MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE FERHAT ABBAS –SETIF1-
UFAS (ALGERIE)

MEMOIRE

Présenté à la Faculté de Technologie
Département d'Electronique

Pour l'obtention du Diplôme de
MAGISTER

Option: Communication

Réalisé Par

AZIZA Yassamine

THEME

Modélisation AR et ARMA de la Parole pour une Vérification Robuste du Locuteur dans un Milieu Bruité en Mode Dépendant du Texte

Soutenu le 13 /10 /2013 devant la commission d'examen :

Mr. A. ZEGADI	PROF	à l'université de Sétif	Président
Mr. FERHAT HAMIDA A.H	PROF	à l'université de Sétif	Rapporteur
Mr. N. BOUZIT	PROF	à l'université de Sétif	Examinateur
Mr. N. BOUKAZOULA	MCCA	à l'université de Sétif	Examinateur

Dédicaces

Je dédie ce modeste travail aux étoiles de ma vie

Ma mère** et **mon père

Mon mari

*A tous les membres de la famille **HADDANA** et **AZIZA***

A toutes mes amies

Et à tous ceux qui m'aiment

YASSAMINE

Remerciements

Je tiens à remercier en premier lieu le Dieu Le tout puissant qui m'a accordé la volonté, la santé et le courage d'accomplir ce travail.

J'exprime ma vive reconnaissance à monsieur FERHAT HAMIDA A.H, professeur au département d'électronique pour son suivi, ces conseils scientifiques et méthodiques, ses soutiens moraux et matériels qu'il m'a apporté le long des années de recherches et pour l'intérêt dont il a fait preuve dans l'accomplissement de ce travail, qu'il trouve ici ma profonde gratitude.

Je remercie aussi Messieurs les membres du jury, Mr Zegadi A, Mr Bouzit N, Mr Boukazoula N, qui ont accepté de m'honorer en acceptant d'examiner, de juger et d'évaluer mon mémoire.

A toute personne qui a contribué de près ou de loin à l'élaboration de ce travail, je dis, MERCI.

Enfin, j'exprime mes vifs remerciements à mon mari, mes chers parents, beaux parents, mes frères et mes amies, qui m'ont encouragé et aidé dans les périodes difficiles et pénibles durant la réalisation de ce travail.

Résumé

Extraire d'un signal parole les paramètres pertinents les plus performants pour une vérification du locuteur définit la motivation principale de cette thèse. Ces paramètres représentent les caractéristiques du conduit vocal caractérisant la voix d'une personne.

Pour cela, nous nous intéressant à la modélisation pôles zéros *ARMA* qui est une méthode performante et peu utilisée vu sa non linéarité donc sa complexité.

Dans un premier temps, différentes méthodes d'analyses sont appliquées, *AR, LPCC, ARMA, CARMA et MFCC* dans un milieu bruité et non bruité. Ce travail est poursuivi par la modélisation du locuteur en utilisant l'approche statistique *HMM,* pour chaque locuteur chaque modèle est entrainé sur la phrase prononcé (mot de passe), il est basé sur le calcul d'un score de vraissemblance.

En l'occurrence du test nous avons obtenue un taux de reconnaissance de 60% pour le modèle *AR* et 80% pour le modèle *ARMA* dans un milieu non bruité et un taux de 60% pour le modèle *AR et 70%* pour le modèle *ARMA* Dans un milieu bruité.

SOMMAIRE

Liste des figures

Liste des tableaux

Listes des Acronymes et Symboles

Acronymes

AR	Auto Regressive
ARMA	Auto Regressive Moving Average
DTW	Dynamique Time Warping
FR	Faux Rejets
FA	Fausses Acceptations
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
IAL	Identification Automatique du Locuteur
IPA	Iterative Pre-Processing Algorithm
LPC	Linear prediction coefficient
LPCC	Linear prediction cepstral coefficient
MA	Moving Average ou Moyenne Ajustée
MFCC	Mel Frequency Cepstral Coefficient
RAL	Reconnaissance Automatique du Locuteur
SMA	Algorithme Steiglitz-McBride
TFCT	Transformée de Fourier à Court Terme
VAL	Vérification Automatique du Locuteur

INTRODUCTION GENERALE

Support pour le transport de l'information, vecteur de notre identité et témoin de nos émotions, la communication parlée est l'interface majeure des interactions humaines. Plus qu'un simple moyen de porter un message, la voix est un outil privilégié pour l'expression et l'affirmation de soi, pour convaincre, séduire ou entraîner l'adhésion de l'auditeur.

Il a fallu attendre le milieu des années soixante et la diffusion des moyens de calcul numérique pour voir un enrichissement très important entrainant des recherches théorique nouvelles sur le traitement de la parole.

Le traitement de la parole regroupe trois grands axes qui sont l'analyse, le codage et la reconnaissance. Il a largement bénéficié, des outils du traitement numérique du signal.

Comme tout signal, la parole doit être modélisée pour tout traitement. Ceci revient à trouver une fonction de transfert qui décrit au mieux le signal.

Les caractéristiques propres de la parole ont fait que le modèle le plus utilisé fut le modèle *AR (Auto Regressive)* où un segment est représenté par une fonction de transfert tous pôles. Bien que le modèle *AR* soit facile à obtenir et nécessite un temps de calcule très réduit, il a l'inconvénient majeur de mal représenter les sons nasalisés qui nécessite des zéros dans la fonction de transfert.

Les spécialistes du traitement de la parole savent depuis très longtemps que le meilleur modèle est le modèle *ARMA* (*Auto Regressive Moving Average*), modèle où la transmittance contient des pôles et des zéros. Cependant, il est rarement utilisé car il nécessite un calcul très volumineux. Les ordinateurs de l'époque n'étaient pas accessibles et n'avaient pas la puissance de calcul de ceux de nos jours. Les dernière années, la modélisation *ARMA* est utilisé surtout dans l'analyse et la reconnaissance de la parole sous l'impulsion des vitesses phénoménales que les ordinateurs d'aujourd'hui ont atteint.

Pour vérifier que le modèle *ARMA* de la parole est meilleur que le modèle *AR*, nous avons utilisé ces deux derniers dans la vérification du locuteur en mode dépendant du texte.

La vérification du locuteur consiste à déterminer si un locuteur est bien celui qu'il prétend être. Dans ce type d'applications, il s'agit donc de trancher entre les deux hypothèses : soit le locuteur est bien le locuteur autorisé, c'est à dire celui dont l'identité est revendiquée, soit nous avons affaire à un imposteur qui cherche à se faire passer pour un locuteur autorisé. Les applications classiquement envisagées pour la vérification du locuteur correspondent donc à l'idée de "serrure vocale" qui peut être utilisée, par exemple, pour valider des transactions bancaires effectuées par téléphone, ou pour compléter un dispositif d'accès (à un bâtiment, un système informatique).

Pour cela nous avons utilisé les modèles de Markov cachés (*HMM*) comme des reconnaisseurs.

Le mémoire est constitué de quatre chapitres :

Le premier chapitre est une introduction générale au domaine du signal de la parole, décrit le signal de parole, ses caractéristiques, son système de production et quelques rappelles essentiels.

Le deuxième chapitre est consacré à l'étude des différentes méthodes des modèles *AR* (Durbin et Schur) et *ARMA* (itératives et non itératives)

Le troisième chapitre entame de la reconnaissance automatique du locuteur, les principales méthodes utilisées en reconnaissance. Nous étudions de prés la méthode de Modèle de Markov Caché *HMM*. Cette méthode nous a permis de créer des modèles des locuteurs à base d'une approche statistique.

Enfin, le quatrième chapitre détaille les différentes étapes utilisées dans notre application. Nous démarrons par l'acquisition du signal de parole des différents locuteurs, la modélisation de ce dernier avec les modèles *AR, LPCC, ARMA, CARMA, MFCC* jusqu'à la création de leurs modèles statistiques *HMM* et aussi la comparaison entres les résultats de la phase d'apprentissage et la phase de test. Enfin nous prendrons une décision sur tous ces modèles d'analyse du signal parole.

Chapitre I

GENERALITE SUR LE SIGNAL DE LA PAROLE

I.1. INTRODUCTION [1]

Le traitement de la parole est une science située au croisement du traitement du signal numérique et du traitement du langage. La parole a la particularité, par rapport aux autres signaux du traitement de l'information, à être produite et perçue instantanément par le cerveau, et pour cela le traitement de la parole tend à remplacer ces fonctions par des systèmes automatiques (*figure I.1*):

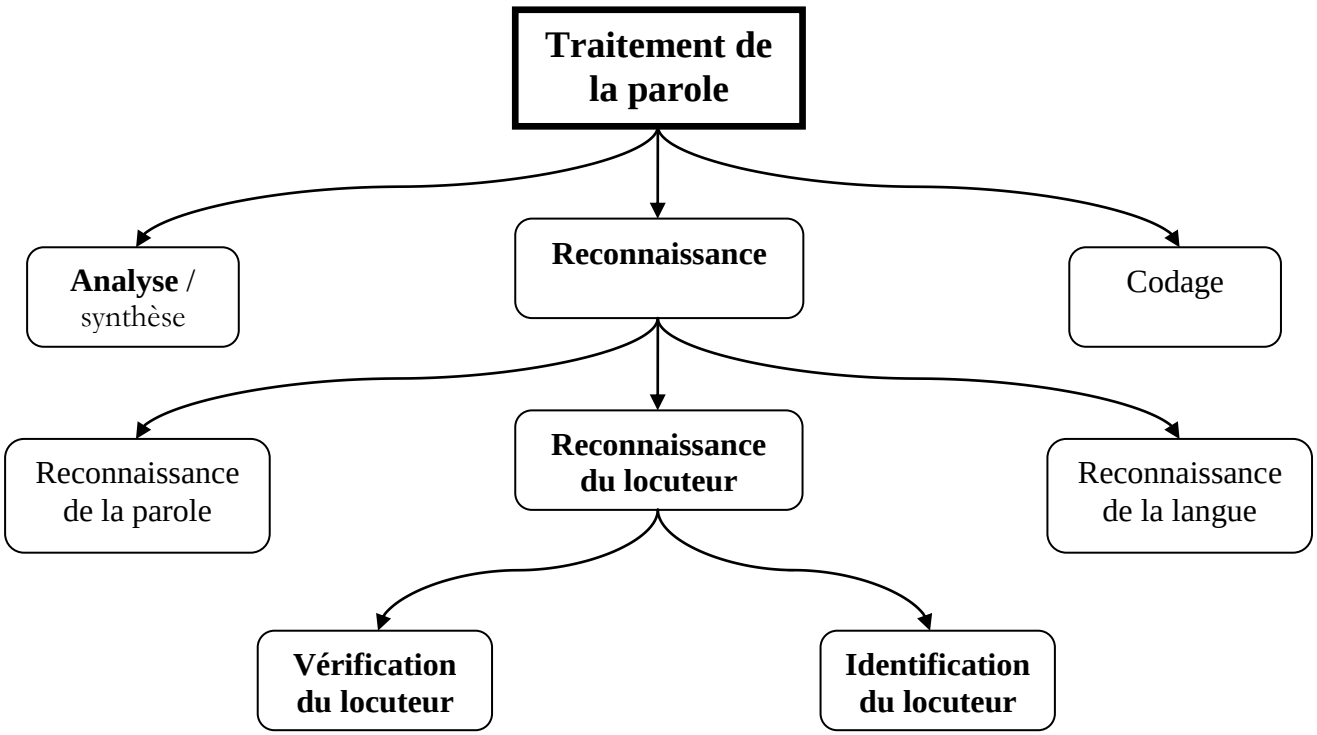


Figure I.1: Traitement de la parole.

- **Les analyseurs de parole** mettent en évidence les caractéristiques du signal vocal tel qu'il est produit. Ils sont utilisés soit comme composant de base de systèmes de codage, de reconnaissance ou de synthèse.
- **Les reconnaisseurs de parole** décodent l'information portée par le signal vocal à partir des données fournies par l'analyse. On les classes en fonction de l'information que l'on cherche à extraire du signal vocal:

La reconnaissance du locuteur qui est l'identification (vérifier que la voix analysée correspond bien à la personne qui est sensée la produire) ou **la vérification** du locuteur (déterminer qui, parmi un nombre fini et préétabli de locuteurs, a produit le signal analysé.).

Il y a aussi la reconnaissance du locuteur **dépendante du texte (**la phrase à prononcer pour être reconnu est fixée dès la conception du système**)**, la reconnaissance avec **texte dicté** (La phrase à prononcer est fixée lors du test), et reconnaissance indépendante du texte **(**La phrase à prononcer n'est pas précisé**)**.

En plus du reconnaisseur de parole monolocuteur, multilocuteur, ou indépendant du locuteur pour reconnaître la voix d'une personne, d'un groupe fini de personnes, ou de reconnaître n'importe qui.

Enfin le reconnaisseur de mots isolés, reconnaisseur de mots connectés, et reconnaisseur de parole continue, selon que le locuteur sépare chaque mot par un silence, qu'il prononce de façon continue une suite de mots prédéfinis, ou qu'il prononce n'importe quelle suite de mots de façon continue.

