- L'énergie est passée à travers chaque filtre Mel. Soit Sk l'énergie du signal à la sortie du filtre k, nous avons maintenant mp (le nombre de filtres) ;
- Le logarithme de Sk est calculé ;
- Finalement, les coefficients sont calculés en utilisant la iDCT (inverse Discret Cosine Transform), donc, nous revenons vers le temporel et on obtient une matrice de covariance digonale.

$$C_i = racine\ (2/m_p \left\{ \sum_{k=1}^{m_p} \log(S_k) \cos[\frac{i\left(k-\frac{1}{2}\right)\pi}{m_p}] \right\}_{pour\ i=1\ldots N} \tag{4.12}$$

Où N est le nombre de coefficients MFCC.

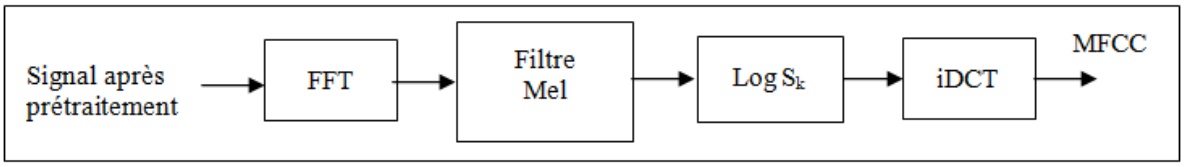


Figure 4.2 : Calcul des coefficients MFCC

4.3 Modèle de Mélange de Gaussiennes

La reconnaissance du locuteur par mélange de gaussiennes (GMM) consiste à modéliser le vecteur acoustique d'un locuteur suspect par une somme pondérée de composantes gaussiennes [91].

Un mélange de gaussiennes est une somme pondérée de M densité gaussiennes. Soit un vecteur acoustique x de dimension D, le mélange de gaussiennes est défini comme suit :

$$p(x|\lambda) = \sum_{m=1}^{M} \pi_m \, b_m(x) \tag{4.13}$$

Où $b_m(x)$ représente les densités de probabilités gaussiennes paramétrées par le vecteur moyenne μ_m et une matrice de covariance Σ_m, et π_m représente le poids des mélanges avec :

$$\sum_{m=1}^{M} \pi_m = 1 \tag{4.14}$$

$$b_m(x) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_m|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(x-\mu_m)(\Sigma_m)^{-1}(x-\mu_m)\right] \tag{4.15}$$

4.3.1 Modèle universel (UBM)

Le modèle universel ou du monde (UBM) est un grand GMM entrainé par une grande quantité de la base de données de la population potentielle (P) utilisant l'algorithme Expectation Maximisation (EM), afin de construire un modèle statistique de cette population.

4.3.1.1 Estimation du modèle UBM par l'algorithme (EM)

L'algorithme (EM) est une méthode itérative pour trouver un maximum local. Cette technique s'applique à tous les modèles à mélange gaussienne. En utilisant un ensemble des données à modéliser qui sont les vecteurs acoustiques. Ladite méthode maximise la fonction de vraisemblance d'une façon itérative et le modèle sera changé d'une itération à une autre.

4.3.2 Modèle du suspect

Généralement, dans le domaine criminalistique, les enquêteurs et les experts forensiques ne disposent pas suffisamment les données d'apprentissage pour estimer le modèle du locuteur suspect. Pour remédier à ce problème, On peut démarrer du modèle universel qui représente les paramètres acoustiques de la population potentielle, par l'adaptation du modèle (UBM) en utilisant les données d'apprentissage du locuteur suspect, ainsi que la méthode maximum à posteriori (MAP). L'adaptation MAP se fait en deux étapes : la première étape est identique à l'algorithme (EM). Ici les paramètres statistiques sont calculés à partir des données du suspect pour chaque mélange de l'UBM. Dans l'étape suivante, les nouveaux paramètres acoustiques sont combinés avec les anciens paramètres à partir des paramètres de mélange UBM.

Pour pallier au manque de données d'apprentissage dans l'estimation GMM, l'adaptation MAP du modèle universel est utilisée. L'adaptation d'un modèle GMM consiste à modifier les paramètres initiaux du modèle vers un autre modèle plus spécifique, par rapport aux données d'adaptation [92]. Etant donné un corpus d'apprentissage avec une séquence des vecteurs acoustiques $X = \{x_1, x_3, \dots, x_N\}$ appliqués sur les vecteurs moyens μ_i du modèle UBM pour obtenir les vecteurs moyens adaptés $\hat{\mu}_i$.

$$\hat{\mu}_i = \alpha_i E_i(X) + (1 - \alpha_i)\mu_i, \quad i = 1, \dots, M \tag{4.16}$$

Avec $E_i(X)$la moyenne i du GMM client et est un coefficient de pondération qui permet d'affecter plus ou moins de poids aux paramètres estimés sur les données d'apprentissage.

$$\alpha_i = \frac{n_i(X)}{n_i(X)+r} \tag{4.17}$$

n_i est le nombre de trames associées à la Gaussienne i défini par l'équation suivante :

$$n_i(X) = \sum_{t=1}^{N} P(i/x_t) \tag{4.18}$$

$$E_i(X) = \frac{1}{n_i}\sum_{t=1}^{N} P(i/x_t)x_t \tag{4.19}$$

$$P(i/x_t) = \frac{\lambda_i P_i(x_t)}{\sum_{j=1}^{M} \lambda_i P_j(x_t)} \tag{4.20}$$

Où λ_i et $p_i(x)$ sont respectivement le poids du mélange et la densité de probabilité de i^{ime} mélange parmi *M* mélange de l'UBM, et r est un facteur de contrôle de degré d'adaptation.

4.4 Fonction MODGD proposée pour l'estimateur MMSE du spectre à court terme du signal de parole bruité

Un signal de parole ne peut être représenté complètement dans le domaine spectral que si les informations d'amplitude et de phase sont spécifiées. Cependant, les informations extraites du spectre de phase sont plus complexes que les informations extraites du spectre d'amplitude, car le spectre de phase est généralement discontinu ou enveloppé entre $[-\pi, \pi]$ [93][94]. Une fonction à valeurs multiples est utilisée pour en faire une fonction continue ; c'est ce qu'on appelle la phase déroulée (déroulage) [87]. La fonction retard de groupe est principalement utilisée pour extraire les informations contenues dans le spectre de phase.

Soit $x(n)$ un signal de parole, sa transformée de Fourier est donnée par l'équation 3.2.
La fonction de retard de groupe $\tau(\omega)$ d'un signal $x(n)$ est définie comme la dérivée négative du spectre de phase comme suit :

$$\tau_X(\omega) = -\frac{d\theta(\omega)}{d\omega} \tag{4.21}$$

La fonction de retard de groupe peut également être estimée à partir du signal vocal en utilisant l'équation suivante :

$$\tau_X(\omega) = \frac{X_R(\omega)\hat{X}_R(\omega) + X_I(\omega)\hat{X}_I(\omega)}{|X(\omega)|^2} \tag{4.22}$$

Où, *R* et *I* désignent respectivement la partie réelle et la partie imaginaire, $x(n) \leftrightarrow X(\omega)$ et $\hat{x}(n) \leftrightarrow \hat{X}(\omega)$ sont les paires de Transformée de Fourier, et $\hat{x}(n) = nx(n)$.

La fonction de retard de groupe nécessite que le signal de parole soit un système à minimum de phase ou que les pôles de la fonction de transfert soient dans le cercle unitaire [87][95].

Pour un signal à minimum de phase, nous pouvons montrer que le log amplitude et les spectres de phase continue sont liés comme suit :

$$ln|X(w)| = \frac{1}{2}c(0) + \sum_{n=1}^{\infty} c(n)\cos(nw) \tag{4.23}$$

$$\theta(w) = -\sum_{n=1}^{\infty} c(n)\sin(nw) \tag{4.24}$$

Où *c(n)* sont les coefficients cepstraux. Prendre la dérivée négative de l'équation 4.24, nous obtenons la fonction de retard de groupe pour les signaux de phase minimum :

$$\tau(w) = \sum_{n=1}^{\infty} nc(n)\cos(nw) \tag{4.25}$$

Les équations 4.23 et 4.24 montrent que pour les signaux à minimum de phase, l'amplitude logarithmique et la phase sont liés par des coefficients cepstraux. De plus, à partir de l'équation 4.25, nous trouvons que la fonction de retard de groupe est la Transformée de Fourier du cepstre pondéré.

En lissant le spectre d'amplitude $X(\omega)$ [87] dans l'équation 4.22, pour supprimer les zéros près du cercle d'unité. Finalement, nous définissons une fonction de retard de groupe modifiée MODGD qui est donnée comme suit :

$$\tau_X(\omega) = \left(\frac{\tau_s(\omega)}{|\tau_s(\omega)|}\right)(|\tau_s(\omega)|^{\alpha}) \tag{4.26}$$

où,

$$\tau_s(\omega) = \frac{X_R(\omega)\hat{X}_R(\omega) + X_I(\omega)\hat{X}_I(\omega)}{|S(\omega)|^{2\gamma}} \tag{4.27}$$

et $|S(\omega)|$ est une version lissée de $|X(\omega)|$, α et γ sont deux paramètres introduits pour contrôler la plage dynamique de MODGDF, sachant que, $0 < \alpha \leq 1$ et $0 < \gamma \leq 1$. La longueur de la fenêtre de lissage cepstral est contrôlée par le paramètre lifter_{ω}.

Les paramètres MODGD sont calculés, en utilisant le Discrete Cosine Transform (DCT) :

$$\begin{aligned}\hat{X}_k^2 &= E\{\tau_{Xk}^2/\tau_Y(w_k)\} \\ &= \int_0^{\infty} \tau_{Xk}^2 f_{\tau_{Xk}}(\tau_{Xk}/\tau_{Yk}(w_k))d\tau_{Xk} \\ &= \frac{\xi_k}{1+\xi_k}\left(\frac{1}{\gamma_k} + \frac{\xi_k}{1+\xi_k}\right)\tau_{Yk}^2\end{aligned} \tag{4.29}$$

où,

$$\xi_k = \frac{\sigma_x^2(k)}{\sigma_d^2(k)}, \quad \gamma_k = \frac{\tau_{Yk}^2}{\sigma_d^2(k)} \tag{4.30}$$

$$\sigma_x^2(k) = E\{\tau_{Xk}^2\}, \quad \sigma_d^2(k) = E\{\tau_{Dk}^2\} \tag{4.31}$$

Enfin, la densité de probabilité a posteriori $f_{X_k}(X_k/Y(w_k))$ devient :

$$f_{\tau_{Xk}}(\tau_{Xk}/\tau_{YK}(w_k)) = \frac{\tau_{Xk}}{\sigma_k^2}\exp\left(-\frac{\tau_{Xk}^2 + s_k^2}{2\sigma_k^2}\right) I_0\left(\frac{\tau_{Xk}s_k}{\sigma_k^2}\right) \tag{4.32}$$

De plus, les équations 3. 20, 3.21 et 3.22 peuvent être écrites comme suit :

$$\tau_{Yk}^2 \approx \tau_{Xk}^2 + \tau_{Dk}^2 \tag{4.33}$$

$$f_{\tau_{Xk}^2}(\tau_{Xk}^2) = \frac{1}{\sigma_x^2(k)} e^{-\frac{\tau_{Xk}^2}{\sigma_x^2(k)}} \tag{4.34}$$

$$f_{\tau_{Dk}^2}(\tau_{Dk}^2) = \frac{1}{\sigma_d^2(k)} e^{-\frac{\tau_{Dk}^2}{\sigma_d^2(k)}} \tag{4.35}$$

Où $\sigma_x^2(k)$ et $\sigma_d^2(k)$ sont donnés par l'équation 4.31. La densité de probabilité a posteriori du spectre de l'amplitude au carré de la parole propre, devient ainsi :

$$f_{\tau_{Xk}^2}(\tau_{Xk}^2/\tau_{Yk}^2) = \frac{f_{\tau_{Yk}^2}(\tau_{Yk}^2/\tau_{Xk}^2) f_{\tau_{Xk}^2}(\tau_{Xk}^2)}{f_{\tau_{Yk}^2}(\tau_{Yk}^2)}$$

$$= \begin{cases} \psi_k e^{-\frac{\tau_{Xk}^2}{\lambda(k)}}, & \text{si } \sigma_x^2(k) \neq \sigma_d^2(k) \\ \frac{1}{\tau_{Yk}^2}, & \text{si } \sigma_x^2(k) = \sigma_d^2(k) \end{cases} \tag{4.36}$$

Où $\lambda(k)$ est donné par les équations 3.24 et 4.31.

et

$$\psi_k = \frac{1}{\lambda(k)\left\{1 - exp\left[-\frac{\tau_{Yk}^2}{\lambda(k)}\right]\right\}} \tag{4.37}$$

Enfin, l'estimateur MMSE modifié est donné par :

$$\hat{X}_k^2 = E\{\tau_{Xk}^2/\tau_{Yk}^2\}$$
$$= \int_0^{\tau_{Yk}^2} \tau_{Xk}^2 f_{\tau_{Xk}^2}(\tau_{Xk}^2/\tau_{Yk}^2) d\tau_{Xk}^2$$
$$= \begin{cases} \left(\frac{1}{\upsilon_k} - \frac{1}{e^{\upsilon_k} - 1}\right)\tau_{Yk}^2, & \text{si } \sigma_x^2(k) \neq \sigma_d^2(k) \\ \frac{1}{2}\tau_{Yk}^2, & \text{si } \sigma_x^2(k) = \sigma_d^2(k) \end{cases} \tag{4.38}$$

υ_k est donné par l'équation 3.27, et, γ_k est donné par l'équation 4.30.

4.5 Protocole expérimental pour l'amélioration de la parole

Des tests de qualité objectifs approfondis ont été réalisés pour évaluer les performances de la méthode d'estimation MMSE-MODGD proposée en utilisant dix (10) phrases extraites de la Base de Données NOIZEUS [96]. Dans cette base de données, les signaux de bruit sont générés en ajoutant le bruit des bases de données AURORA et NOISEX-92 aux signaux propres, à un SNR global de 0 dB, 5 dB et 10 dB. La taille de trame choisie est de 20 ms avec un recouvrement de 50 %. Une fréquence d'échantillonnage de 8 kHz et une fenêtre de Hamming ont été utilisées. Les méthodes utilisées pour la comparaison avec le MMSE-MODGD proposé sont les estimateurs suivants : Maximum de vraisemblance (ML), MMSE, log MMSE, maximum a posteriori (MAP), incorporation de la probabilité de présence de la parole dans MMSE (MMSE-ISP), incorporation de la probabilité de présence de la parole dans log MMSE (log MMSE-ISP) et Wiener [97].

L'évaluation objective a été réalisée comme proposé dans [98]. Les tests effectués pour évaluer la méthode proposée comprennent des mesures liées à la perception du signal de parole sur une échelle de distorsion du signal (SIG) à cinq points (1-5), le bruit de fond sur une échelle de cinq points (1-5) (BAK) et la qualité globale (OVRL) basée sur le score moyen d'opinion (MOS) allant de 1 à 5. Les autres mesures utilisées sont le SNR segmentaire (SegSNR), le spectre de la pente pondérée (WSS), l'Evaluation Perceptuelle de la Qualité de la Parole (PESQ) et le rapport de vraisemblance logarithmique (LLR) [98].

4.5.1 Résultats et discussion

Sur la base d'une étude comparative utilisant des spectrogrammes, nous pouvons remarquer que la méthode MMSE-MODGD proposée donne de bons résultats par rapport aux autres méthodes à savoir : ML, MMSE, log-MMSE, MAP, MMSE-ISP, log MMSE-ISP et Wiener. Cette bonne performance obtenue par l'approche proposée est confirmée par l'évaluation objective.

Les Tableaux 1, 2 et 3 présentent les résultats des évaluations utilisant les mesures objectives : SIG, BAK, OVRL, PESQ, SegSNR, WSS et LLR en utilisant 10 phrases extraites de la base de données NOIZEUS. La méthode MMSE-MODGD proposée est comparée à ML, MMSE, logMMSE, MAP, MMSE-ISP, log MMSE-ISP et Wiener, dans un contexte de dégradation par un bruit blanc, d'usine et de bavardage, respectivement. Les scores LLR et WSS indiquent une perte de parole et doivent donc être minimes. Les résultats présentés dans les tableaux montrent clairement que les scores SIG, BAK et OVRL, qui reflètent le niveau de perception du signal de parole et la qualité globale, sont généralement plus élevés pour la méthode MMSE-MODGD que pour les autres méthodes. Les résultats montrent également

que ces évaluations confirment que l'amélioration de la parole basée sur la méthode MMSE-MODGD produit un SNR segmentaire plus élevé, un PESQ plus élevé et un WSS plus faible que les autres méthodes.

Tableau 4.1 : Evaluation objectives de la technique MMSE-MODGD par rapport à ML, MMSE, Log-MMSE, MAP, MMSE-ISP, Log-MMSE-ISP et Wiener, dont le corpus de test est corrompu par le bruit blanc. Les valeurs moyennes ont été obtenues en utilisant 10 phrases extraites de la base de données NOIZEUS. Les meilleures performances sont indiquées en gras.

Mesures objectives	**Entrée SNR dB**	**Bruit blanc**							
		ML	**MMSE**	**Log-MMSE**	**MAP**	**MMSE-ISP**	**Log-MMSE-ISP**	**Wiener**	**MMSE-MODGD**
SIG	0	1.31	1.29	1.09	1.19	1.19	0.91	0.93	**1.36**
[1 to5]	5	1.83	1.90	1.61	1.68	1.68	1.27	1.30	**1.95**
	10	2.43	**2.46**	2.12	2.19	2.20	1.74	1.84	2.38
BAK	0	1.60	1.59	1.55	1.62	1.64	1.57	1.58	**1.71**
[1 to 5]	5	1.94	1.99	1.84	1.85	1.86	1.70	1.72	**2.03**
	10	2.33	2.36	2.17	2.16	2.17	1.98	2.01	**2.39**
OVRL	0	**1.30**	1.29	1.16	1.24	1.25	1.06	1.08	**1.30**
[1 to 5]	5	1.72	1.75	1.59	1.63	1.62	1.29	1.33	**1.78**
	10	2.22	2.29	2.09	2.06	2.05	1.72	1.80	**2.32**
PESQ	0	1.54	1.70	1.57	1.68	1.62	1.63	1.68	**1.72**
	5	1.81	2.08	1.82	1.94	1.76	1.71	1.87	**2.10**
	10	2.16	**2.40**	2.14	2.23	2.08	1.98	2.21	**2.40**
SegSNR	0	-4.18	-1.95	-1.95	-3.03	-1.46	**-1.13**	**-1.13**	-1.32
	5	-1.79	0.35	-1.26	-1.61	-0.73	-0.34	-0.34	**1.23**
	10	0.78	2.27	-0.43	0.70	-0.01	0.34	0.34	**4.03**
WSS	0	72.54	104.33	108.63	72.30	109.70	105.64	108.12	**71.18**
	5	63.80	94.14	104.25	64.21	102.32	98.81	100.74	**59.00**
	10	54.20	80.55	93.20	55.23	91.86	89.76	85.98	**53.12**
LLR	0	1.99	1.83	2.10	1.90	2.12	2.12	2.13	**1.81**
	5	1.72	1.55	1.86	1.60	1.85	1.89	1.87	**1.45**
	10	1.42	1.31	1.71	1.29	1.69	1.72	1.62	**1.24**

Tableau 4.2 : Evaluation objectives de la technique MMSE-MODGD par rapport à ML, MMSE, Log-MMSE, MAP, MMSE-ISP, Log-MMSE-ISP et Wiener, dont le corpus de test est corrompu par le bruit d'usine. Les valeurs moyennes ont été obtenues en utilisant 10 phrases extraites de la base de données NOIZEUS. Les meilleures performances sont indiquées en gras.

Mesures objectives	**Entrée SNR dB**	**Bruit usine**							
		ML	**MMSE**	**Log-MMSE**	**MAP**	**MMSE-ISP**	**Log-MMSE-ISP**	**Wiener**	**MMSE-MODGD**
SIG	0	1.41	1.33	1.22	1.25	1.23	1.01	1.13	**1.58**
[1 to5]	5	1.95	1.93	1.60	1.72	1.80	1.34	1.39	**2.05**
	10	2.49	2.55	2.18	2.29	**2.64**	1.89	2.04	2.47
BAK	0	1.65	1.67	1.62	1.77	1.44	1.30	1.72	**1.82**
[1 to 5]	5	2.04	2.11	1.92	1.95	1.96	1.74	1.54	**2.13**
	10	2.45	2.47	2.37	2.46	2.16	2.08	2.21	**2.56**
OVRL	0	1.82	1.49	1.15	1.22	1.28	1.17	1.22	**1.87**
[1 to 5]	5	1.83	1.50	1.19	1.33	1.31	1.34	1.33	**1.92**
	10	2.12	2.20	2.07	2.00	2.01	1.55	1.88	**2.34**
PESQ	0	1.04	1.07	1.76	1.83	1.72	1.74	1.98	**2.02**
	5	1.81	2.01	1.80	1.93	1.77	1.81	2.07	**2.29**
	10	2.36	2.45	2.04	2.32	2.17	1.99	2.47	**2.54**
SegSNR	0	-3.13	-1.73	-1.82	-3.13	-1.70	-1.10	**-1.03**	-1.77
	5	-0.99	1.05	-1.06	-1.33	-0.22	0.04	-0.19	**1.28**
	10	1.02	2.50	0.88	1.09	0.30	1.50	0.95	**3.73**
WSS	0	72.03	98.33	102.63	67.30	103.10	100.60	102.02	**66.11**
	5	53.70	84.12	94.52	54.22	98.12	88.80	100.01	**50.99**
	10	45.20	78.52	90.21	52.22	81.87	83.73	83.93	**44.14**
LLR	0	1.88	1.72	2.05	1.87	2.02	2.09	2.17	**1.49**
	5	1.39	1.67	**1.12**	1.57	1.66	1.98	1.77	1.33
	10	1.30	**1.01**	1.05	1.19	1.59	1.53	1.44	1.11

Tableau 4.3 : Evaluation objectives de la technique MMSE-MODGD par rapport à ML, MMSE, Log-MMSE, MAP, MMSE-ISP, Log-MMSE-ISP et Wiener, dont le corpus de test est corrompu par le bruit bavardage. Les valeurs moyennes ont été obtenues en utilisant 10 phrases extraites de la base de données NOIZEUS. Les meilleures performances sont indiquées en gras.

Mesures objectives	**Entrée SNR dB**	**Bruit bavardage**							
		ML	**MMSE**	**Log-MMSE**	**MAP**	**MMSE-ISP**	**Log-MMSE-ISP**	**Wiener**	**MMSE-MODGD**
SIG	0	2.33	1.88	1.92	2.29	2.22	1.84	1.95	**2.67**
[1 to5]	5	2.74	2.73	2.27	2.59	2.51	2.04	2.25	**3.09**
	10	**3.52**	2.93	2.77	3.03	3.06	2.63	2.81	3.28
BAK	0	1.77	1.48	1.55	1.73	1.71	1.58	1.62	**1.84**
[1 to 5]	5	2.07	1.85	1.84	1.99	1.99	1.82	1.89	**2.19**
	10	**2.42**	2.25	2.16	2.28	2.29	2.09	2.14	**2.42**
OVRL	0	2.17	1.52	1.60	1.90	1.84	1.54	1.64	1.93
[1 to 5]	5	2.30	1.99	1.95	2.18	2.13	1.79	1.95	**2.54**
	10	2.76	2.48	2.39	2.57	2.59	2.27	2.40	**2.89**
PESQ	0	1.85	1.55	1.79	**1.87**	1.69	1.64	1.83	**1.87**
	5	2.12	1.95	1.98	2.13	1.93	1.85	2.12	**2.15**
	10	2.34	2.28	2.30	2.38	2.28	2.16	2.38	**2.48**
SegSNR	0	-4.19	-3.09	-2.01	-3.46	-1.67	**-1.42**	**-1.42**	-2.59
	5	-1.72	-1.10	-1.49	-1.30	-1.02	-0.57	-0.57	**-0.23**
	10	0.93	0.67	-1.00	0.82	-0.39	0.04	0.04	**2.81**
WSS	0	69.77	100.07	107.02	65.20	108.35	105.06	104.81	**58.80**
	5	**50.10**	92.07	92.55	58.87	95.87	95.17	87.61	64.38
	10	51.12	73.35	84.21	49.12	83.88	83.96	73.73	**41.42**
LLR	0	0.98	1.21	1.20	0.99	1.25	1.26	1.14	**0.96**
	5	0.80	1.33	1.23	0.85	1.28	1.29	1.07	**0.78**
	10	0.69	0.85	1.06	0.63	0.98	1.03	0.79	**0.60**