Les synthétiseurs de parole est une technique informatique de synthèse sonore qui permet de créer de la parole artificielle à partir de n'importe quel texte. Pour obtenir ce résultat, elle s'appuie à la fois sur des techniques de traitement linguistique, notamment pour transformer le texte orthographique en une version phonétique prononçable sans ambiguïté, et sur des techniques de traitement du signal pour transformer cette version phonétique en son numérisé écoutable sur un haut parleur. Il y a deux types de synthétiseurs : les synthétiseurs de parole à partir d'une représentation numérique, inverses des analyseurs, dont la mission est de produire de la parole à partir des caractéristiques numériques d'un signal vocal telles qu'obtenues par analyse, et les synthétiseurs de parole à partir d'une représentation symbolique (texte ou concept), inverse des reconnaisseurs de parole et capables en principe

de prononcer n'importe quelle phrase sans qu'il soit nécessaire de la faire prononcer par un locuteur humain au préalable.

- Enfin, le rôle des codeurs est de permettre la transmission ou le stockage de parole avec un débit réduit, ce qui passe tout naturellement par une prise en compte judicieuse des propriétés de production et de perception de la parole [1].

I.2. LA PAROLE

La parole est un signal continu, d'énergie finie, non stationnaire. Sa structure est complexe et variable dans le temps:

- Tantôt périodique (plus exactement pseudopériodique) pour les sons voisés,
- Tantôt aléatoire pour les sons fricatifs,
- Tantôt impulsionnelle dans les phases explosives des sons occlusifs.

I.2.1. ARCHITECTURE ET FONCTIONNEMENT DE L'APPAREIL VOCAL [3] [4]

I.2.1.1. L'APPAREIL VIBRATEUR

L'air est la matière première de la voix, en expulsant l'air pulmonaire à travers la trachée, le système respiratoire joue le rôle d'une soufflerie. Il s'agit du souffle phonatoire produit, soit par l'abaissement de la cage thoracique, soit dans le cadre de la projection vocale par l'action des muscles abdominaux.

L'extrémité supérieure de la trachée est entourée par un ensemble de muscles et de cartilages mobiles qui constituent le larynx. Le larynx se trouve au carrefour des voies aériennes et digestives (*Figure I.2*), entre le pharynx et la trachée, et en avant de l'œsophage. Les plis vocaux, communément nommés cordes vocales, sont deux lèvres symétriques (structures fibreuses) placées en travers du larynx. Ces lèvres se rejoignent en avant et sont plus au moins écartées l'une de l'autre sur leur partie arrière (structure en forme de V) ; l'ouverture triangulaire résultante est nommée glotte.

Le larynx et les plis vocaux forment notre « appareil vibrateur ». Lors de la production d'un son qualifié de non-voisé (ou sourd), comme c'est le cas, par exemple, pour les phonèmes [s] ou [f], les plis vocaux sont écartés et l'air pulmonaire circule librement en

direction des structures en aval. En revanche, lors de la production d'un son voisé (ou sonore), comme c'est le cas, par exemple, pour les phonèmes [z], [v] et pour les voyelles, les plis vocaux s'ouvrent et se ferment périodiquement, obstruant puis libérant par intermittence le passage de l'air dans le larynx. Le flux continu d'air pulmonaire prend ainsi la forme d'un train d'impulsions de pression, nos cordes vocales vibrent. Le dernier élément principal de notre appareil vibrateur est l'épiglotte. Lors de la déglutition, cette dernière agit comme un clapet qui se rabat sur le larynx, conduisant les aliments vers l'œsophage en empêchant leur passage dans la trachée et les poumons.

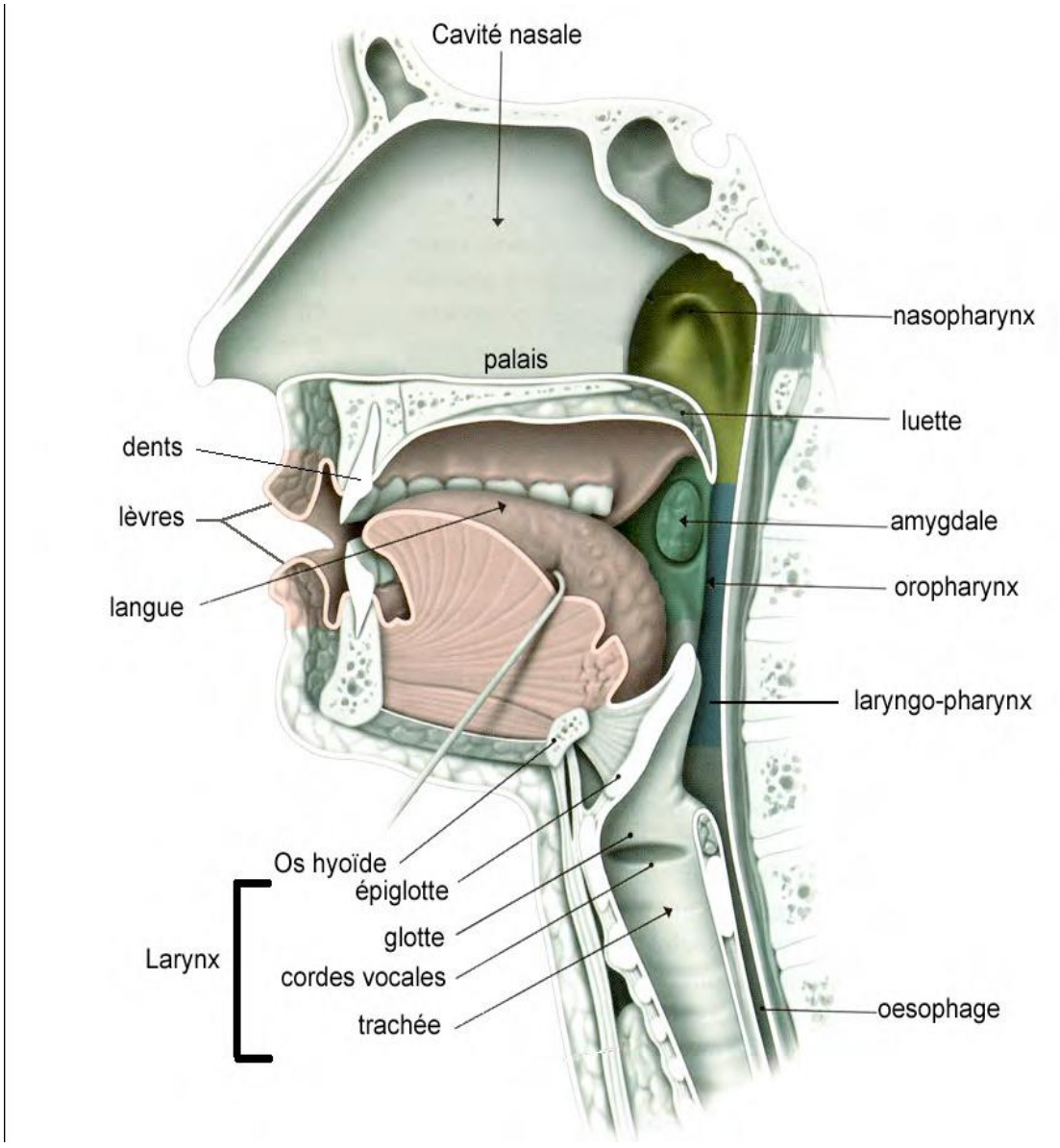


Figure I.2: Vue schématique de l'appareil vocal [4].

I.2.1.2. LE RESONATEUR

L'air pulmonaire, ainsi modulé par l'appareil vibrateur, est ensuite appliqué à l'entrée du conduit vocal. Ce dernier est principalement constitué des cavités pharyngiennes (laryngophaynx et oropharynx situés en arrière-gorge) et de la cavité buccale (espace qui s'étend du larynx jusqu'aux lèvres). Pour la réalisation de certains phonèmes, le voile du palais (le velum) et la luette qui s'y rattache, s'abaissent, permettant ainsi le passage de l'air dans les cavités nasales (fosses nasales et rhinopharynx ou nasopharynx). Ces différentes cavités forment un ensemble que nous qualifierons ici de résonateur. Si l'appareil vibrateur peut être décrit comme le lieu de production de la voix, le résonateur apparaît alors comme le lieu de naissance de la parole. Il abrite en effet des organes mobiles, nommés articulateurs, qui en modifiant sa géométrie et donc ses propriétés acoustiques, mettent en forme le son laryngé (ou son glottique) en une séquence de sons élémentaires. Ces derniers peuvent être interprétés comme la réalisation acoustique d'une série de phonèmes, unités linguistiques élémentaires propres à une langue. Les articulateurs principaux sont la langue, les lèvres, le voile du palais et la mâchoire [4].

I.3. CLASSIFICATION DES SONS DE LA PAROLE

Une décomposition simplifiée du signal de la parole doit ressortir deux types de sons: voisés et non voisés.

- *Les sons voisés*, tels que des voyelles, sont produits par le passage de l'air de poumons à travers la trachée qui met en vibration les cordes vocales. Ce mode, qui représente 80% du temps de phonation, est caractérisé en général par une quasi-périodicité, une énergie élevée et une fréquence fondamentale (pitch). Typiquement, la période fondamental des différents sons voisés varie entre *2ms* et *20ms*.

- *Les sons non voisés*, comme certaines consonnes, dans ce cas les cordes vocales ne vibrent pas, l'air passe à haute vitesse entre les cordes vocales. Le signal produit est équivalent à un bruit blanc.

I.4. PARAMETRES DU SIGNAL DE PAROLE

Le signal vocal est généralement caractérisé par trois paramètres: sa fréquence fondamentale, son énergie et son spectre.

I.4.1. LA FREQUENCE FONDAMENTALE

Elle représente la fréquence du cycle d'ouverture/fermeture des cordes vocales. Cette fréquence caractérise seulement les sons voisés, elle peut varier [1] :

- De 80Hz à 200Hz pour une voix masculine,
- De 150Hz à 450Hz pour une voix féminine,
- De 200Hz à 600Hz pour une voix d'enfant.

I.4.2. L'ENERGIE

Elle est représentée par l'intensité du son qui est liée à la pression de l'air en amont du larynx. L'amplitude du signal de la parole varie au cours du temps selon le type de son, et son énergie dans une trame est donnée par :

$$E = \sum_{n=0}^{N-1} s^2(n) \tag{I.1}$$

Avec N : la taille de la trame.

I.4.3. LE SPECTRE

L'enveloppe spectrale ou spectre représente l'intensité de la voix selon la fréquence, elle est généralement obtenue par une analyse de Fourier à court terme.

La quasi stationnarité du signal de parole permet de mettre en œuvre des méthodes efficaces d'analyse et de modélisation utilisées pour le traitement à court terme du signal vocal sur des fenêtres de durée généralement comprise entre *20ms* et *30ms* appelées trames, avec un recouvrement entre ces fenêtres qui assure la continuité temporelle des caractéristiques de l'analyse.

La transformée de Fourier à court terme (*TFCT*) d'un signal échantillonné est par définition la transformée du signal pondéré.

$$\hat{S}(k) = \hat{S}\left(f = \frac{k}{N} \right) = \sum_{n=0}^{N-1} s(n) \cdot w(n) \cdot \exp(-j2\pi nk/N) \qquad 0 \le k \le N-1 \tag{I.2}$$

Où : N : Le nombre de points prélevés.

$S(k)$: Spectre complexe.

$s(n)$: Segment analysé**.**

$w(n)$: Fenêtre de temps.

Le spectre de puissance (appelé aussi densité spectrale de puissance de la transformé de Fourier) est donné par:

$$\left|\hat{S}(k)\right|^2, 0 \le k \le \frac{N}{2} \tag{I.3}$$

I.5. PROPRIETES STATISTIQUES DU SIGNAL VOCAL

L'audiogramme ou l'évolution temporelle du signal vocal ne fournit pas directement les traits acoustiques du signal, il est nécessaire de mener un ensemble de calculs statiques. Le signal vocal peut être vu comme un processus aléatoire non stationnaire.[1]

I.5.1. DENSITE DE PROBABILITE

Si N_ξ représente le nombre d'échantillons *x(n)* dont l'amplitude est comprise entre *[ξ-Δξ/2, ξ+Δξ/2]* alors que $n \in [-N, N]$, la densité de probabilité du signal *x* supposé ergodique et stationnaire est donnée par:

$$P_x(\xi) = \lim_{N\to\infty}\left[N_\xi / (2N+1)\right] \tag{I.4}$$

I.5.2. LA VALEUR MOYENNE ET LA VARIANCE

La valeur moyenne d'un système stationnaire vaut:

$$\mu_x = \int_{-\infty}^{+\infty} \xi P_x(\xi) d\xi = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n) \tag{I.5}$$

Pour le signal vocal cette moyenne est supposé nulle, elle ne contient aucune information utile. La variance est donnée par:

$$\sigma_x^2 = \int_{-\infty}^{+\infty} \xi^2 P_x(\xi) d\xi = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x^2(n) \tag{I.6}$$

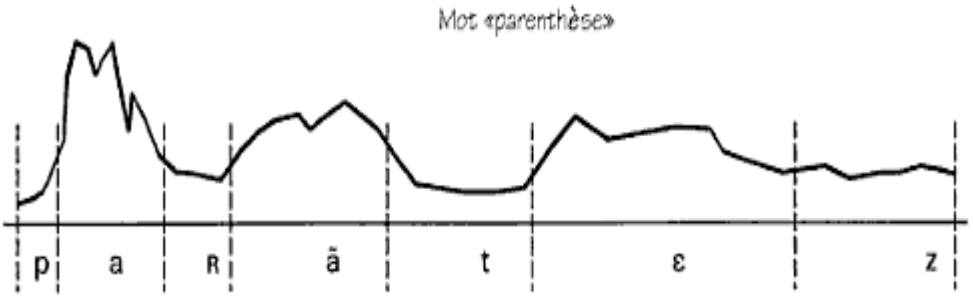


Figure I.3: Variance du mot parenthèse.[1]

I.5.3. FONCTION D'AUTOCORRELATION

La fonction d'autocorrélation d'un signal ergodique et stationnaire s'exprime par:

$$\phi_{xx}(k) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)x(n+k) \tag{I.7}$$

L'estimation sur un nombre fini N d'échantillons peut être calculé par:

$$\phi_{xx}(k) = \frac{1}{N-k} \sum_{n-0}^{N-k} x(n)\,x(n+k) \tag{I.8}$$

La fonction d'autocovariance est définie et estimée par des formules identiques après avoir soustrait la moyenne μ_x et comme $\mu_x = 0$ alors ces deux notions peuvent être confondues. On a aussi : $\sigma_x^2 = \emptyset_{xx}(0)$.