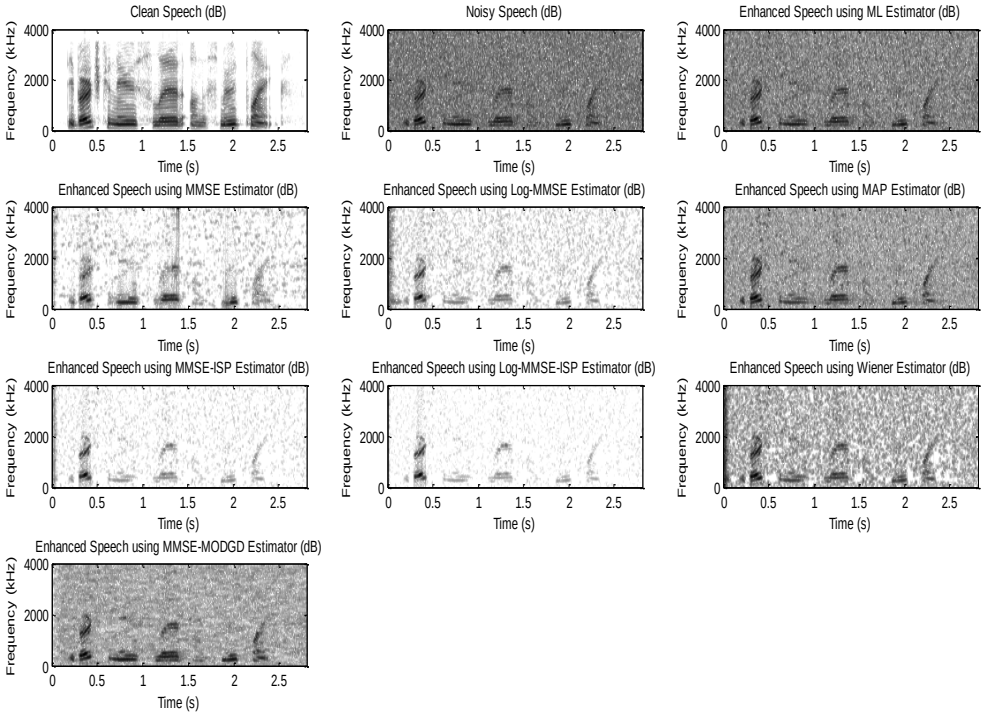


Figure 4.3 : spectrogrammes de la parole propre, parole corrompue par le bruit blanc, avec le SNR = 0 dB et méthodes d'amélioration de la parole

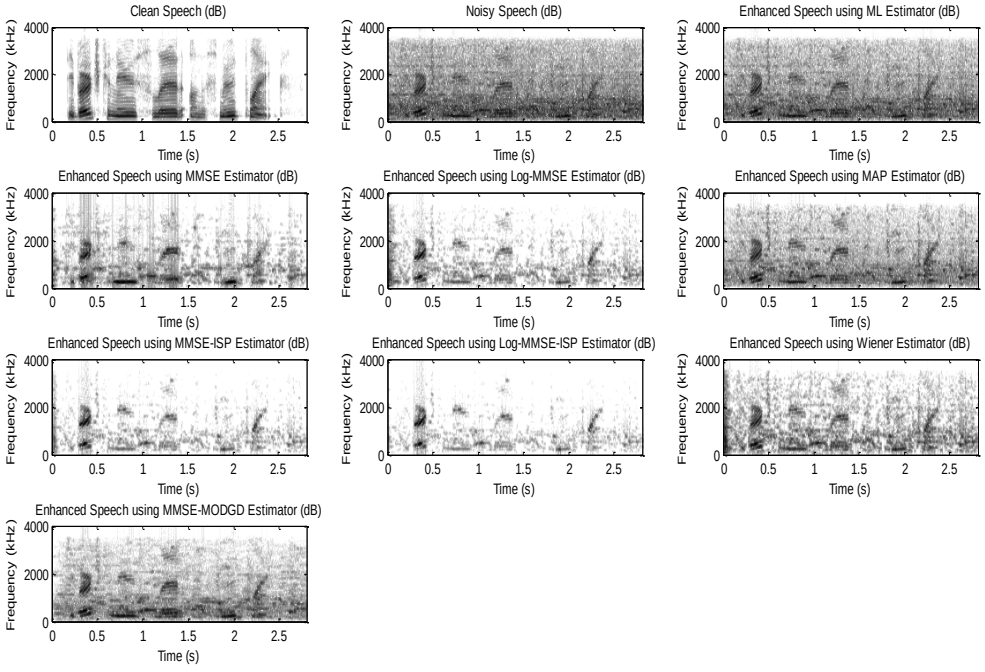


Figure 4.4 : spectrogrammes de la parole propre, parole corrompue par le bruit d'usine, avec le SNR = 0 dB et méthodes d'amélioration de la parole

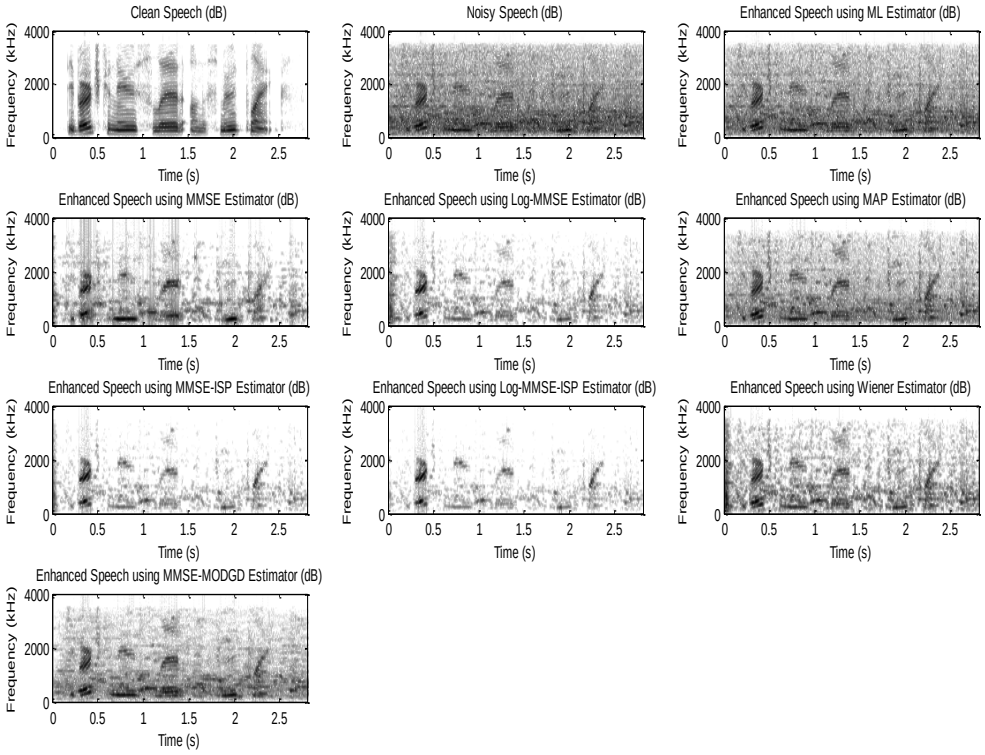


Figure 4.5 : spectrogrammes de la parole propre, parole corrompue par le bruit de bavardage, avec le SNR = 0 dB et méthodes d'amélioration de la parole

4.6 Protocole expérimental pour la configuration du système FASR

Généralement, il existe deux contraintes dans les scénarios FASR. Le premier est la non-collaboration des suspects et le second est le nombre limité de suspects connus de la personne ciblée (personne qui souffre des agissements des autres). En raison de ces contraintes, le nombre de suspects utilisés pour développer de tels systèmes (FASR) est vraiment limité.

Dans ce travail, toutes les expériences ont été réalisées sur le corpus NIST 2000, qui consiste en la parole téléphonique spontanée échantillonnée à 8 kHz. Pour l'extraction de caractéristiques, un vecteur 23 MFCC est trouvé à partir de la parole pré-accentuée en utilisant une fenêtre de Hamming de 20 ms, avec un recouvrement de 50%.

Vingt (20) locuteurs ont été choisis comme suspects dans ce corpus, la Base de Données de Référence du locuteur suspect (R) contient un (1) enregistrement d'une durée de 2 min pour chaque suspect, avec 75 % de la durée de cet enregistrement est destinée à la modélisation et 25 % aux tests (traces).

Le segment de test est divisé en 4 sections, pour avoir 4 traces pour chaque suspect. La Base de Données Potentielle (P) utilisée est un ensemble de 420 locuteurs du même corpus

cité ci-dessus. Le GMM-UBM se composait de 256 composants de mélange gaussien conçu via l'algorithme Expectation Maximisation (EM) en utilisant 10 itérations [91].

Vingt modèles suspects ont été créés grâce au GMM-UBM en utilisant une adaptation Maximum A Posteriori (MAP) avec un facteur de pertinence r=16, 256 mélanges gaussiens et une quantité de données d'adaptation de 14 heures est utilisée [99].

Selon le Figure 2.9, qui explique l'approche méthodologique FASR adoptée dans notre travail, nous avons besoin de 3 bases de données :

- Base de données potentielle (base de données UBM) : contient 420 locuteurs (420*2 min = 14 heures) ;
- Trace-database (T) : contient 20 locuteurs, chaque locuteur a 4 traces de (0,25*2min)/4= 7,5 s. Ainsi, le total des vrais essais (H0) est de 20*4 = 80 et le total des faux essais (H1) est de 4*20*20-80 (vrais essais) = 1520 ;
- Base de données de données de référence : contient 20 locuteurs, chaque locuteur a 0,75*2 min = 1,5 min.

Les mesures de performance ont fourni une valeur numérique unique décrivant les performances en termes de précision, de pouvoir discriminant et d'étalonnage de la méthode LR (Probabilités de preuves trompeuses, PMEH0 et PMEH1), Equal Proportion Probability (EPP) [42][100]. Les valeurs utilisées pour les fonctions MODGD sont la longueur de la fenêtre de lissage cepstral lifterω = 8, $\alpha = 0.4$ et $\gamma = 0.9$.

L'algorithme appliqué dans notre travail pour élaborer le système FASR est donné selon l'organigramme illustré dans la Figure 4.6.