Le coefficient d'autocorrélation est défini par $\rho_x(k) = \frac{\emptyset_{xx}(k)}{\emptyset_{xx}(0)}$ sa valeur est comprise entre 1 et -1.

I.5.4. DENSITE SPECTRALE DE PUISSANCE

Densité spectrale de puissance $S_{xx}(\theta)$ est la transformée de fourrier de la fonction d'autocorrélation:

$$S_{xx}(\theta)=\sum_{k}\phi_{xx}(k)\exp(-jk\theta),\qquad \theta=\omega T_e \tag{I.9}$$

$$\hat{S}_{xx}(\theta)=\sum_{k=-k}\phi_{xx}(k)w(k)\exp(-jk\theta) \tag{I.10}$$

Où $w(k)$ est une fonction fenêtre.

I.6. MODELISATION DE LA PRODUCTION DE LA PAROLE [2]

L'absence de couplage entre la glotte et le conduit vocal permet de modéliser séparément la source et le système de production. Pour les sons voisés, la source est un train périodique d'onde de forme particulière (montée rapide en pression suivie d'une chute plus graduelle). Ce train d'ondes est modélisé par la réponse d'un passe bas d'ordre 2 à pôles réels et dont la fréquence de coupure est de l'ordre de *100Hz*, sa transmittance est de la forme : [2]

$$G(z)=\frac{A}{\left(1+\alpha z^{-1}\right)\left(1+\beta z^{-1}\right)} \tag{I.11}$$

Pour les sons non voisés la source est un bruit blanc.

Le conduit vocal peut être assimilé à une succession de tubes acoustiques élémentaires. L'étude de la propagation d'une onde acoustique plane conduit à une modélisation par une cascade de résonateurs dont la transmittance est de la forme:

$$V(z)=\frac{B}{\prod_{k=1}^{k}\left(1+b_{1k}z^{-1}+b_{2k}z^{-2}\right)} \tag{I.12}$$

Chaque résonateur correspond à un formant dont la fréquence centrale est donnée par :

$$F_k=\frac{1}{2\pi}f_s\cos^{-1}\left[\frac{-b_{1k}/2}{\sqrt{b_{2k}}}\right] \tag{I.13}$$

Où f_s est la fréquence d'échantillonnage.

Le son est finalement émis à travers l'ouverture des lèvres, celle-ci représente une charge acoustique, le rayonnement des lèvres peut être modélisé par la transmittance:

$$R(z)=C\left(1-z^{-1}\right) \tag{I.14}$$

Qui exprime que la pression de l'onde observée à une certaine distance des lèvres est proportionnelle à la dérivée du débit volumique aux lèvres.

En résumé la transmittance globale entre le train d'impulsion et le signal émis serait

$$T(z)=G(z)V(z)R(z)=\frac{\sigma\left(1-z^{-1}\right)}{\left(1+\alpha z^{-1}\right)\left(1+\beta z^{-1}\right)\prod_{k=1}^{k}\left(1+b_{1k}z^{-1}+b_{2k}z^{-2}\right)} \tag{I.15}$$

Si l'on considère que l'un des pôles de *G(z)* est très voisin de l'unité, on obtient la forme simplifiée:

$$T(z)=\frac{\sigma}{\left(1+\alpha z^{-1}\right)\prod_{k=1}^{k}\left(1+b_{1k}z^{-1}+b_{2k}z^{-2}\right)}=\frac{\sigma}{A(z)} \tag{I.16}$$

On a posé :

$$A(z)=\left(1+\alpha z^{-1}\right)\prod_{k=1}^{k}\left(1+b_{1k}z^{-1}+b_{2k}z^{-2}\right)=1+\sum_{i=1}^{2k+1}a_i z^{-i} \tag{I.17}$$

La transmittance de ce modèle est dite tous pôles, son inverse le polynôme *A(z)*, est la transmittance du filtre inverse. Les limitations de ce modèle sont évidentes:

En premier lieu, la source est soit un train périodique, soit un bruit blanc. Les sons fricatifs voisés ne peuvent pas être produits par ce modèle.

En second lieu, la production de sons nasalisés fait intervenir deux cavités associées en parallèle, la transmittance correspondante est donc de la forme:

$$\frac{\sigma_1}{A_1(z)}+\frac{\sigma_2}{A_2(z)}=\frac{\sigma_1 A_2(z)+\sigma_2 A_1(z)}{A_1(z)A_2(z)} \tag{I.18}$$

et elle présente des zéros en *z* distincts de l'origine.

La transmittance tous pôles (I.16) est la base de la modélisation par prédiction linéaire, la présence d'un numérateur qui ne serait pas une simple constante complique énormément l'estimation des paramètres du modèle. On spécule donc sur l'identité:

$$1-\alpha z^{-1}\cong\frac{1}{1+\alpha z^{-1}+\alpha^2 z^{-2}+\ldots} \tag{I.19}$$

Pour substituer à un zéro en $z = a(0 < |a| < 1)$ un ou deux pôles, en d'autre terme on accepte de surestimer le degré du dénominateur pour pouvoir assimiler le numérateur à une constante.

I.7. CONCLUSION

Le signal vocal ne peut être considéré comme quasi stationnaire que sur des intervalles de temps de durée limitée. On est donc amené à considérer des tranches successives et à estimer un modèle *AR* ou *ARMA* pour chacune d'elles.

Chapitre II

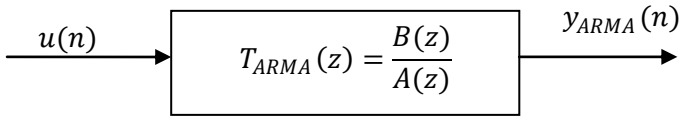
MODELISATION AR *ET* ARMA *DU SIGNAL DE LA PAROLE*

II.1. INTRODUCTION

La modélisation d'un signal $x(n)$ consiste à lui associer un filtre linéaire qui, soumit à une excitation particulière reproduit ce signal le plus fidèlement possible. L'objectif essentiel de la modélisation d'un signal est de permettre la description de son spectre par un ensemble très limité de paramètres.

La modélisation classique d'un signal aléatoire est basée sur un filtre de transmittance rationnelle $T(z) = B(z)/A(z)$ de degré $(q)/(p)$ excité par un bruit blanc $u(n)$de moyenne μ_u nulle et de variance σ_u^2 unitaire *Figure (II.1)*

Dans le cas général $p > 0\ et\ q > 0$, on parle de modélisation *ARMA*. Lorsque $B(z) = 1\ (p > 0, q = 0)$, il s'agit d'une modélisation *AR* : $T(z)$ est une fonction tous pôles. Lorsque $A(z) = 1\ (p = 0, q > 0)$, la transmittance $T(z)$ est celle d'un filtre *RIF*: il s'agit d'une modélisation *MA* (*Moving Average* ou *Moyenne Ajustée*).



$\mu_u = 0, \sigma_u^2 = 1$

Figure II.1 : Modèle *ARMA* pour un signal aléatoire stationnaire [1].

Le modèle le plus utilisé est le modèle *AR*. Ses capacités de modélisation (aptitude à approcher le spectre du signal) sont suffisantes pour beaucoup d'application.

Le modèle *ARMA* est évidemment celui qui possède la plus grande capacité de modélisation pour un degré donné, mais son estimation est complexe. Quant au modèle *MA*, ses capacités de modélisation sont limitées.

II.2. MODELISATION *AR* DU SIGNAL DE LA PAROLE [1,2]

Le signal de parole n'est bien sûr pas stationnaire et on va devoir travailler sur des tranches de temps de durée limitée, avec recouvrement : une procédure classique consiste à calculer un modèle pour des tranches successives de 30*ms* décalées de 10*ms*: si la fréquence d'échantillonnage est de 10*kHz*, l'estimation de chaque modèle est faites sur 300 échantillons.

Il existe deux mécanismes particuliers qui ont engendré le signal vocal:

- Un signal voisé est engendré par un filtre *AR* excité par un train d'impulsion dont la fréquence correspond à celle des vibrations des cordes vocales (pitch).
- Un signal non voisé est engendré par un modèle *AR* excité par du bruit blanc.

Le modèle est illustré sur la *figure (II.2).*

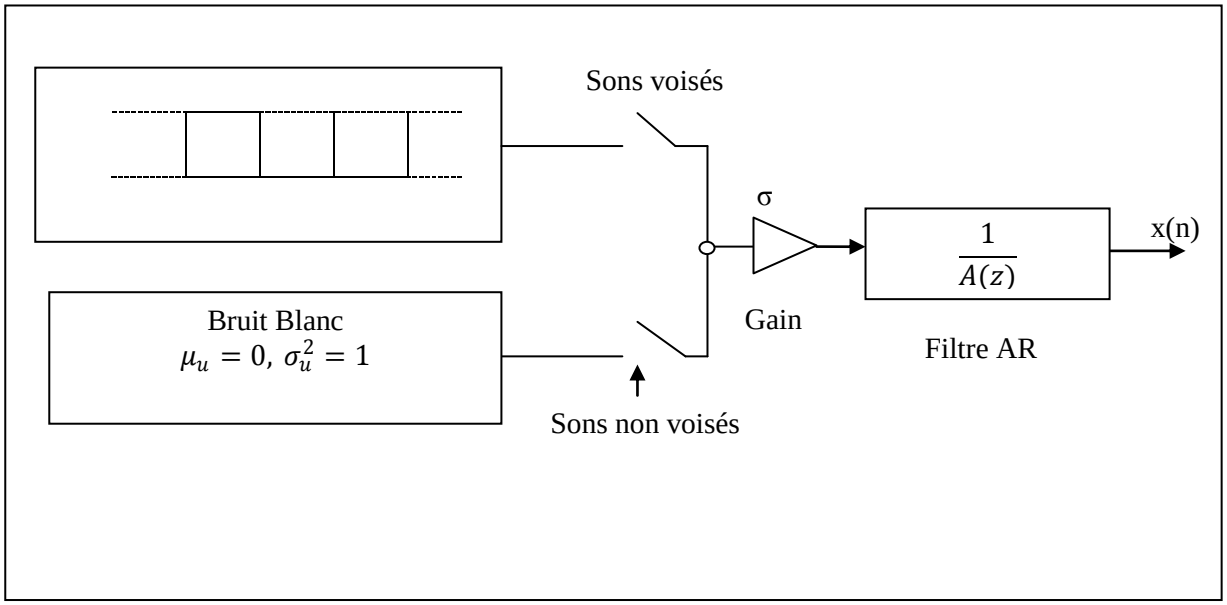


Figure II.2: Modèle *AR* pour le signal vocal [1].

II.2.1. PREDICTION LINEAIRE

La prédiction linéaire comme nous le verrons conduit à un modèle *AR*, elle est basée sur l'hypothèse que chaque échantillon du signal original $x(n)$ peut être approché par une combinaison linéaire des *p* échantillons qui le précèdent [5]:

$$x(n) = -a(1)\cdot x(n-1) - a(2)\cdot x(n-2) \ldots - a(p)\cdot x(n-p) + e(n) \tag{II.1}$$

Dans cette expression les coefficients $a(i)$ sont les coefficients de prédiction d'ordre *p*, et le signal:

$$e(n) = \sum_{i=0}^{p} a(i)\cdot x(n-i) \qquad a(0) = 1 \tag{II.2}$$

est l'erreur de prédiction (ou résidu) d'ordre *p*.

L'estimation des coefficients de prédiction est basée sur la minimisation de la variance de l'erreur de prédiction:

$$\begin{aligned} \sigma_e^2 &= E\left[e(n)^2\right] \\ &= E\left[\sum_{i=0}^{p} a(i)\cdot x(n-i)\cdot \sum_{j=0}^{p} a(j)\cdot x(n-j)\right] \\ &= E\left[\sum_{i,j}^{p} a(i)\cdot a(j)\cdot x(n-i)\cdot x(n-j)\right] \\ &= \sum_{i,j}^{p} a(i)\cdot a(j)\cdot \phi_x(i-j) \end{aligned} \tag{II.3}$$

Où $\phi_x(k)$ est la fonction d'autocorrélation du signal x et $\phi_x(0) = \sigma_x^2$.

La minimisation de (II.3) par rapport à $a(i)$ conduit à

$$\begin{aligned} \frac{\partial E\left[e^2(n)\right]}{\partial a_i} &= \frac{\partial}{\partial a_i}\left[a_i^2\phi_x(0) + 2a_i\sum_{j=0,i\neq j}^{p} a_j\phi_x(j-i) + terme\right] = 0 \\ \frac{\partial E\left[e^2(n)\right]}{\partial a_i} &= 2a_i\phi_x(0) + 2\sum_{j=0,i\neq j}^{p} a_j\phi_x(j-i) = 0 \end{aligned} \tag{II.4}$$

Condition d'optimalité:

$$\sum_{j=0}^{p} a_j \phi_x(j-i) = 0 \qquad pour \qquad i = 1,2,...,p \qquad et \qquad a_0 = 1$$

Cette condition s'écrit aussi:

$$\phi_x(i) = \sum_{j=0}^{p} a_j \phi_x(j-i) \qquad avec \qquad \phi_x(0) = \sigma_x^2$$

La valeur minimisée de la variance de l'erreur de prédiction σ_e^2 qui sera notée $\alpha_p = \sigma_{e,min}^2$ vaut:

$$\sigma_{e,\min}^2 = \sum_{j=0}^{p} a_j \phi_x(j) = \alpha_p \tag{II.5}$$

On obtient donc les équations de *Yule-Walker*:

$$\begin{bmatrix} \sigma^2 & \phi(1) & \phi(2) & \cdots & \cdots & \phi(m) \\ \phi(1) & \sigma^2 & \phi(1) & & & \phi(m-1) \\ \phi(2) & \phi(1) & \sigma^2 & & & \phi(m-2) \\ \vdots & & & & & \vdots \\ \vdots & & & & & \vdots \\ \phi(m-1) & & & & \sigma^2 & \phi(1) \\ \phi(m) & \phi(m-1) & \cdots & \cdots & \phi(1) & \sigma^2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ a_m(1) \\ a_m(2) \\ \vdots \\ \vdots \\ \vdots \\ a_m(m) \end{bmatrix} = \begin{bmatrix} \alpha_m \\ 0 \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \tag{II.6}$$

Ces équations sont connues sous le nom d'équations de *Yule Walker*. La matrice des coefficients d'autocorrélation, qui y apparait, a une forme bien particulière: tous ses éléments situés sur des parallèles à la diagonale principale sont identique, on dit que c'est une matrice de *Toeplitz*. Elle est de plus symétrique.

Ce système est évidemment redondant puisqu'il comporte *p+1* équations en *p* inconnues: il ne sera pas résolu sous cette forme qui par contre convient parfaitement pour l'élaboration des algorithmes de résolution.

Interprétation

Dans ce qui précède, le signal $e(n)$ a été considéré comme une erreur de prédiction dont on a minimisé la variance σ_e^2 pour estimer les coefficients de prédiction $a(i)$.

Cependant on peut aussi interpréter la relation (II.1) comme définissant un filtre excité par un signal $e(n)$ et produit un signal $x(n)$.

Si $E(z)\ et\ X(z)$ désignent les transformées respectives de $e(n)\ et\ x(n)$, cette relation (II.1) conduit à:

$$\begin{aligned} X(z) &= \frac{1}{1+a_p(1)z^{-1}+a_p(2)z^{-2}+\cdots a_p(p)z^{-p}}E(z) \\ &= \frac{1}{A(z)}\cdot E(z) = T(z)\cdot E(z) \end{aligned} \qquad \text{(II.7)}$$

La transmittance $T(z)$ est celle d'un filtre numérique *RII* tous pôles, c'est la définition du modèle *AR* associé au signal $x(n)$.

II.2.2. ALGORITHME DE RESOLUTION

Les algorithmes décrits dans cette section sont de nature récursive sur l'ordre de la prédiction, on construit successivement les solutions pour *m=1,2*….jusqu'à *p*.

II.2.2.1. METHODE DE LEVINSON-DURBIN

L'algorithme de *Levinson* concerne l'inversion d'une matrice de *Toeplitz* symétrique. Considérons le système d'ordre *m* (II.6) [1]:

$$\begin{bmatrix} \sigma^2 & \phi(1) & \phi(2) & \cdots & \cdots & \phi(m) \\ \phi(1) & \sigma^2 & \phi(1) & & & \phi(m-1) \\ \phi(2) & \phi(1) & \sigma^2 & & & \phi(m-2) \\ \vdots & & & & & \vdots \\ \vdots & & & & & \vdots \\ \phi(m-1) & & & & & \phi(1) \\ \phi(m) & \phi(m-1) & \cdots & \cdots & \phi(1) & \sigma^2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ a_m(1) \\ a_m(2) \\ \vdots \\ \vdots \\ \vdots \\ a_m(m) \end{bmatrix} = \begin{bmatrix} \alpha_m \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \qquad \text{(II.8)}$$

Et le système

$$\begin{bmatrix} \sigma^2 & \phi(1) & \phi(2) & \cdots & \cdots & \phi(m) \\ \phi(1) & \sigma^2 & \phi(1) & & & \phi(m-1) \\ \phi(2) & \phi(1) & \sigma^2 & & & \phi(m-2) \\ \vdots & & & & & \vdots \\ \vdots & & & & & \vdots \\ \phi(m-1) & & & & & \phi(1) \\ \phi(m) & \phi(m-1) & \cdots & \cdots & \phi(1) & \sigma^2 \end{bmatrix} \cdot \begin{bmatrix} b_m(1) \\ b_m(2) \\ \vdots \\ \vdots \\ \vdots \\ b_m(m) \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \beta_m \end{bmatrix} \quad \text{(II.9)}$$

Où : α_m et β_m sont les variances de l'erreur de prédiction avant et arrière d'ordre *m* respectivement.

a_m et b_m sont les vecteurs des coefficients de prédiction arrière avant et arrière d'ordre *m* respectivement.

$$\beta_m = \alpha_m \quad et \quad b_m = a_m^T \quad \text{(II.10)}$$

Pour trouver la solution d'ordre *m+1*, ajoutons une composante nulle en bas du vecteur $[1, a_m]^T$ dans (II.6) et en tête du vecteur $[b_m, 1]^T$ dans (II.9). Etant donné (II.10) et la structure particulière de la matrice $\phi^{(m+1)}$, on peut écrire d'une part:

$$\begin{bmatrix} \sigma^2 & \phi(1) & \cdots & \cdots & \phi(m) & \phi(m+1) \\ \phi(1) & \sigma^2 & \phi(1) & & \phi(m-1) & \phi(m) \\ \phi(2) & \phi(1) & \sigma^2 & & \phi(m-2) & \phi(m-1) \\ \vdots & & & & & \vdots \\ \vdots & & & & & \vdots \\ \phi(m-1) & & & & & \phi(1) \\ \phi(m) & \phi(m-1) & & \phi(1) & \sigma^2 & \phi(1) \\ \phi(m+1) & \phi(m) & \cdots & \cdots & \phi(1) & \sigma^2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ a_m(1) \\ a_m(2) \\ \vdots \\ \vdots \\ \vdots \\ a_m(m) \\ 0 \end{bmatrix} = \begin{bmatrix} \alpha_m \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ \mu_m \end{bmatrix} \quad \text{(II.11)}$$

Et d'autre part :

$$\begin{bmatrix} \sigma^2 & \phi(1) & \cdots & \cdots & \phi(m) & \phi(m+1) \\ \phi(1) & \sigma^2 & \phi(1) & & \phi(m-1) & \phi(m) \\ \phi(2) & \phi(1) & \sigma^2 & & \phi(m-2) & \phi(m-1) \\ \vdots & & & & & \vdots \\ \vdots & & & & & \vdots \\ \phi(m-1) & & & & & \phi(1) \\ \phi(m) & \phi(m-1) & & \phi(1) & \sigma^2 & \phi(1) \\ \phi(m+1) & \phi(m) & \cdots & \cdots & \phi(1) & \sigma^2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ b_m(1) \\ b_m(2) \\ \vdots \\ \vdots \\ \vdots \\ b_m(m) \\ 0 \end{bmatrix} = \begin{bmatrix} \tau_m \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ \beta_m \end{bmatrix} \qquad \text{(II.12)}$$

Rappelant que : $\alpha_m = \sum_{i=0}^{m} a_m(i).\phi(i) = \sigma_x^2 + \sum_{i=1}^{m} a_m(i).\phi(i)$, expression que l'on peut vérifier en observant le système (II.11) d'autre part: $\beta_m = \alpha_m$.

En réalité, on n'obtient bien sur pas de la sorte la solution d'ordre *m+1* par suite de la présence des constantes non nulles μ_m et τ_m, la solution correcte est en fait une combinaison linéaire des deux solutions proposées:

$$a_{m+1}(i) = a_m(i) + k_{m+1} \cdot b_m(i) \qquad i = 0,1,...,m \qquad \text{(II.13)}$$

Soit

$$A_{m+1}(z) = A_m(z) + k_{m+1} \cdot B_m(z) \qquad \text{(II.14)}$$

Il suffit de choisir k_{m+1} tel que:

$$\begin{gathered} \mu_m + k_{m+1} \cdot \beta_m = 0 \\ et \\ \alpha_m = \beta_m = \frac{-\mu_m}{k_{m+1}} \end{gathered} \qquad \text{(II.15)}$$

Pour retrouver la forme canonique du système d'ordre *m+1* dans (II.11) et (II.12). Les constantes μ_m et τ_m valent respectivement:

$$\begin{aligned} \mu_m &= \sum_{i=0}^{m} a_m(i) \cdot \phi(m+1-i) \\ \tau_m &= \sum_{i=1}^{m+1} b_m(i) \cdot \phi(i) \end{aligned} \qquad \text{(II.16)}$$

Sachant que : $b_m(i) = a_m(m+1-i) \qquad i = 1,2,\ldots.m+1$

Alors: $\mu_m = \tau_m$

Si l'on tient compte de (II.15) et(II.16), on obtient aisément:

$$k_{m+1} = -\frac{\mu_m}{\alpha_m} = -\left[\sum_{i=0}^{m} a_m(i)\cdot\phi(m+1-i)\right]\cdot(1/\alpha_m) \qquad \text{(II.17)}$$

Ainsi que:

$$\alpha_{m+1} = \alpha_m \cdot \left(1+k_{m+1}^2\right) \qquad \text{(II.18)}$$

II.2.2.2. ALGORITHME DE LEVINSON

L'algorithme commence par le calcul de la fonction d'autocorrélation dans laquelle on fait un choix d'une valeur convenable pour le nombre d'échantillons N. Pour un signal stationnaire, on a:

$$\phi_x(i-j) = \phi_x(|i-j|) = \phi_x(k) \qquad k = 1,2,...,p \qquad \text{(II.19)}$$

Viennent ensuite :

- ✓ *Initialisation :*

$$a_m(0) = 1, \quad m = 1,2,...p \qquad \alpha_0 = \phi_x(0) = \sigma_x^2$$

- ✓ *Récursion :*

- Pour : $m = 0,1,....p-1$

$$k_{m+1} = -(1/\alpha_m)\cdot\sum_{i=0}^{m-1} a_{m-1}(i)\cdot\phi_x(m-1)$$

- Pour : $i = 1,2,\dots,m-1$

$$a_m(i) = a_{m-1}(i) + k_m \cdot a_{m-1}(m-i)$$
$$a_{m+1}(m) = k_{m+1}$$
$$\alpha_{m+1} = \alpha_m\left(1-k_{m+1}^2\right)$$

II.2.2.3. METHODE DE SCHUR :

Dans la méthode de *Levinson* on passe des systèmes (II.8) et (II.9) d'ordre m au système (II.11) et (II.12) d'ordre *m+1* en ajoutant un zéro à la fin du vecteur a_m et au début du vecteur b_m. La solution d'ordre *m+1* est une combinaison linéaire des solutions d'ordre *m* [2].

Dans la méthode de *Schur*, on passe directement à des systèmes d'ordre *m* aux systèmes d'ordre *m+1* en ajoutant *p-m+1* zéros pour obtenir le système (II.20) d'ordre *p* :

$$\phi^{(p)}\cdot\begin{bmatrix} 1 & 0 \\ & \\ & \\ a_m & b_m \\ & \\ & \\ 0 & 1 \\ -- & -- \\ 0 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} a_m & \tau_m(m+1) \\ 0 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \\ \mu_m(m+1) & \beta_m \\ --- & --- \\ \mu_m(m+2) & \tau_m(m+2) \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \mu_m(p) & \tau_m(p) \end{bmatrix} \begin{matrix} \\ \\ \\ (m+2) \\ \\ \\ \\ \\ \\ (p-m+1) \\ \\ \\ \end{matrix} \qquad \text{(II.20)}$$

$$\mu_m(j) = \sum_{i=0}^{m} a_m(i)\cdot\phi(i-j) \qquad \text{(II.21)}$$

$$\tau_m(j) = \sum_{i=1}^{m+1} a_m(m+1-i)\cdot\phi(i-j) \qquad j = m+2, m+3, ..., p \qquad \text{(II.22)}$$

Ces éléments vont jouer le rôle des variables auxiliaires sur laquelle va porter l'itération. Si l'on multiplie chaque membre de (II.20) par la matrice : $\begin{bmatrix} 1 & k_{m+1} \\ k_{m+1} & 1 \end{bmatrix}$

On obtient :

$$\phi^{(p)} \cdot \begin{bmatrix} 1 & k_{m+1} \\ a_m + k_{m+1} \cdot b_m & k_{m+1} \cdot a_m + b_m \\ k_{m+1} & 1 \\ -- & -- \\ 0 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \begin{matrix} \\ (m+2) \\ \\ \\ \\ (p-m+1) \end{matrix} = \cdots$$

$$\cdots = \begin{bmatrix} \alpha_{m+1} & 0 \\ 0 & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \\ \mu_m(m+2) \cdot k_{m+1} \cdot \tau_m(m+2) & k_{m+1} \cdot \mu_m(p) + \tau_m(p) \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \mu_m(p) + k_{m+1} \cdot \tau_m(p) & k_{m+1} \cdot \mu_m(p) + \tau_m(p) \end{bmatrix} \begin{matrix} \\ (m+2) \\ \\ \\ \\ \\ \\ (p-m+1) \end{matrix} \quad \text{(II.23)}$$