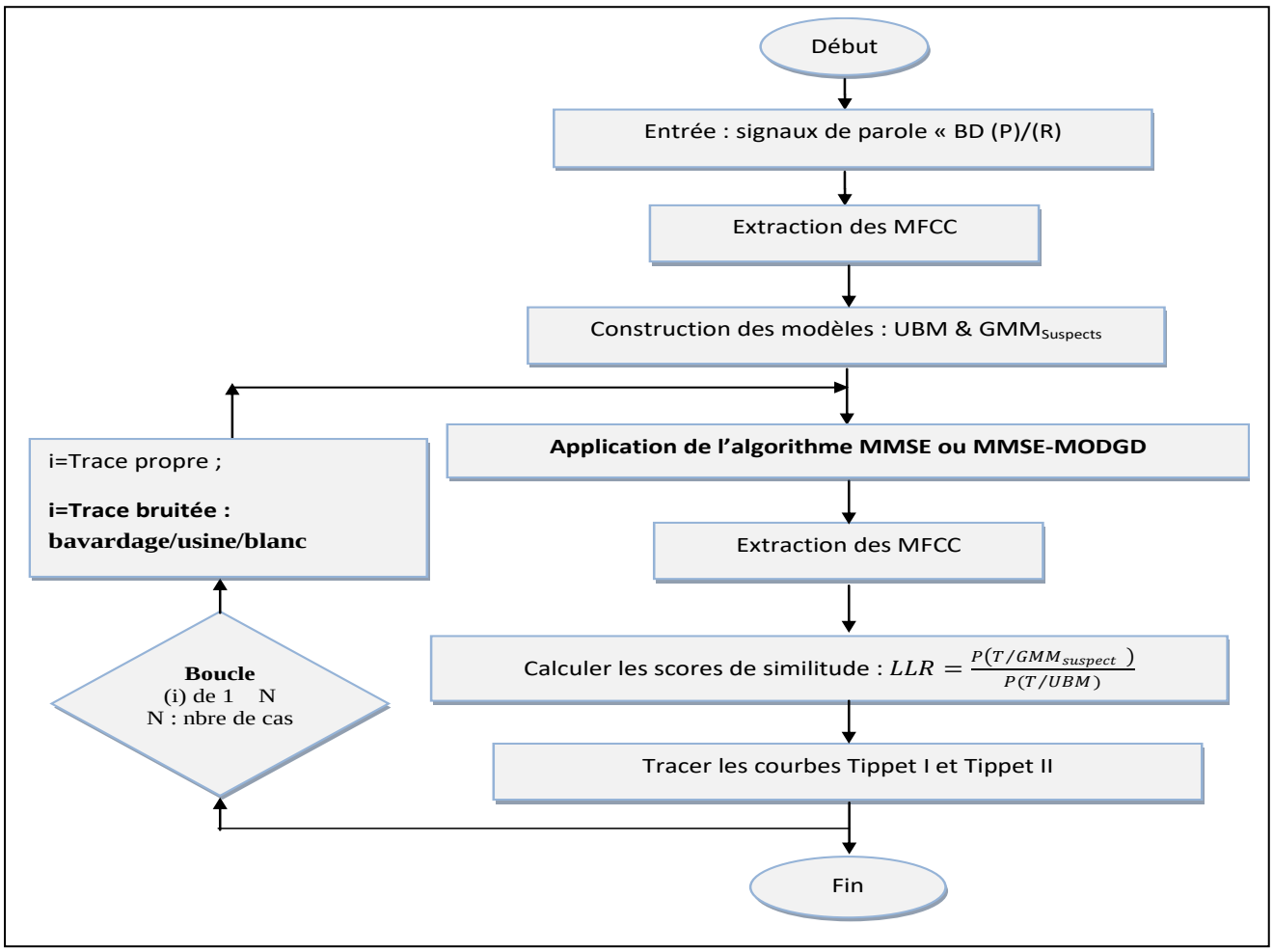


Figure 4.6 : Organigramme simplifié du système FASR

4.7 Résultats du système FASR classique

Cette section évalue les résultats obtenus dans des environnements propres et bruités.

4.7.1 Performances FASR dans les conditions propres

Une évaluation du FASR basée sur les performances du GMM-UBM en termes d'EPP, $PMEH_0$ et $PMEH_1$ a été réalisée dans un environnement propre.

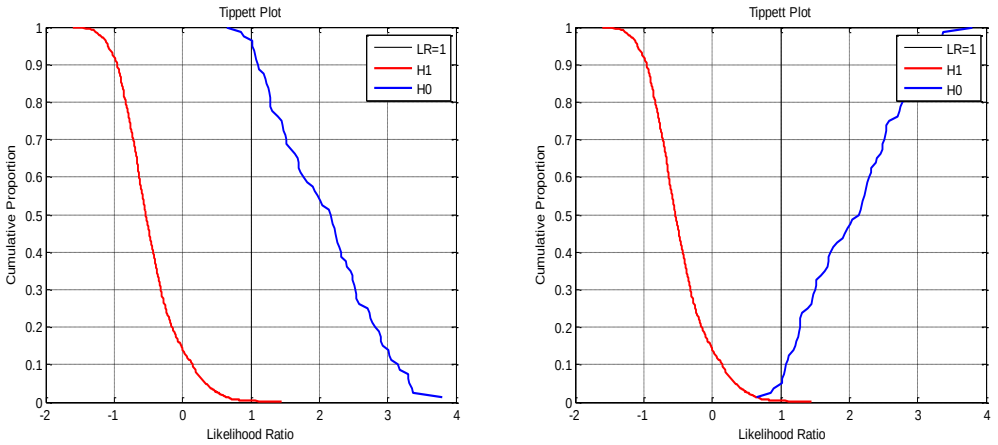


Figure 4.7 : Courbes de Tippett du FASR en conditions propres

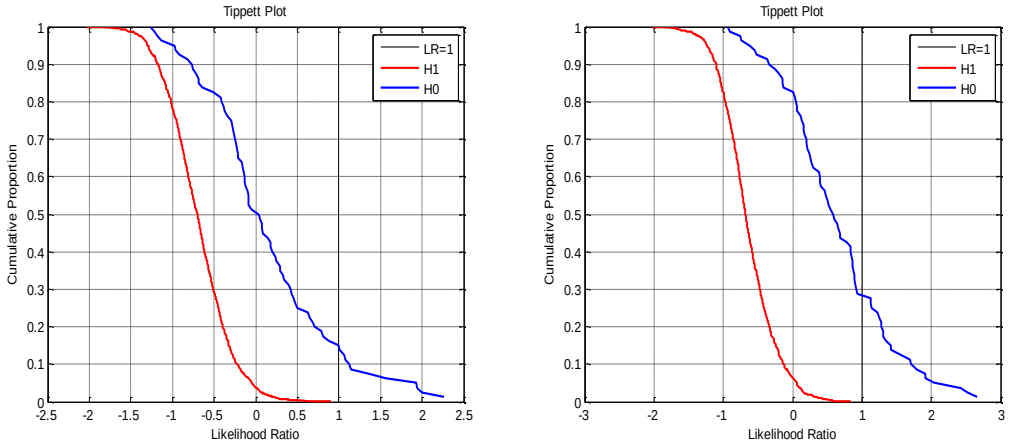
Tableau 4.4 : Evaluation des résultats obtenus dans les conditions propres propres

EPP (%)	H_0 vraie (%)		H_1 vraie (%)	
	LR<1	LR>1	LR<1	LR>1
1.25	4	96	99.60	0.40

D'après la Figure 4.7 et le Tableau 4.4, les résultats sont très satisfaisants, en termes d'EPP, $PMEH_0$ et $PMEH_1$. Par conséquent, EPP = 1,25 %, le LR dépasse 1 dans 96 % des cas lorsque H_0 est vrai et dans seulement 0,4 % des cas lorsque H_1 est vrai.

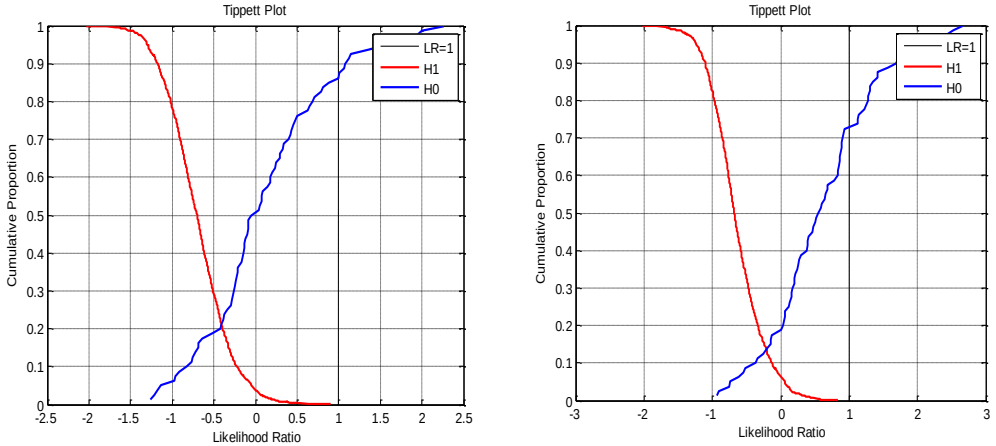
4.7.2 Performances FASR dans les conditions bruitées

Différents bruits ont été choisis arbitrairement dans ce travail (bavardage, usine et blanc) qui ont été ajoutés au corpus de test (traces) pour produire des vecteurs de caractéristiques bruités. Le Tableau 4.5 et les Figures 4.8 au 4.13, présentent les performances du FASR, avec, (a) : SNR = 0 dB et (b) : SNR = 5 dB.



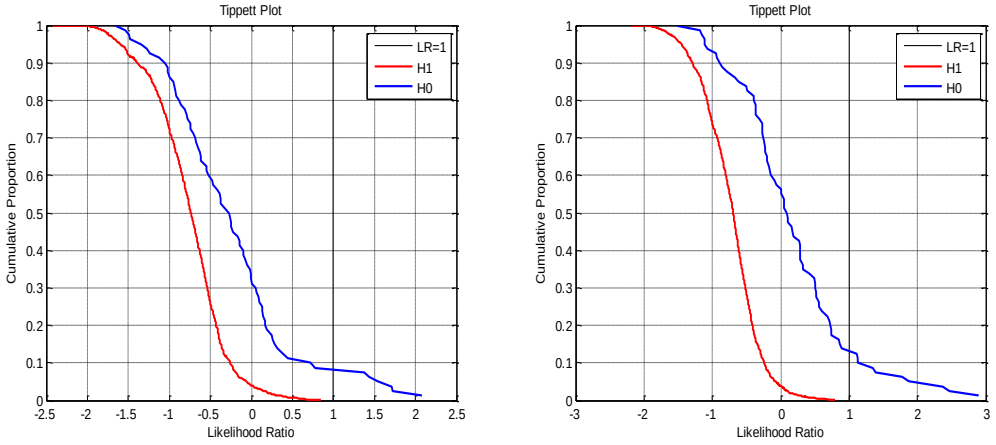
(a) (b)

Figure 4.8 : Courbes de Tippett de type I de FASR dans le bruit de bavardage



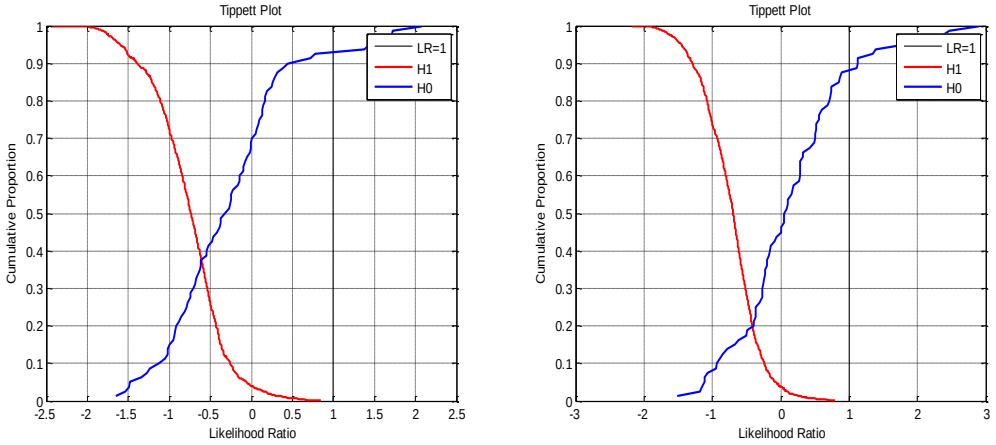
(a) (b)

Figure 4.9 : Courbes de Tippett de type II de FASR dans le bruit de bavardage



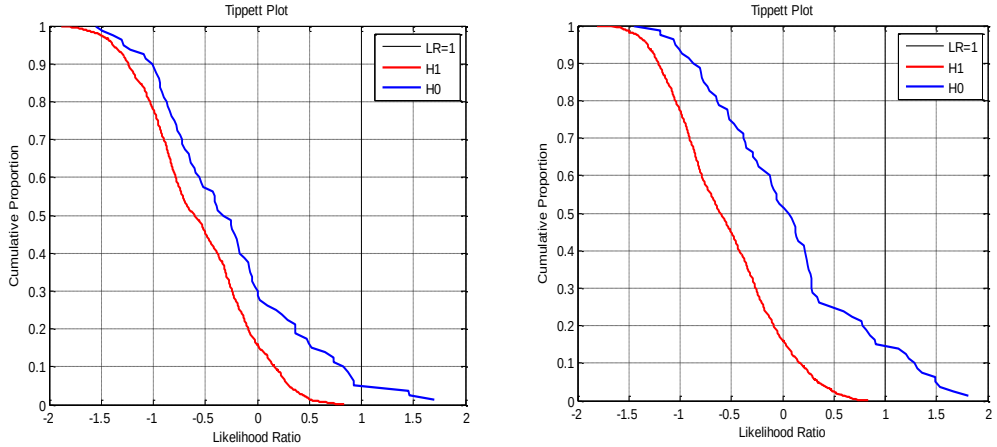
(a) (b)