On a posé pour cela:

$$\mu_m(m+1) + \alpha_m \cdot k_{m+1} = 0 \quad \text{(II.24)}$$

Soit:

$$k_{m+1} = -(1/\alpha_m) \cdot \mu_m(m+1) \quad \text{(II.25)}$$

Dans la seconde matrice du premier membre de (II.23) les *(m+2)* élément de tête de la première colonne constituent la solution d'ordre *m+1*. Si l'on décale tous les éléments de la

seconde colonne d'une place vers le bas en introduisant un zéro en tête, on obtient l'expression:

$$\phi^{(p)}\cdot\begin{bmatrix} 1 & 0 \\ \vdots & k_{m+1} \\ \vdots & \vdots \\ a_m + k_{m+1}\cdot b_m & k_{m+1}\cdot a_m + b_m \\ \vdots & \vdots \\ k_{m+1} & \vdots \\ 0 & 1 \\ -- & -- \\ 0 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ a_{m+1} & b_{m+1} \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 1 \\ -- & -- \\ 0 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} \\ \\ \\ (m+3) \\ \\ \\ \\ \\ \\ \\ (p-m-1) \\ \\ \end{matrix} \qquad \text{(II.26)}$$

Qui n'est autre que la forme générale itérée du premier membre de (II.20) . Le système complet devient donc:

$$\phi^{(p)}\cdot\begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ a_{m+1} & b_{m+1} \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 1 \\ -- & -- \\ 0 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \alpha_{m+1} & \tau_{m+1} \\ 0 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \\ \mu_{m+1}(m+2) & \alpha_{m+1} \\ --- & --- \\ \mu_{m+1}(m+3) & \tau_{m+1}(m+3) \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \mu_{m+1}(p) & \tau_{m+1}(p) \end{bmatrix} \begin{matrix} \\ \\ \\ (m+3) \\ \\ \\ \\ \\ \\ \\ (p-m-2) \\ \\ \end{matrix} \qquad \text{(II.27)}$$

Or dans le second membre de (II.26), la première colonne n'a pas été affectée, tandis que la seconde a été décalée d'une place vers le bas avec une introduction d'un zéro en tête. Ceci est dû à la structure particulière de la matrice $\phi^{(p)}$ et au fait que les *(p-m-2)* dernier éléments de la matrice multiplicande sont des zéros. On en tire les récurrences:

$$\begin{aligned}\mu_{m+1}(i) &= \mu_m(i) + k_{m+1}\cdot\tau_m(i)\\ \tau_{m+1}(i) &= k_{m+1}\cdot\mu_m(i) + \tau_m(i)\end{aligned} \qquad \text{(II.28)}$$

L'initialisation de ces variables résulte du système (II.23) dans lequel on pose *m=0*:

$$\phi^{(p)}.\begin{bmatrix}1 & 0\\ 0 & 1\\ 0 & 0\\ \vdots & \vdots\\ \vdots & \vdots\\ \vdots & \vdots\\ \vdots & \vdots\\ \vdots & \vdots\\ 0 & 0\end{bmatrix} = \begin{bmatrix}\alpha_0 & \mu_0(1)\\ \mu_0(1) & \alpha_0\\ \mu_0(2) & \tau_0(2)\\ \vdots & \vdots\\ \vdots & \vdots\\ \vdots & \vdots\\ \vdots & \vdots\\ \vdots & \vdots\\ \mu_0(p) & \tau_0(p)\end{bmatrix} \qquad \text{(II.29)}$$

Soit:

$$\mu_0(i) = \tau_0(i+1) = \phi(i), \qquad = 0,1,2,...,p \qquad \text{(II.30)}$$

Ces résultats conduisent à l'algorithme de *Schur*.

II.2.2.4. ALGORITHME DE SCHUR

On définit deux vecteurs:

$$U = \left[u(0), u(1), ..., u(p)\right]^T$$
$$V = \left[v(0), v(1), ..., v(p)\right]^T$$

- ✓ *Initialisation:*

$$U = V = \left[\phi(0), \phi(1), ..., \phi(p)\right]^T$$

- ✓ *Récursion:*

– Pour : $m = 1,2,\dots,p$

– Pour : $i = 0,1,\dots,p-m$

$$v(i) \Leftarrow v(i+1)$$

$$k_m = v(0)/u(0) \quad \{si \quad |k_m| \geq 1 \quad ou \quad m = p : stop\}$$

– Pour : $i = 0,1,\dots,p-m$ effectuer:

$$u(i) \Leftarrow u(i) + k_m \cdot v(i)$$
$$v(i) \Leftarrow k_m \cdot u(i) + v(i)$$

II.2.2.5. COEFFICIENTS DE CORRELATION PARTIELLE ET PARAMETRE LPC :

Les coefficients $k_m (m = 1,2,\dots,p)$ calculés par l'algorithme de *Levinson* ou de *Schur* caractérisent complètement les coefficients de prédiction $a_m(i)(m = 1,2,\dots,p; i = 1,2,\dots,m)$. Ils sont appelés coefficients de corrélation partielle**.** Les paramètres $a_m(i)\ et\ k_m$ sont appelés couramment paramètre *LPC***.**

II.3. MODELISATION *ARMA* DU SIGNAL DE LA PAROLE

Les sons nasalisés introduisent des zéros dans le spectre de la parole. Et reproduire des segments de parole nasalisées à partir d'un modèle *AR* (qui n'a pas de zéros) sonnent souvent artificielle. Les tentatives d'estimer les sons nasalisées en utilisant un ordre élevé du modèle *AR* n'ont pas été très réussies à cause des formants parasites qui sont produites. La contribution de la région nasale nécessite un modèle *ARMA* (ou pôle-zéro) [6].

Par conséquent, nous avons étudié le modèle *ARMA* pour le filtre du conduit vocal. Ceci peut être décrit par la fonction de transfert :

$$V(z) \equiv \frac{N(z)}{D(z)} = \frac{\sum_{k=0}^{q} b_k z^{-k}}{1 + \sum_{k=1}^{p} a_k z^{-k}} \qquad \text{(II.31)}$$

où les a_k sont les coefficients des pôles, b_k sont les coefficients des zéros, et p et q sont les ordres des pôles et des zéros, respectivement. Le signal estimé $\hat{y}_n$ peut être exprimé en fonction du signal original y_n et du signal d'entrée u_n comme:

$$\hat{y}_n = -\sum_{k=1}^{p} a_k y_{n-k} + \sum_{k=0}^{q} b_k u_{n-k} \qquad \text{(II.32)}$$

Donc l'erreur entre le signal original y_n et le signal estimé $\hat{y}_n$ est:

$$e_n = y_n + \sum_{k=1}^{p} a_k y_{n-k} - \sum_{k=1}^{p} b_k u_{n-k} \tag{II.33}$$

II.3.1. METHODES DE MODELISATION ARMA

Dans la modélisation *ARMA,* les pôles et les zéros sont déterminés en minimisant l'erreur quadratique totale comme dans le cas *AR*. Cependant, les équations obtenues par la mise des dérivés de l'erreur égale à zéro sont non linéaires.

L'erreur estimée dans (II.33) peut être minimisé en utilisant soit les méthodes itératives soit les méthodes non itératives.

II.3.1.1. LES METHODES NON ITERATIVES

Les techniques non itératives résolvent les pôles et les zéros séparément en deux étapes successives. Cependant, ils ont l'avantage du faible coût de calcul. Les deux principales méthodes non itératives étudiées ici sont *Shanks* [7,8] et la *LPC* inverse [8].

D'abord, le modèle *AR* est utilisé pour identifier les pôles dans les deux méthodes.

A. IDENTIFICATION DES ZEROS PAR LA METHODE DE SHANKS

Shanks a proposé un procédé dans lequel le numérateur de *V(z)* est estimé par un critère des moindres carrés. Supposons que *V(z)* est donné par (II.31) alors:

$$v(n) = \sum_{k=0}^{R} b_k f(n-k) \tag{II.34}$$

Où : *f(n)* est la séquence tous pôles de la transformée en *z*

$$F(z) \equiv \frac{1}{D(z)}$$

Dans la méthode de *Shanks, D(z)* est estimé par la covariance de *LPC* et *N(z)* est estimé en minimisant:

$$E \equiv \sum_{n=0}^{\infty} \left| v(n) - \sum_{k=0}^{R} b_k \tilde{f}(n-k) \right|^2 \tag{II.35}$$

Où:

$$\tilde{F}(z) \equiv \frac{1}{\tilde{H}_v(z)}$$

Ce qui conduit aux équations linéaires :

$$\sum_{k=0}^{R} \tilde{b}_k \phi_{\tilde{f}\tilde{f}}(k,r) = \phi_{v\tilde{f}}(0,r), \qquad r = 0...R \tag{II.36}$$

Dans laquelle :

$$\phi_{x_1x_2}(t,u) \equiv \sum_{n=0}^{\infty} x_1(n-t)\, x_2(n-u) \tag{II.37}$$

Est la corrélation entre $x_1(n)$ et $x_2(n)$.

Une interprétation de la méthode de *Shanks* comme une approximation polynomiale des moindres carrés est établie par la réécriture de (II.35) dans le domaine fréquentiel. Appliquant le théorème de *Parseval* et la définition de $\tilde{F}(z)$

$$E = \int_{-\pi}^{\pi} \left| V(z)\tilde{H}_v(z) - \sum_{k=0}^{R} b_k z^{-k} \right|^2 \left|\tilde{F}(z)\right|^2 \frac{d\omega}{2\pi}$$

Où $z=e^{jw}$ Rappelons :

$$\tilde{E}_v(z) = V(z)\tilde{H}_v(z)$$

On trouve que :

$$E = \int_{-\pi}^{\pi} \left|\tilde{E}(z) - N_s(z)\right|^2 \left|\tilde{F}(z)\right|^2 \frac{d\omega}{2\pi} \tag{II.38}$$

Où :

$$N_s(z) \equiv \sum_{k=0}^{R} b_k z^{-k}$$

B. *IDENTIFICATION DES ZEROS PAR LPC INVERSE*

Une deuxième méthode d'estimation de *N(z)* a été proposée par *Makhoul* [5]. Cette approche implique l'inversion du spectre de *v(n)*, puis on estime ses zéros par le modèle *AR*. Si *V(z)* est donné par (II.31) alors

$$V^{-1}(z) \equiv \frac{1}{V(z)} = \frac{D(z)}{N(z)} \tag{II.39}$$

Alors les pôles de $V^{-1}(z)$ sont les zéros de *V(z)* et vice versa. Quand *LPC* est appliqué sur $V^{-1}(z)$, le polynôme prédicteur obtenu est une estimation de *N (z)*.

II.3.1.2. LES METHODES ITERATIVES

Les méthodes itératives, comme l'algorithme *Steiglia-McBride (SMA)* [9] et l'algorithme *IPA* (*Iterative Pre-Processing Algorithm*) [10], sont plus coûteuses en termes de calcul que les méthodes non itératives, mais fournissent des solutions optimales.

Comme la méthode non itérative de *Shanks*, ces deux méthodes itératives déterminent les paramètres du modèle *ARMA* de la réponse impulsionnelle du filtre conduit vocal. Comme pour les méthodes non itératives, une opération de filtrage passe-haut est également appliquée en premier. La méthode *SMA* et la méthode du *IPA* réestiment les paramètres du modèle à chaque itération en utilisant le résultat de l'itération précédente. Ces méthodes sont des prolongements de la méthode de *Kalman*, Où les paramètres *p* + *q* sont des modèles obtenues par régression linéaire sur l'entrée et la sortie du modèle.

A. METHODE DE STEIGLITZ MCBRIDE

Cette méthode estime les pôles et les zéros de la fonction de transfert du conduit vocal simultanément.

$$N(z) = \alpha_0 + \alpha_1 z^{-1} + \ldots + \alpha_{n-1} z^{-(n-1)}$$

$$D(z) = 1 + \beta_1 z^{-1} + \ldots + \beta_n z^{-n}$$

Kalman a suggéré de trouver les 2n coefficients par régression linéaire sur l'entrée et la sortie du modèle. Dans *figure (II.3)*, si *x* et *w* sont les échantillons d'entrée et sortie respectivement, la minimisation suivante est impliquée:

$$\sum e_j^2 = \frac{1}{2\pi j}\oint |XN - WD|^2 \frac{dz}{z} = \min \tag{II.40}$$

Où:

$$X = X(z) = \sum x_j z^{-j}$$

$$W = W(z) = \sum w_j z^{-j}$$

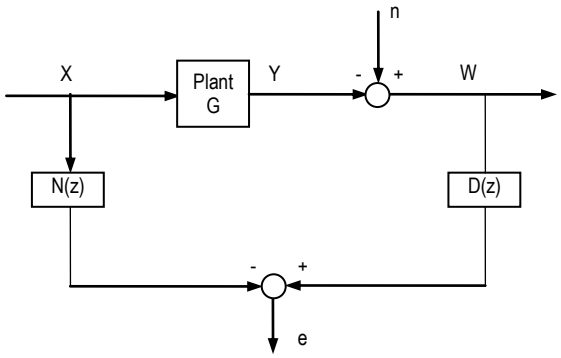


Figure II.3 : L'erreur de la régression linéaire [11].

- ✓ *Solution des équations de régression linéaire :*

L'erreur de *figure (II.3)* au temps *j* est donné par:

$$e_j = \sum_{i=0}^{n-1} \alpha_i x_{j-i} - \sum_{i=1}^{n} \beta_i w_{j-i} - w_j \tag{II.41}$$

si le coefficient du vecteur δ' et le vecteur entrée-sortie q_j' sont définis par :

$$\delta' = [\alpha_0, ..., \alpha_{n-1}, -\beta_1, ..., -\beta_n]$$

$$q_j' = [x_j, ..., x_{j-n+1}, w_{j-1}, ..., w_{j-n}]$$

Ceci devient :

$$e_j = q_j'\delta - w_j \tag{II.42}$$

Où (′) désigne la transposée.

$$grad\left(\sum e_j^2\right) = \partial / \partial\delta\left(\sum e_j^2\right) = 2\sum\left(\partial e_j / \partial\delta\right) e_j = 2\sum q_j e_j = 0 \qquad \text{(II.43)}$$

La substitution de (II.42) dans (II.43) donne:

$$\left(\sum q_j q_j'\right)\delta = \sum w_j q_j$$

Si la matrice de corrélation *2nX2n* est définie par:

$$Q = \sum q_j q_j'$$

Et le vecteur de corrélation 2n calculé à partir de x et *w* par :

$$c = \sum w_j q_j$$

La solution du problème de minimisation originale est:

$$\delta = Q^{-1} c \qquad \text{(II.44)}$$

Où:

$$\delta = \begin{bmatrix} \alpha \\ -\beta \end{bmatrix}$$

Le problème de minimisation schématisé sur la *figure (II.3)* est facile à résoudre et l'erreur résiduelle de (II.40) ne dispose d'aucune interprétation physique réelle. Un problème plus significatif serait de minimiser l'erreur indiquée dans la *figure (II.4)*

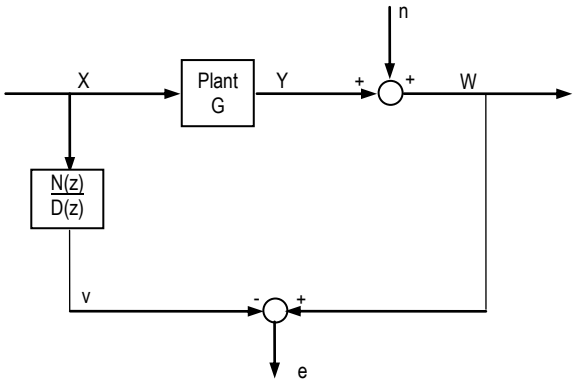


Figure II.4 : L'erreur correcte du modèle [11].

$$\sum e_j^2 = \frac{1}{2\pi j}\oint \left| X\,\frac{N}{D} - W \right|^2 \frac{dz}{z} = \min \qquad \text{(II.45)}$$

C'est l'erreur quadratique de la moyenne entre la sortie prédite et la sortie observée de ***plant***. Malheureusement, l'équation (II.45) est un problème de régression non linéaire.

L'idée de la procédure itérative est schématisée sur la *figure (II.5)*

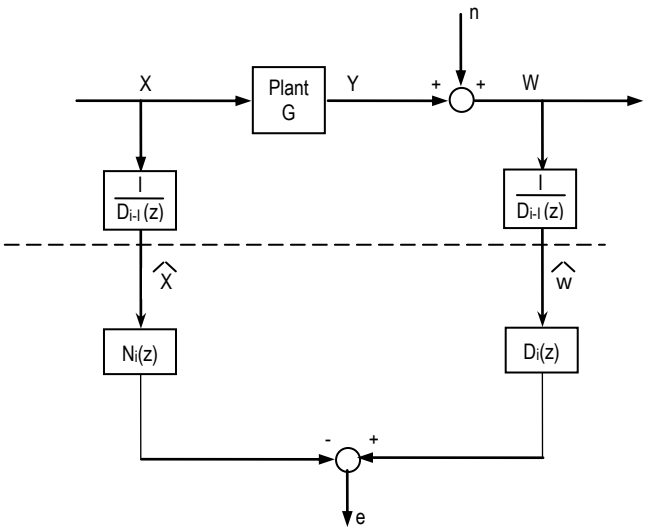


Figure II.5 : Schémas de la procédure de *Steiglitz Mc Bride* [11].

D'abord le problème de minimisation (II.40) est résolu en utilisant (II.44) et les entrées-sorties. Le résultat est une première estimation de N*(z)* et *D(z)* (estimation de *Kalman*). Appelons ceci $N_1(z)$ et $D_1(z)$. L'entrée et la sortie *x* et *w* sont ensuite filtrés par le filtre numérique $1/D_1(z)$, donnant une nouvelle entrée et sortie pré filtrées $\hat{x}$ et $\hat{w}$. Ces deux

derniers sont utilisés à la place de l'entrée-sortie originale dans (II.44) et une nouvelle estimation $N_2(z)$ et $D_2(z)$ est obtenue. Le filtre numérique $1/D_2(z)$ est utilisé pour trouver $\hat{x}$ et $\hat{w}$ de *x* et *w*, et ainsi de suite. Le dénominateur précédent est utilisé pour pré filtrer l'entrée et la sortie, afin que $N_i(z)$ et $D_i(z)$ sont trouvées tel que:

$$\oint \left| X\frac{N_i}{D_{i-1}} - W\frac{D_i}{D_{i-1}} \right|^2 \frac{dz}{z} = \oint \left| X\frac{N_i}{D_i} - W \right|^2 \left| \frac{D_i}{D_{i-1}} \right|^2 \frac{dz}{z} = \min \quad i = 1,2,3.... \quad and \quad D_0 = 1 \qquad \text{(II.46)}$$

- ✓ *Itération mode **2 :***

Etant donné L'erreur *E(z)* est donné par:

$$E(z) = X(z)\frac{N(z)}{D(z)} - W(z) \qquad \text{(II.47)}$$

Ses dérivés partielles sont données par:

$$\frac{\partial E(z)}{\partial \alpha_i} = \frac{X(z)}{D(z)} z^{-i} = \hat{X}(z) z^{-i}$$

$$\frac{\partial E(z)}{\partial \beta_i} = \frac{-X(z)N(z)}{D^2(z)} z^{-i} = -\frac{V(z)}{D(z)} z^{-i} = -\hat{V}(z) z^{-i} \qquad \text{(II.48)}$$

Un nouveau vecteur $p_j^{'}$ est défini par:

$$p_j' = \left[\hat{x}_j, ..., \hat{x}_{j-n+1}, -\hat{v}_{j-1}, ..., -\hat{v}_{j-n} \right]$$

Le gradient de l'erreur réel devient:

$$grad\left(\sum e_j^2\right) = 2\sum p_j e_j = 2\sum \left(p_j q_j' \delta - w_j p_j \right) \qquad \text{(II.49)}$$

Où la seconde égalité est vraie uniquement à la convergence. La procédure est ensuite identique à celui du mode 1, sauf que les définitions de *Q* et *c* devient:

$$Q = \sum p_j q_j' \quad et \quad c = \sum w_j p_j$$

La méthode de *Steiglitz McBride* a donné une amélioration significative sur l'estimation de *Kalman*.

B. MÉTHODE IPA (ITERATIVE PREPROCESSING ALGORITHME)

Supposons une séquence observée y(nT) ayant N échantillons d'un système constitués de K plus un bruit non corrélé e(nT) de telle sorte que [10]:

$$y(nT) = \sum_{i=1}^{k} c_i \exp(S_i nT) + e(nT) \tag{II.50}$$

Pour : *n=0,1,....N-1*, où S_i sont les pôles, c_i sont les résidus, et T est la période d'échantillonnage. L'équation (II.50) devient:

$$y_n = \sum_{i=1}^{k} c_i z_i^n + e_n \tag{II.51}$$

Où $z_i = exp(s_i T)$. L'approche directe de la minimisation d'une erreur par rapport aux pôles et résidus est non linéaire et difficile à résoudre. Le système a surtout une fonction de transfert d'ordre K donné par :

$$\begin{aligned} H(z) &= \frac{A(z)}{B(z)} = \frac{a_0 z^k + a_1 z^{k-1} + \ldots + a_{k-1} z}{b_0 z^k + b_1 z^{k-1} + \ldots + b_k} \\ &= w_0 + w_1 z^{-1} + w_2 z^{-2} + \ldots \end{aligned} \tag{II.52}$$

Où b0=1 et w_n sont des données sans bruit. Notez que les z_i sont les racines du polynôme B(z). En termes d'équation matricielle, l'équation (II.52) devient:

$$Wx + w = a \tag{II.53}$$

Où:

$$W = \begin{bmatrix} \square & & & 0 & 0 \\ & & & 0 & w_0 \\ & & & w_0 & w_1 \\ & & \iddots & \iddots & \vdots \\ & \iddots & \iddots & & \vdots \\ w_0 & w_i & \cdots & \cdots & w_{k-1} \\ w_1 & w_2 & \cdots & \cdots & w_k \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ w_{N-k-1} & w_{N-k} & \cdots & \cdots & w_{N-2} \end{bmatrix}$$

$$w = \left[w_0, w_1, ..., w_{N-1} \right]^T$$
$$x = \left[b_k, b_{k-1}, ..., b_1 \right]^T$$
$$\tilde{a} = \left[a_0, a_1, ..., a_{k-1}, 0, 0, ..., 0 \right]^T = \left[a^T, 0^T \right]^T$$

L'équation (II.53) peut être réécrite comme:

$$Bw = \tilde{a} \tag{II.54}$$

Où :

$$B = \begin{bmatrix} 1 & & & & & & \\ b_1 & 1 & & & & & \\ . & . & . & & & & \\ . & & . & . & & & \\ b_k & .. & .. & b_1 & 1 & & \\ & . & & . & . & & \\ & & . & & . & . & \\ & & & . & & . & . \\ & & & b_k & .. & b_1 & 1 \end{bmatrix}$$

Dans une expérience, on observe $y_n = w_n + e_n$ pour n=0,1,….N-1. Sous forme vectorielle :

$$y = w + e \tag{II.55}$$

Où: $y = [y_0, \ldots . y_{N-1}]^T$ et $e = [e_0, \ldots . , e_{N-1}]^T$. Étant donné y, on cherche à estimer les coefficients du polynôme dénominateur dans (II.52). Soit Y la matrice, qui est réalisé sous la même manière que W avec y_N comme éléments. Puis, à partir de (II.53) nous avons:

$$Yx + y = \tilde{a} + d \tag{II.56}$$

Où $d = (Y - W)x = Be$ est appelé le vecteur d'erreur de l'équation. Maintenant (II.56) peut être réécrite comme:

$$Be = Yx + y - \tilde{a} \tag{II.57}$$

Elle est non singulière et inversible parce que la matrice *B* est triangulaire inférieure et les éléments de la diagonale sont des 1. L'inverse de la matrice *B* est facile à obtenir:

$$F = B^{-1} = \begin{bmatrix} f_0 & & & & & & & \\ f_1 & f_0 & & & & & & \\ f_2 & f_1 & f_0 & & & & & \\ \cdot & \cdot & \cdot & \cdot & & & & \\ & & & \cdot & \cdot & & & \\ \cdot & \cdot & & & \cdot & \cdot & & \\ f_{N-2} & f_{N-3} & \cdot & \cdot & \cdot & f_1 & f_0 & \\ f_{N-1} & f_{N-2} & \cdot & \cdot & \cdot & f_2 & f_1 & f_0 \end{bmatrix} \quad \text{(II.58)}$$

Où :

$$f_0 = 1$$

et :

$$f_i = \begin{cases} -\sum_{i=1}^{j} b_i f_{j-i} & 1 \le j \le K \\ -\sum_{i=1}^{k} b_i f_{j-i} & K+1 \le j \end{cases}$$

Notons que f_i, sont les coefficients de filtre inverse tel que $1/B(z)=\sum_{i=0}^{\infty} f_i z^{-i}$, et sont calculés de manière récursive. En introduisant la matrice F, l'équation (II.57) devient :

$$e = F\left[Yx + y - \tilde{a}\right] = FYx + Fy - F_L a \quad \text{(II.59)}$$

Où F_L est la matrice constituée des K premières colonnes de F. Les Multiplications de la matrice FY et Fy sont rien d'autre que le filtrage inverse de la séquence de données et la construction de la matrice et du vecteur est de la même manière que Y et y, respectivement.

Soit : $\tilde{Y}$=FY et $\tilde{y}$=Fy, alors (II.59) devient :

$$e = \tilde{Y}x + \tilde{y} - F_L a \quad \text{(II.60)}$$

Parce que F (et ainsi F_L) est une fonction de b_i, la minimisation d'une somme de l'erreur quadratique $J = e * e$ est malheureusement un problème non linéaire. Nous utilisons l'astérisque (*) pour indiquer le conjugué de la transposé. Nous fixons F comme une matrice constante et mettons à zéro les dérivées de J par rapport à x et a, nous obtenons deux conditions:

$$F_L^* F_L a - F_L^* \tilde{Y} x = F_L^* \tilde{y} \tag{II.61}$$

$$-\tilde{Y}^* F_L a + \tilde{Y}^* \tilde{Y} x = -\tilde{Y}^* \tilde{y} \tag{II.62}$$

Pour un F fixe, le problème est linéaire en a et x. Une procédure itérative peut être utilisée pour minimiser J. en prenant $F^{(0)}$ comme matrice identité, nous obtenons La première estimation de $a^{(1)}$ et $x^{(1)}$. Avec ce $x^{(1)}$, on peut construire $F^{(1)}$ pour obtenir une nouvelle estimation de $a^{(2)}$ et $x^{(2)}$, et ainsi de suite. A chaque itération, l'estimation précédente des coefficients du dénominateur est utilisée pour obtenir de nouvelles estimations, de sorte que $a^{(m)}$ et $x^{(m)}$ sont trouvés en résolvant un système d'équations linéaires:

$$\begin{bmatrix} \left(F_L^{(m-1)}\right) * F_L^{(m-1)} & -\left(F_L^{(m-1)}\right) * \tilde{Y}^{(m-1)} \\ -\left(\tilde{Y}^{(m-1)}\right) * F_L^{(m-1)} & \left(\tilde{Y}^{(m-1)}\right) * \tilde{Y}^{(m-1)} \end{bmatrix} \begin{bmatrix} a^{(m)} \\ x^{(m)} \end{bmatrix} = \begin{bmatrix} \left(F_L^{(m-1)}\right) * \tilde{y}^{(m-1)} \\ -\left(Y^{(m-1)}\right) * \tilde{y}^{(m-1)} \end{bmatrix} \tag{II.63}$$

Pour m=1,2,…. Et $F^{(0)}$ = I. Si la convergence est obtenue, $J = e * e$ est minimisée. On peut vérifier que (II.63) est une forme explicite de la SMA [9,11] avec une entrée impulsionnelle. A chaque itération, il est nécessaire de résoudre un système d'équations linéaires 2K.