Figure 4.10 : Courbes de Tippett de type I de FASR dans le bruit de l'usine



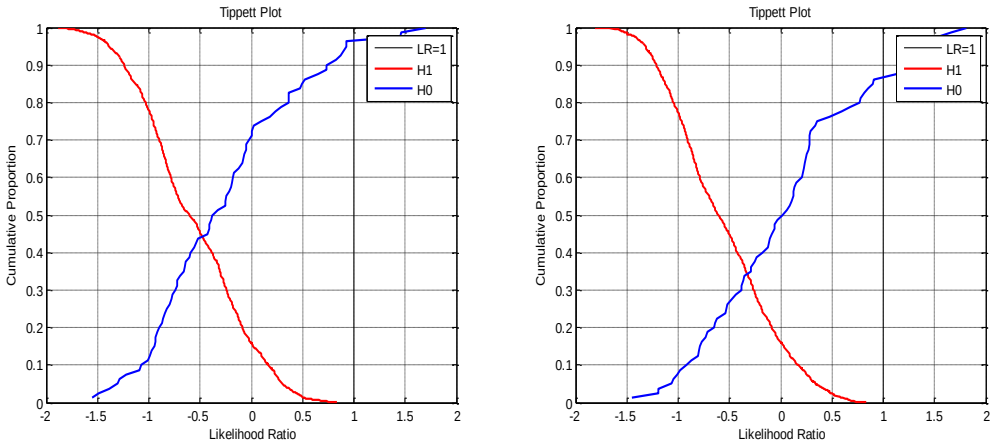
(a) (b)

Figure 4.11 : Courbes de Tippett de type II de FASR dans le bruit de l'usine



(a) (b)

Figure 4.12 : Courbes de Tippett de type I de FASR dans le bruit blanc



(a) (b)

Figure 4.13 : Courbes de Tippett de type II de FASR dans le bruit blanc

Table 4.5 : Evaluation des résultats obtenus dans les conditions bruitées

Type de bruit	SNR (dB)	EPP (%)	H_0 vraie (%)		H_1 vraie (%)	
			LR<1	LR>1	LR<1	LR>1
Bavardage	0	20.00	85	15	100	0
	5	13.75	71	29	100	0
Usine	0	37.50	91	09	100	0
	5	20.00	86	14	100	0
Blanc	0	43.75	95	05	100	0
	5	33.75	85	15	100	0

Le Tableau 4.5 résume les performances du FASR en conditions bruités, en termes d'EPP, $PMEH_0$ et $PMEH_1$. On peut remarquer que les performances du FASR diminuent avec la diminution du SNR et augmentent avec des valeurs des SNR élevées, et les performances de la parole corrompue par le bruit de bavardage sont moins dégradées par rapport aux autres bruits. Cela peut s'expliquer par le fait que le bruit de bavardage est un chevauchement de plusieurs sons provenant de deux ou plusieurs locuteurs [101]. Ses traits sont comme ceux de la voix. Il ne couvre que le spectre des basses fréquences. Par conséquents, seules les informations dans les régions à basse fréquence sont affectées par ce bruit. Alors que les bruits d'usine et blanc sont caractérisés par une intensité élevée. Ils couvrent le spectre des basses et hautes fréquences et ils affectent toutes les informations existantes dans le signal vocal. Les performances sont pires lors de l'utilisation de ces deux types de bruits (usine et blanc) que ceux obtenus sous le bruit de bavardage.

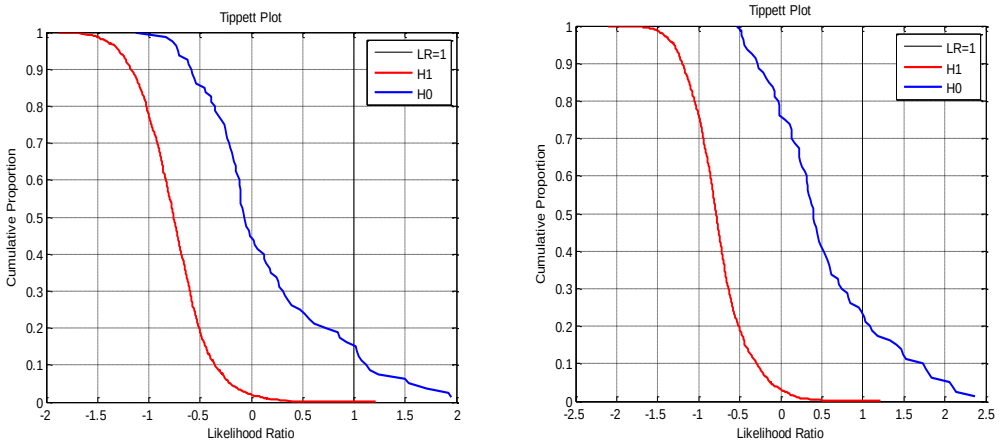
4.8 Résultats du système FASR amélioré

Dans cette section, les performances de ce système sont calculées en utilisant deux méthodes d'amélioration de parole : MMSE et Notre approche MMSE-MODGD.

4.8.1 Performances FASR utilisant l'estimateur MMSE

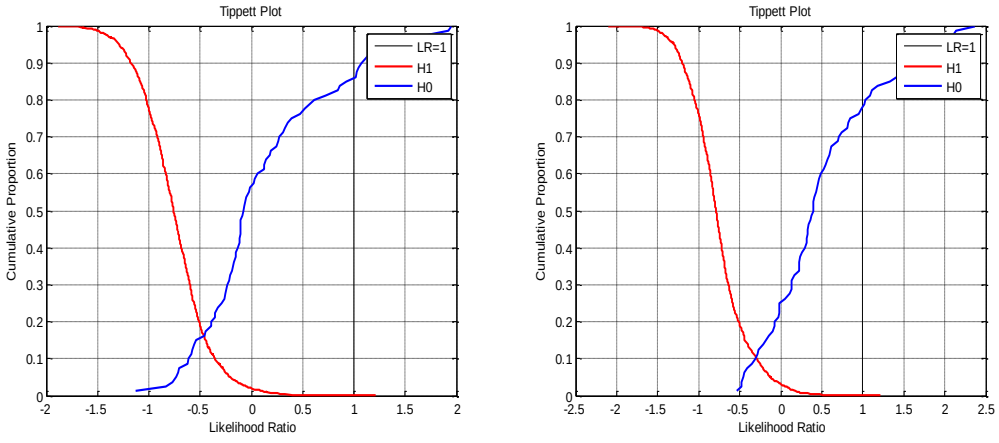
Le Tableau 4.6 et les Figures de 4.14 au 4.19 indiquent les résultats obtenus lors de l'utilisation de l'algorithme d'amélioration de l'amplitude MMSE (uniquement les informations contenues dans l'amplitude), avec, (a) : SNR = 0 dB et (b) : SNR = 5 dB.

Les résultats présentés dans le Tableau 4.6, lors de l'application de l'amélioration de la parole MMSE sur des tests bruités (traces) de parole, indiquent une amélioration des performances représentées par la diminution de l'EPP avec l'évolution de LR.



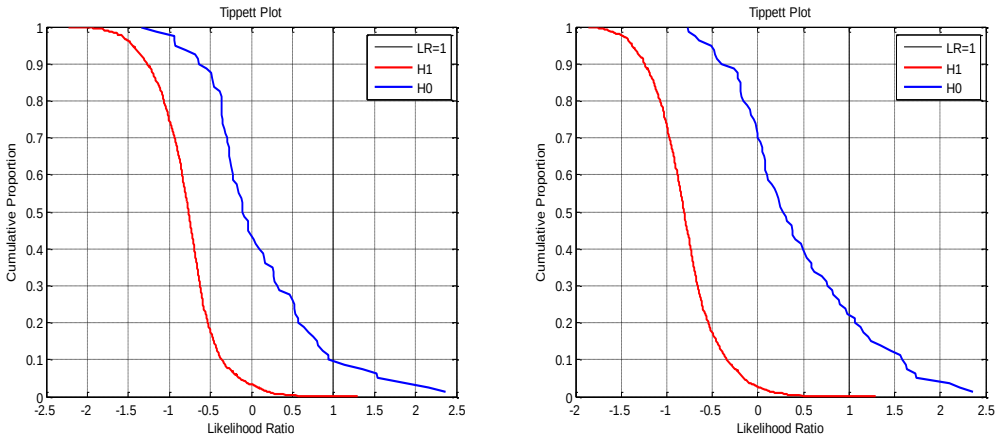
(a) (b)

Figure 4.14 : Courbes de Tippett de type I de FASR dans le bruit de bavardage



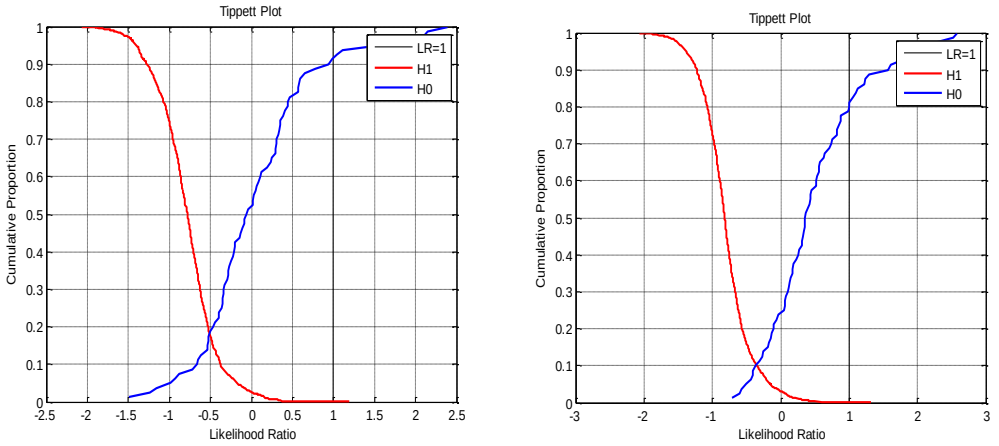
(a) (b)

Figure 4.15 : Courbes de Tippett de type II de FASR dans le bruit de bavardage



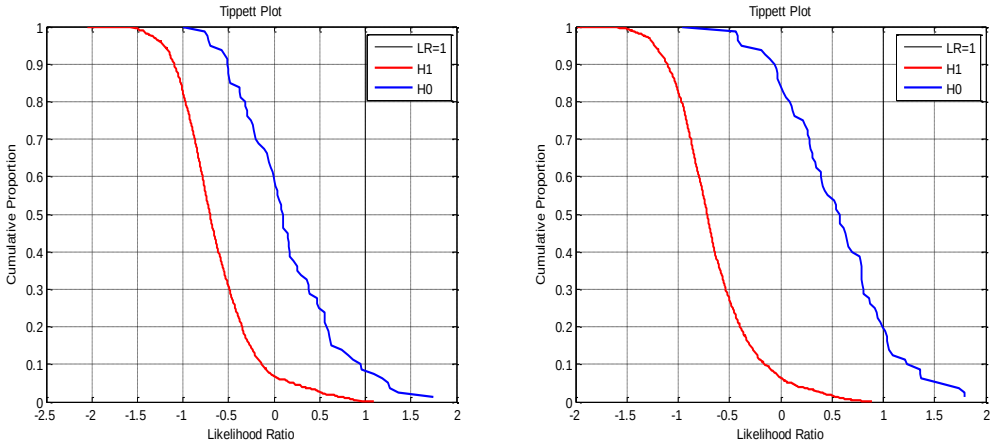
(a) (b)