Comme le rang de la matrice est FL est K, la matrice FL*FL est hermitienne positive et inversible. De (II.61) :

$$a = \left(F_L^* F_L\right)^{-1} \left(F_L^* \tilde{Y} x + F_L^* \tilde{y}\right) \tag{II.64}$$

Substituons (II.64) dans (II.62), nous obtenons :

$$\tilde{Y}^* Q \tilde{Y} x = -\tilde{Y}^* Q \tilde{y} \tag{II.65}$$

Où $Q = I - F_L (F_L * F_L)^{-1} F_L^*$. Avec un *F* fixe, cette équation est linéaire en *x*. *A* chaque itération, nous avons besoin de résoudre un système de *K* équations linéaires seulement.

II.4. CONCLUSION

Le modèle *AR* basé sur une transmittance tous pôles assure une bonne représentation du signal vocale pour une valeur suffisante de *p*. Une représentation plus précise encore peut être obtenue par une modélisation basée sur une transmittace pôles-zéros, il s'agit du modèle *ARMA*.

Chapitre III

LA RECONNAISSANCE AUTOMATIQUE DU LOCUTEUR (RAL)

III.1. INTRODUCTION

La reconnaissance automatique du locuteur est interprétée comme une tâche particulière de reconnaissance de formes. Ce domaine regroupe les problèmes relatifs à l'identification ou à la vérification du locuteur sur la base de l'information contenue dans le signal acoustique: il s'agit de reconnaître une personne d'après sa voix. Le champ d'application est très vaste, allant des applications domestiques aux applications militaires, en passant par des applications judiciaires. L'authentification offre en effet de nombreuses applications potentielles comme sécurisation accrue des téléphones portables, contrôle supplémentaire au niveau d'une application sur un site comme l'accès sécurisé à un bâtiment ou remplacement du mot de passe sur les ordinateurs. Le principal avantage de cette technique est d'autoriser une authentification à distance. Ces applications concernent la vérification du locuteur à travers le réseau téléphonique pour accéder à un service (p.ex. validation de transactions par le téléphone) ou pour identifier un interlocuteur [12].

III.2. IDENTIFICATION AUTOMATIQUE DU LOCUTEUR

L'identification automatique du locuteur (*IAL*) a été une des premières utilisations de la *RAL*. En *IAL,* la liste des locuteurs à identifier est connue du système. Le système doit pouvoir décider, à partir d'un échantillon de voix, à quelle identité connue du système correspond l'échantillon [12].

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

En fonction de l'application, deux types de décisions sont possibles. Prenons le cas d'un centre d'appel téléphonique qui enregistre les conversations de ses employés. Un système d'*IAL* peut être utilisé pour classer chacun des enregistrements en fonction de l'employé concerné. Dans ce cas, la liste des locuteurs possibles est connue du système (tous les employés), et aucun autre locuteur non employé ne peut être concerné. Dans ce cas, l'application présuppose que l'ensemble des locuteurs possible est fermé et connu du système. Le système d'*IAL* choisira alors, parmi la liste, le modèle de locuteur le plus ressemblant à l'enregistrement de test. En revanche, dans une application utilisant un ensemble ouvert de locuteurs possibles, le système d'*IAL* ne connait pas tous les locuteurs possibles. Dans ce cas, en plus de déterminer le locuteur le plus vraisemblable, le système a la possibilité de rejeter l'échantillon de test en ne renvoyant aucune identité connue pour cet échantillon.

III.3. VERIFICATION AUTOMATIQUE DU LOCUTEUR

La vérification automatique du locuteur (*VAL*) permet de décider si l'identité revendiquée par un locuteur est compatible avec sa voix. Il s'agit donc de trancher entre deux hypothèses : soit le locuteur est bien le locuteur autorisé (on l'appelle aussi locuteur client), c'est-à-dire que son identité correspond à celle revendiqué, soit le locuteur est un imposteur qui cherche à se faire passer pour la personne qu'il n'est pas. À partir d'un échantillon de voix de référence, et d'un échantillon de voix de test, le système va donc devoir dire si oui ou non les deux locuteurs correspondent.

Les systèmes de *VAL* sont très dépendants des différences entre les échantillons de voix de référence et les échantillons de tests. Accepter un locuteur qui devrait être rejeté peut avoir de lourdes conséquences, en particulier dans les applications où un haut niveau de sécurité est demandé (contrôle aux frontières, système bancaire, identification judiciaire, etc.).

III.4. VARIABILITE DE LA VOIX

La variabilité d'une personne à une autre (variabilité inter-locuteurs) démontre les différences du signal de parole en fonction du locuteur. Cette variabilité, utile pour différencier les locuteurs, est également mélangée à d'autres types de variabilité : variabilité intra-locuteur, variabilité due aux conditions d'enregistrement et de transmission du signal de

parole (bruit ambiant, microphone utilisé, lignes de transmission) et variabilité due au contenu linguistique [12].

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

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

Une dégradation croissante des performances a été observée au fur et à mesure que le temps qui sépare la session d'enregistrement de la session de test augmente. De plus, le comportement des locuteurs se modifie lorsque ceux-ci s'habituent au système. Les modèles des locuteurs doivent donc être régulièrement mis à jour avec les nouvelles données d'exploitation du système. Les altérations de la voix dues à l'état physique (fatigue, rhume) ou émotionnel (stress), lorsqu'ils sont importants, peuvent mettre aussi en échec l'efficacité de certains systèmes.

III.5. DEPENDANCE AU TEXTE

On parle de reconnaissance vocale en mode dépendant du texte lorsque le texte prononcé par le locuteur est fixé et connu à l'avance. A l'opposé, lorsque le texte prononcé par le locuteur n'est pas connu à priori, on parle de mode indépendant du texte. Mais cette terminologie ne rend pas bien compte des différentes dépendances au texte possible. Les différents systèmes peuvent être classés, selon le degré croissant d'indépendance au texte, de la façon suivante [12]:

- Système à texte fixé dépendant du locuteur: pour un locuteur donné, le texte est toujours le même d'une session à l'autre. Mais chaque locuteur a un texte différent.

- Système dépendant du vocabulaire: l'utilisateur du système prononce une séquence démos, issues d'un vocabulaire limité (des séquences de chiffres par exemple), mais dont l'ordre peut varier d'une session à l'autre.

- Système dépendant d'événements phonétiques: le vocabulaire n'est pas directement imposé, mais certains événements phonétiques doivent être présents dans la séquence de parole prononcée (p.ex. présence de certaines voyelles ou nasales). Les phrases à prononcer peuvent éventuellement être affichées sur l'écran à chaque session.

- Système à texte imposé par la machine: le texte est différent pour chaque session et pour chaque locuteur, mais affiché à chaque fois par la machine. Le texte est choisi de manière imprédictible pour éviter l'utilisation d'enregistrements par un imposteur.

- Système indépendant du texte: le locuteur est entièrement libre de ce qu'il dit à chaque session.

Les systèmes dépendants du texte donnent généralement de meilleures performances de reconnaissance que les systèmes indépendants du texte car la variabilité due au contenu linguistique de la phrase prononcée est alors neutralisée

III.6. MODELISATION DES LOCUTEURS

De manière à modéliser des caractéristiques qui dépendent du locuteur, nous utilisons des algorithmes capables de capturer les points communs entre différentes représentations de motifs spectraux issus du même locuteur (constituant ainsi un modèle du locuteur), tout en ayant la possibilité de s'adapter aux variations d'échelles fréquentielles et temporelles liées au signal de parole. Ces motifs peuvent être soit des segments de parole déterminés (mots, phonèmes) si nous travaillons en mode dépendant du texte, soit des segments de parole dont on ne connaît pas le contenu phonétique si l'application fonctionne en mode indépendant du texte. Ces algorithmes doivent être couplés avec une mesure qui permettra de donner une valeur de distance (ou de similitude) entre le modèle du locuteur et un motif inconnu dont on cherche à déterminer la provenance. Nous proposons ici un aperçu des méthodes déterministes (comparaison dynamique et quantification vectorielle) et statistiques (modèles à

mélange de distributions gaussiennes et modèles de *Markov* cachés) qui fournissent les meilleurs résultats en reconnaissance vocale [13].

Le problème de reconnaissance du locuteur peut se formuler selon un problème de classification. Différentes approches ont été développées, néanmoins on peut les classer en trois grandes familles:

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

III.6.1. L'APPROCHE CONNEXIONNISTE

Les réseaux de neurones ont été assez largement utilisés en reconnaissance du locuteur. Ils offrent en effet une bonne alternative au problème de la discrimination entre les locuteurs. Ces outils de classification permettent de séparer des classes, dans un espace de représentation donné, de façon non linéaire. L'inconvénient important de l'application de cette technique en identification du locuteur est le coût important lié à l'ajout d'un nouveau locuteur dans la base de référence (ce n'est pas le cas en vérification du locuteur). On peut aussi utiliser les réseaux de neurones en les couplant à d'autres techniques, comme par exemple les modèles de *Markov* cachés. On parle alors de méthodes hybrides [14] [15].

III.6.2. L'APPROCHE VECTORIELLE

III.6.2.1. RECONNAISSANCE DU LOCUTEUR A BASE DTW

La reconnaissance par l'alignement temporel par programmation dynamique *DTW* (*Dynamique Time Warping)* repose sur le principe que chaque mot est représenté par une

prononciation de référence. Compte tenu des décalages temporels entre les différentes prononciations d'un même mot, l'algorithme met en correspondance des séquences de paramètres par distorsion temporelle (*Time warping)*. La programmation dynamique permet d'aligner temporellement une phrase de test avec une phrase d'apprentissage ce qui signifie que c'est une technique exclusivement utilisée en mode dépendant du texte [16][17].

Ce type de technique a été peu à peu abandonné au profit des modèles séquentiels statistiques (comme les modèles de Markov cachés) qui sont moins "rigides" et donc plus robustes vis à vis de la variabilité inhérente au signal de parole.

III.6.2.2. QUANTIFICATION VECTORIELLE

La quantification vectorielle est une méthode non-paramétrique qui permet de décrire un ensemble de données par un faible nombre de vecteurs formant un dictionnaire associé aux données. Le dictionnaire est en général calculé de telle façon que la distance moyenne entre un vecteur issu des données et son plus proche voisin dans le dictionnaire soit la plus petite possible. La quantification vectorielle est une technique de groupage qui est d'autant plus adaptée que les données présentent naturellement des "points d'accumulation" autour desquels la densité de vecteurs issus des données est importante [16][18].

Pour la reconnaissance du locuteur, la mesure de similarité entre deux ensembles de vecteurs acoustiques consiste à évaluer la distance moyenne d'un des deux ensembles de vecteurs acoustiques en utilisant le dictionnaire optimisé pour l'autre ensemble de vecteurs acoustiques par quantification vectorielle.

La caractérisation de la distribution des données obtenue par la quantification vectorielle est en fait voisine de celle fournie par un modèle de mélange de distributions gaussiennes. Les performances des deux types de systèmes sont donc assez proches. Lorsque les données disponibles pour l'apprentissage sont suffisantes, il semble que le modèle de mélange de densités gaussiennes soit plus robuste.

III.6.3. L'APPROCHE STATISTIQUE

III.6.3.1. MESURES STATISTIQUES DU SECOND ORDRE

Cette partie présente une famille de mesures de similarité entre locuteurs. Ces mesures reposent sur les caractéristiques du second ordre d'une séquence de vecteurs, c'est-a-dire sur le vecteur moyen et la matrice de covariance de cette séquence. Plusieurs mesures de distance ont été utilisées, on peut citer : le rapport de vraisemblance, la distance de *Kullbak- Leibler,* maximum de vraisemblance, test de sphéricité, déviation absolue des valeurs propres. Ces mesures donnent des résultats très encourageants sur la parole propre, et, naturellement, voient leurs performances se dégrader sur la parole téléphonique. De part leur relative simplicité, ces mesures peuvent également servir de référence pour évaluer la qualité d'une base de données.