Figure 4.16 : Courbes de Tippett de type I de FASR dans le bruit de l'usine



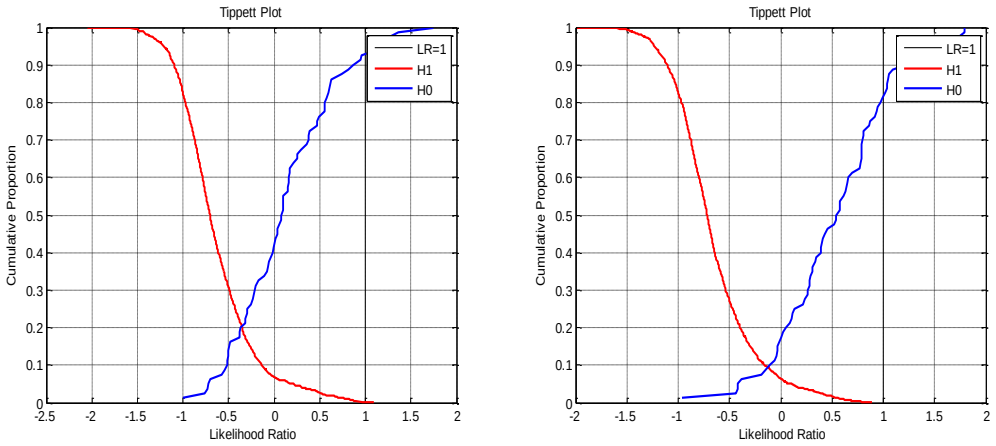
(a) (b)

Figure 4.17 : Courbes de Tippett de type II de FASR dans le bruit de l'usine



(a) (b)

Figure 4.18 : Courbes de Tippett de type I de FASR dans le bruit de blanc



(a) (b)

Figure 4.19 : Courbes de Tippett de type II de FASR dans le bruit blanc

Table 4.6 : Evaluation des résultats obtenus avec l'estimateur d'amplitude MMSE

Type de bruit	SNR (dB)	EPP (%)	H_0 vraie (%)		H_1 vraie (%)	
			LR<1	LR>1	LR<1	LR>1
Bavardage	0	15.39	85.00	15.00	100	0
	5	10.00	76.25	23.75	100	0
Usine	0	17.82	90.00	10.00	100	0
	5	11.25	77.50	22.50	100	0
Blanc	0	20.00	91.25	08.75	100	0
	5	08.81	80.00	20.00	100	0

Cette amélioration s'explique par le fait que l'estimateur de spectre de l'amplitude basé sur le MMSE élimine tout le bruit à large bande en éliminant la plus part des pics larges qui constituent les variances indésirables des ordonnées du spectre [97].

De plus, l'estimateur de spectre de l'amplitude basé sur le MMSE fournit la fonction de densité de probabilité a postériori (PDF) du signal propre étant donné le signal bruité. Ce PDF est un estimateur optimal pour une grande classe de différentes mesures de distorsions entre un signal propre et un signal bruité. Cette mesure de distorsion attribue une distorsion nulle pour les estimations dans le voisinage immédiat du signal propre. Par conséquent, la séparation entre les composants de bruit et de parole est meilleure [1][65]. Cet estimateur a également produit une faible distorsion de parole.

Bien que l'approche FFT basée sur le MMSE, s'est avérée offrir de meilleures performances dans l'amélioration de la parole et les résultats FASR, mais un certain bruit musical est toujours perceptible dans la parole améliorée dans le cas de faibles valeurs de SNR. Donc, le bruit déforme l'enveloppe spectrale FFT, modifie sa pente et ne conserve pas les emplacements des pics de formants supérieurs.

4.8.2 Performances FASR utilisant l'estimateur proposé MMSE-MODGD

Le Tableau 4.7 et les Figures 4.20 au 4.25 montrent les résultats obtenus lors de l'utilisation de l'algorithme d'amélioration de la parole proposé (MMSE-MODGD), en tenant compte les informations contenues dans l'amplitude et la phase, à SNR = 0 dB et SNR = 5 dB.

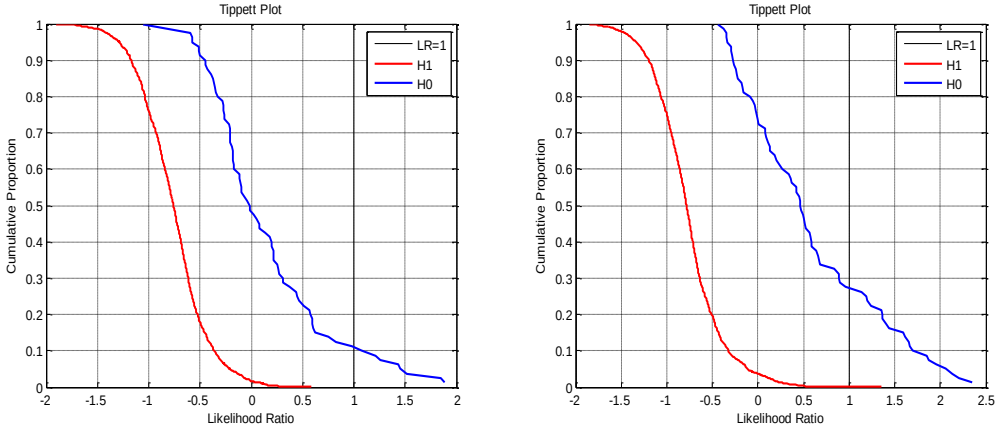


Figure 4.20 : Courbes de Tippett de type I de FASR dans le bruit de bavardage

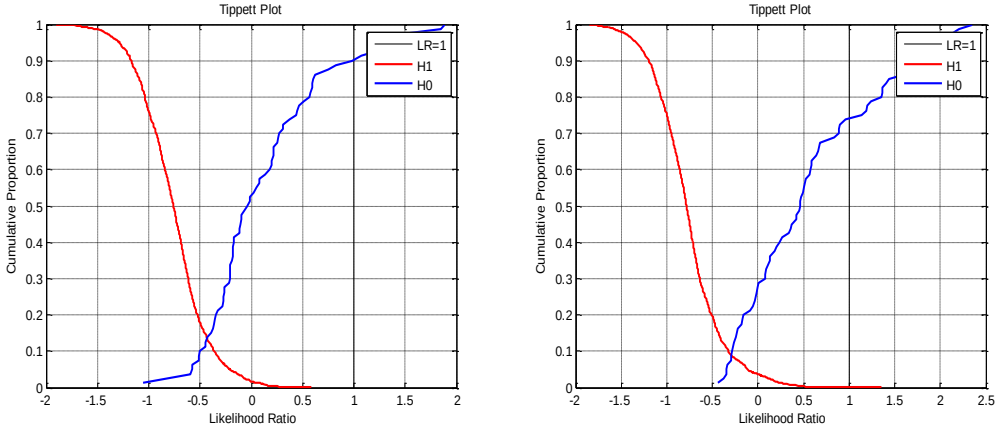


Figure 4.21 : Courbes de Tippett de type II de FASR dans le bruit de bavardage

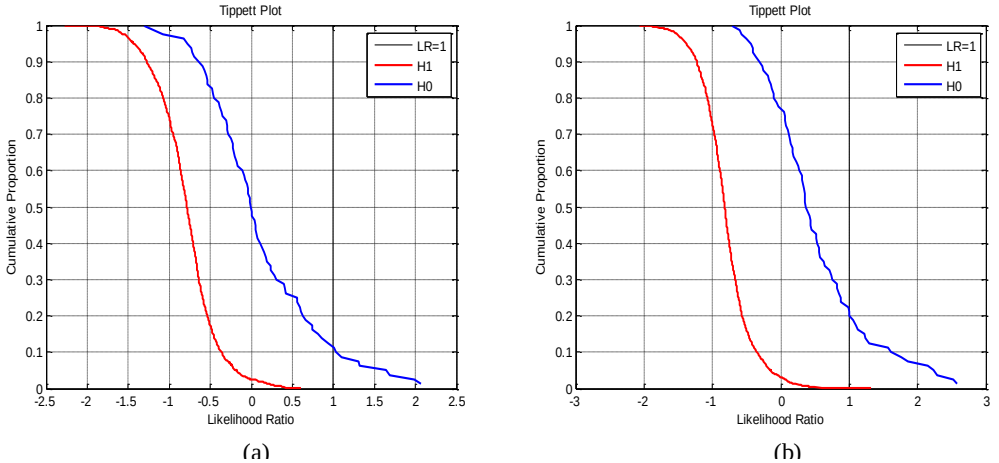
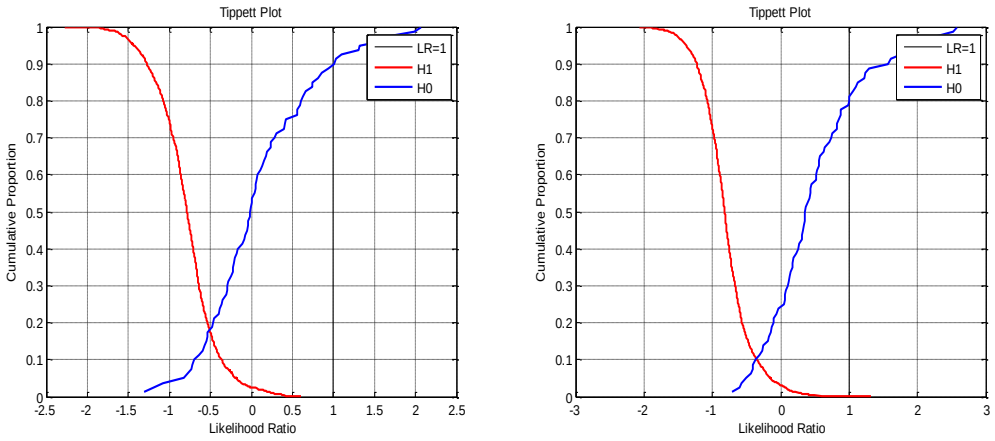
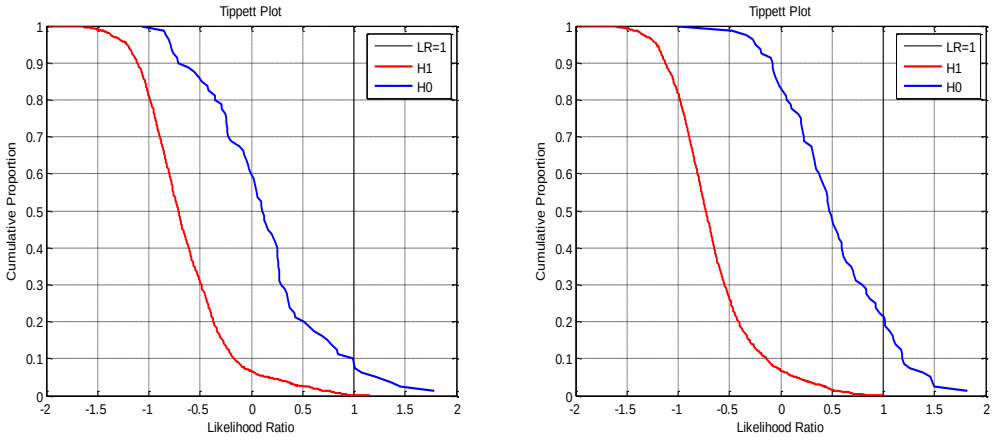


Figure 4.22 : Courbes de Tippett de type I de FASR dans le bruit de l'usine



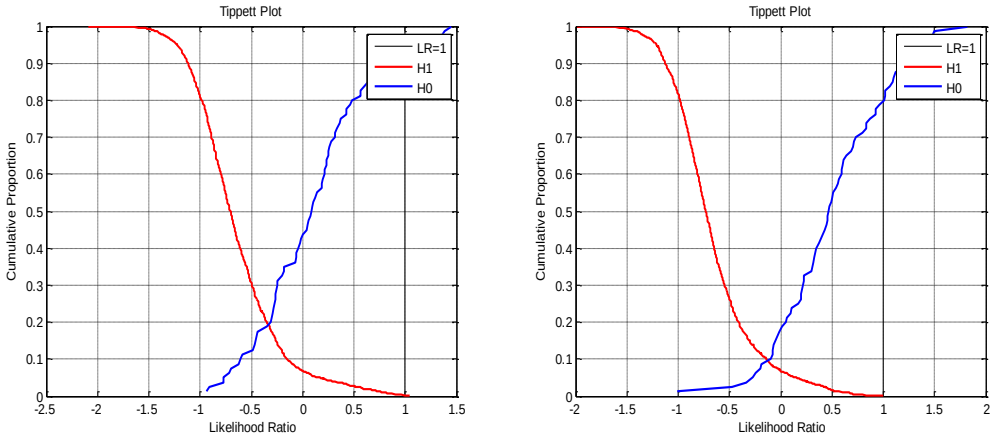
(a) (b)

Figure 4.23 : Courbes de Tippett de type II de FASR dans le bruit de l'usine



(a) (b)

Figure 4.24 : Courbes de Tippett de type I de FASR dans le bruit blanc



(a) (b)

Figure 4.25 : Courbes de Tippett de type II de FASR dans le bruit blanc

Table 4.7 : Evaluation des résultats obtenus avec l'estimateur MMSE-MODGD

Type de bruit	SNR (dB)	EPP (%)	H_0 vraie (%)		H_1 vraie (%)	
			LR<1	LR>1	LR<1	LR>1
Bavardage	0	13.55	88.75	11.25	100	0
	5	8.75	72.50	27.50	100	0
Usine	0	17.50	90	10	100	0
	5	10	79	21	100	0
Blanc	0	18.75	92	8	100	0
	5	8.75	79	21	100	0

Sur la base des résultats du Tableau 4.7, on peut observer qu'en comparant ces résultats avec ceux obtenus à la section 8.1, une amélioration significative des mesures de performances FASR en termes de EPP, $PMEH_0$ et $PMEH_1$ est observée, pour les trois types de bruits (bavardage, usine et blanc). Par conséquent, en termes de l'EPP, les améliorations représentent une réduction de 1,84 % pour le bruit de bavardage et une réduction de 1,25 % pour les autres bruits. Ces résultats sont encourageants étant donné qu'une amélioration de 1 % est significative pour les systèmes de haute sécurité tels que les systèmes FASR, car l'innocence ou l'accusation de l'individu est en jeu.

Cette amélioration apportée par l'ajout de l'estimateur MMSE-MODGD au système FASR s'explique par le fait que, la soustraction du bruit du signal de parole bruité, lors de l'utilisation du spectre de l'amplitude MMSE, ne peut pas éliminer les vallées profondes entourant les pics étroits, qui restent dans le spectre du bruit. Par conséquent, l'excursion des pics de bruit reste importante. Cependant, MMSE-MODGD écarte ces vallées profondes en préservant bien les pics et les vallées du spectre d'amplitude propre en présence de bruit additif (propriétés de la fonction de retard de groupe d'un signal de phase minimale). L'estimateur proposé fournit une distorsion de parole faible et une qualité globale plus élevée.

De plus, dans [87], les auteurs ont indiqué que le spectre MODGD est inversement proportionnel à la puissance du bruit à des fréquences correspondant à des régions à fort bruit, et directement proportionnel à la puissance du signal. Cela indique que le spectre MODGD a tendance à suivre le spectre de signal, plutôt que celui du bruit.

Ainsi, sur la base d'expériences, il a été constaté que le bruit déforme moins la forme du spectre MODGD que le spectre FFT, modifie ses pentes et réduit la plage dynamique du spectre MODGD mois que la FFT. La plupart du temps, les emplacements fréquentiels des pics des formants supérieurs sont conservés dans une certaine mesure dans le spectre MODGD par rapport au spectre FFT en présence de bruit. Par conséquent, notre proposition de MMSE-MODGD retient plus d'information contenue dans le signal de parole bruité que le

MMSE conventionnel [102], pour éviter toute dégradation de l'intelligibilité et des performances FASR.

4.9 Conclusion

Les résultats des expériences montrent que le spectre MODGD a le potentiel de réduire les composantes de bruit dans le signal de parole bruité, puisque les spectres MODGD ont tendance à suivre le spectre de l'amplitude de la parole et s'opposant au spectre de bruit. Par conséquent, on peut conclure que les informations importantes conservées dans le signal de parole amélioré utilisant le spectre MODGD peuvent compléter celles données par le spectre FFT et donner plus de fiabilité et de robustesse au système FASR dans des conditions bruitées.

CONCLUSIONS GENERALES ET PERSPECTIVES

Cette thèse s'inscrit dans le cadre du domaine de Reconnaissance Automatique Criminalistique du Locuteur (FASR), dans les milieux bruités en utilisant un algorithme amélioré de rehaussement de parole (MMSE-MODGD). Ce travail se positionne comme une contribution pour améliorer les performances et la robustesse d'un système FASR, dans des valeurs des SNR faibles pour trois types de bruit (bavardage, usine et blanc). Pour réaliser cet objectif, nous avons étudié ce système dans trois étapes :

- La première a présenté une nouvelle approche de rehaussement de parole par l'estimateur MMSE, basé sur le spectre de retard de groupe modifié (MODGD), les meilleurs résultats ont été obtenus en termes des mesures objectives telles que (SIG, BAK, OVRL, PESQ, SegSNR, WSS et LLR), par rapport aux autres estimateurs à savoir : ML, MMSE, logMMSE, MAP, MMSE-ISP, log MMSE-ISP et Wiener pour les trois bruits cités supra (dans le cas où le SNR = 0 dB, SNR = 5 dB et SNR = 10 dB), ce qui montre l'efficacité d'une telle approche qui pourra élever les performances du système FASR dans les conditions bruitées.

- La deuxième consiste en la mise en œuvre du système FASR classique, en calculant ses performances qui sont représentées par EPP, PMEH0 et PMEH1, et ceci dans les conditions propres et bruitées à SNR = 0 dB et SNR = 5 dB pour les trois types de bruit. Cette étape à son tour est réalisée en deux niveaux :

    - Dans les conditions propres : les performances du système FASR sont très satisfaisantes en termes de : pourcentage de EPP égal à 1,25 %, et le LR qui dépasse 1 dans 96 % des cas lorsque H_0 est vrai et dans seulement 0,4 % des cas lorsque H_1 est vrai ;
    - Dans les conditions bruitées : nous avons ajouté au corpus de test trois types de bruits (Bavardage, usine et blanc), extraits de la base de données NOISEX-92 avec SNR = 0 dB et SNR = 5 dB, nous avons eu une dégradation des performances en termes de l'EPP, de 37,50 % et 43,75 %, respectivement pour le bruit d'usine et le bruit blanc.

- La troisième étape porte sur la mise en œuvre du système FASR amélioré, dans les conditions bruitées (Bavardage, usine et blanc), avec SNR = 0 dB et SNR = 5 dB, en utilisant deux algorithmes d'amélioration de la parole, il s'agit de MMSE et notre approche proposée MMSE-MODGD :

- Dans le cas d'utilisation de l'estimateur MMSE, nous avons une réduction de pourcentage de l'EPP, qui atteint 10 %, 11.25 % et 8.81 % respectivement pour les bruits, bavardage, usine et blanc, dont le SNR=5 ;
- Dans le cas d'utilisation de l'estimateur MMSE-MODGD, nous avons une réduction de pourcentage de l'EPP, qui atteint 8.75 %, pour les bruits, bavardage et blanc et 10 % pour le bruit d'usine, dont le SNR=5.

Dans notre travail, des estimateurs d'amélioration de la parole du signal de parole bruité ont été étudiés sous l'hypothèse que le spectre du signal de parole bruité peut être représenté dans un plan complexe comme la somme du spectre du signal propre et du spectre du bruit. En plus de l'estimateur traditionnel, qui est basé sur les principes du MMSE, l'estimateur amélioré a été proposé en incorporant des spectres de retard de groupe modifiés. En outre, comparé aux performances du FASR, utilisant les estimateurs de puissance spectrale MMSE classiques, le FASR utilisant le MMSE-MODGD proposé a abouti à une qualité d'amélioration de la parole significativement meilleure.

En perspectives, d'autres variantes peuvent contribuer efficacement à l'amélioration des performances de FASR. Dans le cadre d'un projet au moyen terme, il est également important de prendre les paramètres suivants :

- Appliquer une technique robuste lors de la phase de paramétrage ;
- Appliquer une méthode puissante pour l'apprentissage tel que, le Deep learning, ce qui devrait être une approche intéressante pour affiner les modèles des suspects afin d'obtenir de meilleures performances pour le système FASR ;
- Utilisation d'une base de données spécifique au domaine criminalistique ;
- Utilisation d'autres types de bruits existants souvent dans la nature, tels que le bruit de restauration, véhicule, etc.

REFERENCES BIBLIOGRAPHIQUES

[1] Y. Lu and P. C. Loizou, Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty, IEEE transactions on audio, speech, and language processing, Vol. 19, N° 5, pp. 1123-1137, 2010.
[2] International Biometric Group, home page: www. Biometric group.com.
[3] www.bbc/Afrique/monde-50891797.
[4] www.médicalexpo.fr
[5] J. Daugman, High confidence visual Recognition of persons by a test of statistical independence, IEEE transaction on pattern analysis and machine intelligence, Vol. 15, pp. 1148-1161, 1993.
[6] https:// fr.wikipedia.org/wiki/reconnaissance de l'iris
[7] F. Perronnin and J. L. Dugelay, An Introduction to Biometrics Audio and Video-Based Person Authentification, Traitement du Signal, Vol. 9, N° 4, pp. 253-266, 2002.
[8] https://www.clubic.com/pro/legislation-loi-internet/cnil/actualite-437822-cnil-autorise-etude-frappe-clavier.html
[9] G. R. Doddington, M. A. Przybocki, A. F. Martin and D. A. Reynolds, The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective, Speech communication, Vol. 31, N° 2-3, pp. 225-254, 2000.
[10] NIST Speaker Recognition Evaluation, http://www.nist.gov/speech/tests/spk.
[11] https://www.cite-sciences.fr/archives/francais/ala_cite/expositions/biometrie/nonvoyants/programme_details_6_3.htm
[12] P. Verlinde, Contribution à la vérification multimodale d'identité en utilisant la fusion de décisions, 1999.
[13] Y. Chen, S. C. Dass and A. K. Jain, Fingerprint quality indices for predicting authentication performance, In International conference on audio-and video-based biometric person authentication, Springer, Berlin, Heidelberg, pp. 160-170, 2005.
[14] N. Morizet, Reconnaissance biométrique par fusion multimodale du visage et de l'iris, Thèse de doctorat, Télécom ParisTech, 2009.
[15] S. A. Carmen, NIST report to the United States Congress. Summary of NIST standards for biometric accuracy, tamper resistance and interoperability [R/OL], 2001.
[16] O. Ribaux and P. Margot, Science forensique, 2013.
[17] R. H. Bolt, F. S. Cooper, D. M. Green, S. L. Hamlet, J. G. McKnight, J. M. Pickett, O. Tosi, B. D. Underwood and D. L. Hogan, On the Theory and Practice of Voice Identification, National Research Council, National Academy of Sciences : Washington, D.C, 1979.
[18] O. Tosi, Voice Identification: Theory and Legal Applications, University Park Press: Baltimore, Maryland, 1979.
[19] J. F. Bonastre, F. Bimbot, L. J. Boë, J. P. Campbell, D. A. Reynolds and I. Magrin-Chagnolleau, Authentification des personnes par leur voix: un nécessaire devoir de précaution, Journées d'Etudes sur la Parole, pp. 33-36, 2004.
[20] R. H. Bolt, F. S. Cooper, E. E. Jr. David, P. B. Denes, J. M. Pickett and K. N. Stevens, Speaker Identification by Speech Spectrograms : A Scientists' View of its Reliability for Legal Purposes, The Journal of the Acoustical Society of America, Vol. 47, N° 2B, pp. 597-612, 1970.

[21] F. Nolan, The phonetic bases of speaker recognition: Cambridge University Press, Cambridge, 1983.

[22] L. J. Boë, Forensic voice identification in France, Speech Communication, Vol. 31, N° 2-3, pp. 205-224, 2000.

[23] A. Preti, Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur, Diss. Université d'Avignon, 2008.

[24] R. Boite, Traitement de la parole, PPUR presses polytechniques, 2000.

[25] T. Dutoit, Introduction au traitement automatique de la parole, Notes de cours, Faculté Polytechnique de Mons, 2000.

[26] S. Van Vuuren, Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch, In Proceeding of Fourth International Conference on Spoken Language Processing, ICSLP'96. IEEE, Vol. 3, pp. 1788-1791, 1996.

[27] A. Harrag, Extraction des données d'une base : Application à l'extraction des traits du locuteur, Thèse de doctorat, 2018.

[28] Y. Mami, Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence, Thèse de doctorat, Télécom ParisTech, 2003.

[29] G. R. Doddington, Speaker recognition Identifying people by their voices, Proceedings of the IEEE, Vol. 73, N° 11, pp. 1651-1664, 1985.

[30] R. Ajgou, Reconnaissance Automatique du Locuteur à Travers les Canaux Digitaux, Thèse de doctorat. Université Mohamed Khider–Biskra, 2016.

[31] D. Charlet, Authentification vocale par téléphone en mode dépendant du texte (Doctoral dissertation, Thèse de doctorat. Paris, ENST, 1997.

[32] J. Kharroubi, Etude de techniques de classement Machines à Vecteurs Supports pour la vérification automatique du locuteur, Thèse de doctorat, Télécom ParisTech, 2002.

[33] C. Fredouille, Approche statistique pour la reconnaissance automatique du locuteur: informations dynamiques et normalisation bayesienne des vraisemblances, Avignon, 2000.

[34] I. Booth, M. Barlow and B. Watson, Enhancements to DTW and VQ decision algorithms for speaker recognition, Speech Communication, Vol. 13, N° 3-4, pp. 427-433, 1993.

[35] Y. Bennani, F. F. Soulie and P. Gallinari, A connectionist approach for automatic speaker identification, International Conference on Acoustics, Speech, and Signal Processing, IEEE, pp. 265-268, 1990.

[36] D. Meuwly, Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique, Doctoral dissertation, Université de Lausanne, Faculté de droit et des sciences criminelles, 2000.

[37] P. Corsi, Speaker recognition: A survey, In Automatic Speech Analysis and Recognition, pp. 277-308, 1982.

[38] A. Larcher, Modèles acoustiques à structure temporelle renforcée pour la vérification du locuteur embarquée, Avignon, 2009.

[39] L. Ferrer, E. Shriberg, S. Kajarekar and K. Sonmez, Parameterization of prosodic feature distributions for SVM modeling in speaker recognition, IEEE International

Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Vol. 4, pp. IV-233, 2007.

[40] D. Meuwly and A. Drygajlo, Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM), A Speaker Odyssey-The Speaker Recognition Workshop, 2001.

[41] A. Drygajlo, D. Meuwly and A. Alexander, Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition, Eighth European Conference on Speech Communication and Technology, 2003.

[42] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen and T. Niemi, Methodological guidelines for best practice in forensic semi automatic and automatic speaker recognition, Verlag für Polizeiwissenschaft, 2015.

[43] V.K. Mai, Méthodes avancées de traitement de la parole et de réduction de bruit pour les terminaux mobiles, Diss, Ecole Nationale Supérieure Mines-Télécom Atlantique Bretagne Pays de la Loire, 2017.

[44] H. Gustafsson, U. Lindgren, I. Claesson and S. Nordholm, U.S. Patent No. 6,717,991. Washington, DC: U.S. Patent and Trademark Office, 2004.

[45] S. Dixit and D. M. Y. Mulge, Review on speech enhancement techniques, International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 3, N° 8, pp. 285-290, 2014.

[46] C. Verschuur, M. Lutman and N. H. A. Wahat, Evaluation of a non linear spectral subtraction noise suppression scheme in cochlear implant users, Cochlear implants international, Vol. 7, N° 4, pp. 193-6, 2006.

[47] N. Kwatra, A. A. Milani and J. Alderson, U.S. Patent No. 9,824,677. Washington, DC: U.S. Patent and Trademark Office, 2017.

[48] P. Scalart, et J. Vieira Filho, Speech Enhancement Based on a Priori Signal to Noise Estimation, IEEE International Conference on Acoustic, Speech and Signal Processing, Atlanta, États-Unis, Vol. 2, pp. 629–632, 1996.

[49] S. Vihari, A. S. Murthy and D. C. Naik, Comparison of speech enhancement algorithms, Procedia computer science, vol. 89, pp. 666-676, 2016.

[50] D. Jonathan, B. Michael and F. Sébastien, Les ondelettes, Deuxième Candidature en Sciences Physiques, Printemps des sciences, 2003.

[51] R. Benzid, Ondelettes et statistiques d'ordre supérieur appliquées aux signaux uni et bidimensionnels, Thèse de doctorat. Université de Batna 2-Mustafa Ben Boulaid, 2005.

[52] N. Virag, Single channel speech enhancement based on masking properties of the human auditory system, IEEE Trans, Speech and Audio Processing, Vol. 7, pp. 126-137, 1999.

[53] L. Buniet, Traitement Automatique de la Parole en milieu bruité: étude de modèles connexionnistes statiques et dynamiques, Thèse de doctorat, Université Henri Poincaré-Nancy 1, 1997.

[54] W. B. Kheder, Reconnaissance du locuteur en milieux difficiles, Doctoral dissertation, Université d'Avignon, 2017.

[55] C. Plapous, Traitements pour la réduction de bruit, Application à la communication parlée, Thèse de doctorat. Université de Rennes 1, 2005.

[56] V.K. Mai, Advanced methods of speech processing and noise reduction for mobile device, Ecole Nationale Supérieure Mines-Télécom Atlantique, 2017.

[57] A. Jeanvoine, Intérêt des algorithmes de réduction de bruit dans l'implant cochléaire: Application à la binauralité, Thèse de doctorat, Université Claude Bernard-Lyon I, 2012.

[58] S. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Transactions on acoustics, speech, and signal processing, Vol. 27 N° 2, pp. 113-120, 1979.

[59] M. Berouti, R. Schwartz and J. Makhoul, Enhancement of speech corrupted by acoustic noise, IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 4, pp. 208-211, 1979.

[60] R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics, IEEE Transactions on speech and audio processing, Vol. 9, N° 5, pp. 504-512, 2001.

[61] P. J. Wolfe and S. J. Godsill, Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement, EURASIP Journal on Applied Signal Processing, vol. 10, pp. 1-9, 2003.

[62] Y. Ephraim and D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, IEEE transactions on acoustics, speech, and signal processing, Vol. 33, N° 2, pp. 443-445, 1985.

[63] R. Mc Aulay and M. Malpass, Speech enhancement using a soft-decision noise suppression filter, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28, N° 2, pp. 137-145, 1980.

[64] K. Miura, An introduction to maximum likelihood estimation and information geometry, Interdisciplinary Information Sciences, Vol. 17, N° 3, pp. 155-174, 2011.

[65] P.C. Loizou, Speech enhancement: theory and practise. CRC press, 2013.

[66] T. Lotter and P. Vary, Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model, EURASIP Journal on Advances in Signal Processing, Vol. 7, pp. 1-17, 2005.

[67] R. Bassett and J. Deride, Maximum a posteriori estimators as a limit of Bayes estimators, Mathematical Programming, Vol. 174, N° 1, pp. 129-144, 2019.

[68] R. C. Hendriks, R. Heusdens and J. Jensen, Log-spectral magnitude MMSE estimators under super-Gaussian densities, In Tenth Annual Conference of the International Speech Communication Association, 2009.

[69] D. Malah, R. V. Cox and A. J. Accardi, Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments, IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings, ICASSP99 (Cat. No. 99CH36258), Vol. 2, pp. 789-792, 1999.

[70] I. Cohen and B. Berdugo, Speech enhancement for non-stationary noise environments, Signal processing, Vol. 81, N° 11, pp. 2403-2403, 2001.

[71] I. Cohen, Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator, IEEE Signal processing letters, Vol. 9, N° 4, pp. 113-116, 2002.

[72] J. Yang, Frequency domain noise suppression approaches in mobile telephone systems, IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 363-366, 1993.

[73] L. Fontan, De la mesure de l'intelligibilité à l'évaluation de la compréhension de la parole pathologique en situation de communication, Thèse de doctorat. Université Toulouse le Mirail-Toulouse II, 2012.

[74] Y. Hu and P. C. Loizou, A comparative intelligibility study of speech enhancement algorithms, IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Vol. 4, pp. IV-561, 2007.

[75] Y. Hu and P. C. Loizou, Evaluation of objective measures for speech enhancement, Ninth international conference on spoken language processing, 2006.

[76] Y. Hu and P. C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Transactions on audio, speech, and language processing, Vol. 16, N° 1, pp. 229-238, 2007.

[77] A. E. Mahdi and D. Picovici, Advances in voice quality measurement in modern telecommunications, Digital Signal Processing, Vol. 19, N° 1, pp. 79-103, 2009.

[78] S.R. Quackenbush, T.P. Barnwell and M.A. Clement, Objective measures of speech quality, Prentice-Hall, 1988.

[79] D. J. Goodman, C. Scagliola, R. E. Crochiere, L. R. Rabiner and J. Goodman, Objective and subjective performance of tandem connections of waveform coders with an LPC vocoder, The Bell System Technical Journal, Vol. 58, N° 3, pp. 601-629, 1979.

[80] D. Klatt, Prediction of perceived phonetic distance from critical-band spectra: A first step, In ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 7, pp. 1278-1281, 1982.

[81] T. F. Quatieri, Discrete-Time Speech Signal Processing, AV Oppenheim, Ed, 2002.

[82] S. Wang, A. Sekey and A. Gersho, An objective measure for predicting subjective quality of speech coders, IEEE Journal on selected areas in communications, Vol. 10, N° 5, pp. 819-829, 1992.

[83] W. Yang, M. Benbouchta and R. Yantorno, Performance of the modified bark spectral distortion as an objective speech quality measure, Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), vol. 1, pp. 541-544, 1998.

[84] ITU-T recommendation P.862, Perceptual Evaluation of Speech Quality (PESQ), objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, International Telecommunication Union, Geneva, 2001.

[85] S. Djeghiour and M. Guerti, An improved MMSE estimator based modified group delay spectrum for Forensic Automatic Speaker Recognition, International Journal of Speech Technology, Vol. 24, N° 3, PP. 687-699, 2021.

[86] O. O. Akande and P. J. Murphy, Estimation of the vocal tract transfer function with application to glottal wave analysis, Speech Communication, Vol. 46, N° 1, pp. 15-36, 2005.

[87] S. H. K. Parthasarathi, R. Padmanabhan and H. A. Murthy, Robustness of group delay representations for noisy speech signals, International Journal of Speech Technology, Vol. 14, N° 4, 361, 2011.

[88] Calliope, La parole et son traitement automatique, collection technique et scientifique des télécommunications, CNET-ENST, Masson, 1989.

[89] L. R. Rabiner and R. W. Schafer, Digital processing of speech signals, Prentice-hall, 1978.

[90] L. Burget, P. Matejka, P. Schwarz, O. Glembek and J. H. Cernocky, Analysis of feature extraction and channel compensation in a GMM speaker recognition system, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15 N° 7, 2007.

[91] D. A. Reynolds and R. C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE transactions on speech and audio processing, Vol. 3 N° 1, pp. 72-83, 1995.

[92] A. Preti, Surveillance de réseaux professionnels de communication par reconnaissance du locuteur, Ecole Doctorale 166 I2S, Laboratoire d'informatique d'Avignon, 2008.

[93] H. A. Murthy and B. Yegnanarayana, Group delay functions and its applications in speech technology, Sadhana, Vol. 36 N° 5, pp. 745-782, 2011.

[94] H. A. Chowdhury and M. S. Rahman, Formant estimation from speech signal using the magnitude spectrum modified with group delay spectrum, Acoustical Science and Technology, Vol. 42, No 2, PP. 93-102, 2021.

[95] N. Asbai and A. Amrouche, Boosting scores fusion approach using Front-End Diversity and adaboost Algorithm, for speaker verification, Computers & Electrical Engineering, vol. 62, pp. 648-662, 2017.

[96] Y. Hu, P.C. Loizou, NOIZEUS: a noisy speech corpus for evaluation of speech enhancement algorithms, Available at http://www.utdallas.edu/~loizou/speech/noize.

[97] P.C. Loizou, Speech Enhancement Theory and Practice, 1st edn, CRC Press, 2007.

[98] Y. Hu, P. C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Transactions on audio, speech, and language processing, Vol. 16 N° 1, pp. 229-238, 2007.

[99] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, Speaker verification using adapted Gaussian mixture models. Digital signal processing, vol. 10, no 1-3, pp. 19-41, 2000.

[100] R. Haraksim and A. Drygajlo, Measuring performance in forensic automatic speaker recognition: VQ, GMM-UBM, i-vectors, Biosig, 2016.

[101] S. Djeghiour, N. Asbai, O. Kenai and M. Guerti, Forensic Automatic Speaker Recognition under Noisy Environments, IC3E'2018, University of Bouira Algeria, pp.1-5, 2018.

[102] T. Gerkmann and R. C. Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, N° 4, pp. 1383-1393, 2012.