III.6.3.2. LES MELANGES DE GAUSSIENNES

La reconnaissance du locuteur par mélanges de gaussiennes (*ou GMM pour Gaussian Mixture Models)* consiste à modéliser un locuteur par une somme pondérée de composantes gaussiennes. Ainsi une large gamme de distribution peut être parfaitement représentée. Chaque composante des gaussiennes est supposée modéliser un ensemble de classes acoustiques. L'utilisation de ce type de modèle semble être bien prometteuse. Il semble bien modéliser les caractéristiques spectrales des voix des locuteurs, et il est relativement simple à mettre en œuvre. On peut assimiler un modèle *GMM* à un *HMM* à un seul état, on ne modélise donc pas les aspects temporels du signal. Cette méthode est la plus utilisée en reconnaissance du locuteur en mode indépendant du texte [19][20].

III.6.3.3. MODELES DE MARKOV CACHES

Les modèles de Markov (*ou HMM pour Hidden Markov Models)* ont été initialement introduits en reconnaissance de la parole. Puis leur utilisation s'est étendue peu à peu au domaine de reconnaissance du locuteur. Dans cette approche, il ne s'agit plus d'une mesure de distance d'une forme acoustique à une référence, mais de la probabilité que la forme acoustique ait été engendrée par le modèle de référence du locuteur. Le modèle d'un locuteur est constitué de l'association d'une chaîne de Markov, une succession d'états avec des

probabilités de transition d'un état à l'autre, et de lois de probabilités (probabilités d'observation d'un vecteur acoustique dans un état).

Les propriétés statistiques des modèles de Markov cachés en font une des modélisations les plus efficaces actuellement en reconnaissance du locuteur dépendante du texte. Les *HMM* permettent de modéliser des processus stochastiques variant dans le temps. Pour cela, ils combinent les propriétés à la fois des distributions de probabilités et d'une machine à états [21-24].

III.7. MESURE DES PERFORMANCES

L'identification du locuteur : consiste à reconnaître un locuteur parmi un ensemble de locuteurs en comparant son identité vocale à des références connues. Les performances du système d'identification sont données en termes de taux d'identification correcte I_c ou incorrecte I_i soit : [13]

$$I_c = \frac{nombre \quad de \quad testes \quad correctement \quad identifiées}{nombre \quad total \quad de \quad tentatives}$$

Et

$$I_i = \frac{nombre \quad de \quad testes \quad mal \quad identifiées}{nombre \quad total \quad de \quad tentatives}$$

Avec:

$$I_c + I_i = 100\%$$

La vérification du locuteur : consiste, après que le locuteur a décliné son identité, à vérifier l'adéquation du message vocale avec la référence acoustique du locuteur qu'il prétend être. C'est une décision en tout ou rien. Les performances de vérification de locuteur sont données en termes des taux de faux rejets *FR* et du taux de fausses acceptations *FA* et de l'erreur moyenne EA : [12,13]

$$FR = \frac{nombre \quad de \quad tentatives \quad d'abonnés \quad rejetées}{nombre \quad total \quad de \quad tentatives \quad d'abonnés}$$

$$FA = \frac{nombre \quad \text{de} \quad \text{t}entative \quad d'imposteurs \quad acceptés}{nombre \quad totale \quad \text{de} \quad \text{t}entative \quad d'abonnés}$$

$$EM = \frac{FA + FR}{2}$$

Ces deux types d'erreurs n'ont pas toujours la même incidence en termes de sécurité et de 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 tentant 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.

Nous nous intéresserons ici au système de reconnaissance en mode **vérification du locuteur dépendante du texte**. Plusieurs raisons motivent ce choix [25]:

- Le mode de vérification est plus pratique à utiliser lors d'applications à grande échelle car il ne nécessite pas de connaître les autres clients de l'application, ni d'opérer des tests sur tous les clients enregistrés. Il peut donc être considéré comme indépendant du nombre de clients inscrits dans une application.
- Ce mode est plus à même de modéliser les phénomènes d'imposture que le mode d'identification. En effet, en identification dans un ensemble fermé, on cherche à déterminer le client de l'application qui a la probabilité la plus grande d'avoir prononcé la phrase observée. Cependant, dès l'instant où les imposteurs ne se trouvent pas dans la base des clients, il faut passer en identification dans un ensemble ouvert, combinant ainsi les phénomènes d'imposture et d'identification.
- La dépendance au texte permet de mieux contrôler le message prononcé, vu le mélange complexe entre les informations linguistiques et celles dépendantes du locuteur contenues dans le signal de parole.
- La dépendance au texte permet une modélisation plus précise du contenu linguistique et donc de meilleures performances du système de reconnaissance du locuteur.

Des études récentes montrent que la haute performance pour la vérification du locuteur dépendante du texte peut être obtenue en utilisant l'approche HMM [19,20,26-28].

III.8. MODELES DE MARKOV CACHES (*HMM*)

Les vecteurs acoustiques vont servir d'observations dans les Modèles de Markov cachés *HMM*. Le but des *HMM* est de trouver la meilleure séquence de mots d'un lexique qui définit les mots reconnaissables et d'une grammaire qui détermine les séquences de mots valables ou, du moins, les plus probables.

Un *HMM* est un ensemble d'états et de transitions les reliant. L'appellation modèle de Markov caché provient du fait que le chemin emprunté par un processus aléatoire, modélisé par un *HMM*, est inconnu car les états parcourus ne sont pas directement observables [29].

La structure d'un HMM (*figure III.1*) est définie par trois paramètres principaux [30,31] :

1. La matrice de la distribution initiale des états, $\pi = (\pi_i)$ où π_i est la probabilité d'être dans l'état q_i à l'instant initial.
2. La matrice des probabilités de transition, $A = (a_{ij})$ où a_{ij} est la probabilité de passer de l'état q_i `a l'état q_j. L'égalité suivante est toujours vérifiée :

$$\sum_{j \in s} a_{ij} = 1 \qquad i = 1, 2, \ldots S \qquad \text{(III.1)}$$

3. La matrice des probabilités d'´emission des observations définissant l'ensemble des lois d'´emission, $B = (b_i(o_t))$ où $b_i(o_t)$ est la distribution de probabilité d'être dans l'´etat q_i et d'´emettre l'observation o_i. Ces distributions sont souvent de type gaussien.

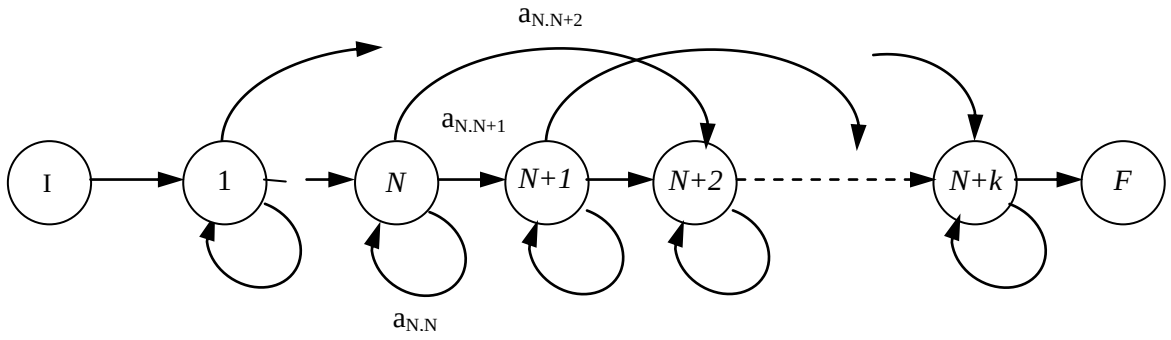


Figure III.1 : La structure *HMM* Left-to-right [24,25,30-32].

Avec ces outils. Un système *HMM* doit répondre aux questions constituant les trois problématiques du processus de reconnaissance de la parole.

Etant donné la séquence d'observations O = (o1, o2, ..., oT) et un *HMM* $\lambda = (A,B, \pi)$ [29,30,32] :

1. Comment calculer P(O | λ), la probabilité de la séquence d'observations, étant donné le modèle *HMM* λ ? (modélisation acoustique).
2. Quelle est la séquence d'états Q = (q1, q2, ..., qT) qui est la plus vraisemblable étant donné la séquence d'observations O? Ce problème correspond au processus de reconnaissance.
3. Comment ajuster les paramètres du modèle *HMM* λ pour maximiser la probabilité P(O | λ)?. Ce problème correspond au processus d'apprentissage.

La reconnaissance de la parole à base des *HMM* est une modélisation stochastique dont l'objectif est de trouver, parmi toutes les séquences de mots W possibles, la séquence de mots $\hat{W}$ la plus probable connaissant les observés O.

$$\hat{W} = \arg\max P(W / O) \tag{III.2}$$

La probabilité P (W / O) est une probabilité dont le calcul repose sur une modélisation du canal acoustique qu'on ne peut pas calculer directement. Cependant, une réécriture ou simplification probabiliste, telle une décision bayesienne, permet de décomposer cette probabilité en l'exprimant autrement. En effet, grâce à la formule de Bayes appliquée à la probabilité P(W | O) (équation III.3), on exprime le problème, cette fois-ci, comme une recherche de la suite de mots W maximisant la probabilité a priori P(W) de leur apparition dans la langue (modélisation linguistique) et que les paramètres acoustiques observés correspondent à cette suite de mots (modélisation acoustique), P(O | W). La formule finale (équation III.4) ne fait pas intervenir P(O), la probabilité d'occurrence de la chaîne acoustique O, car elle est indépendante de W et reste constante quand W varie [32].

$$\arg\max P(W / O) = \arg\max \frac{P(o,W)}{p(O)} = \arg\max \frac{P(W)P(O / W)}{P(O)} \tag{III.3}$$

$$= \arg\max P(W)P(O / W) \tag{III.4}$$

L'approche stochastique permet ainsi d'intégrer les niveaux acoustiques et linguistiques dans un seul processus de décision. Ce processus consiste à chercher le chemin optimal correspondant à la séquence d'état la plus probable au sens de la probabilité de vraisemblance

de la séquence d'observations. Ceci est effectué généralement par l'algorithme de Viterbi qui délivre également la probabilité de vraisemblance sur le meilleur chemin.

III.9. CONCLUSION

D'après ce qu'on vient de voir, la modélisation Markovienne repose sur un formalisme à la fois simple et rigoureux. A ces deux points forts viendra s'ajouter un troisième qui est la robustesse comme on le verra à travers les résultats de reconnaissance exposés dans le dernier chapitre.

Chapitre IV

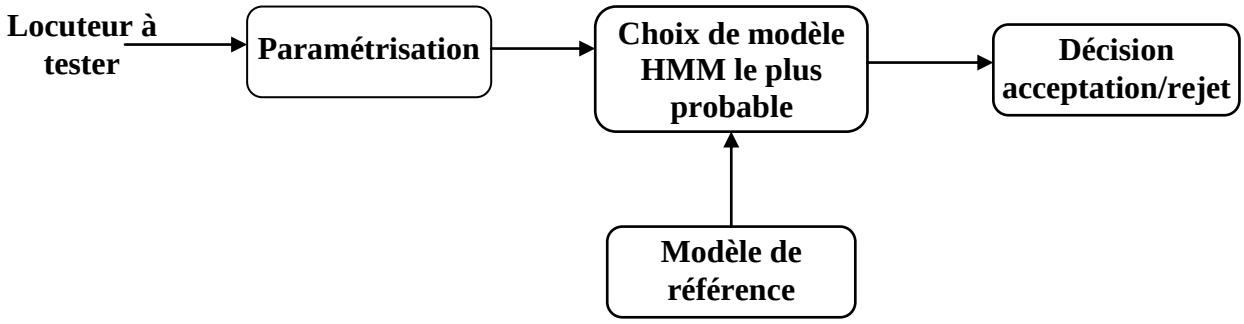
APPLICATION ET RÉSULTATS

IV.1. INTRODUCTION

Dans ce chapitre, nous allons détailler la partie pratique de notre sujet. Pour la réalisation de notre travail, il suffit de faire une bonne acquisition des données et programmer dans MATLAB la phase d'apprentissage et de reconnaissance pour les modélisations: *AR*, *ARMA*, *MFCC* (Mel Frequency cepstral coefficient), *LPCC* (linear prediction cepstral coeffcient), *CARMA* (Cepstral *ARMA*) pour ces systèmes nous avons utilisé les *HMM* comme des reconnaisseurs.

IV.2. DESCRIPTION DU SYSTEME DE VERIFICATION DU LOCUTEUR

1.Phase d'apprentissage



2.Phase de reconnaissance

Figure IV.1 : Les différentes phases du système de vérification du locuteur par *HMM*.

La vérification du locuteur se décompose en deux phases. *(figure IV.1)*. La première phase consiste à obtenir une représentation de l'utilisateur. Elle est appelée apprentissage. Cette phase joue un rôle essentiel dans le processus de vérification.

Lors de cette phase, le système construit une représentation de référence du client, représentation qui sera utilisée par la suite pour autoriser ou non l'accès au service. La deuxième phase est la phase de test. Des données provenant d'un utilisateur souhaitant être vérifié sont soumises au système. Cet utilisateur annonce le mot de passe connue du système. Le test consiste à mesurer la ressemblance entre les données fournies par l'utilisateur et le modèle de référence existant correspondant à l'identité annoncée.

IV.2.1. PHASE D'APPRENTISSAGE

IV.2.1.1. ACQUISITION DU SIGNAL DE PAROLE

L'acquisition du signal parole a été faite grâce au logiciel d'interface Goldwave qui nous a permis d'enregistrer, traiter et visualiser le signal numérisé de la parole sur l'écran.

Le signal parole été échantillonné avec une fréquence d'échantillonage de 10 *kHz* (spectre utile jusqu'à 5 *kHz*), codé en entier sur 16 bits, sur le canal mono, puis stocké sur le disque dans des fichiers d'extension .wav. (*figure IV.2*)

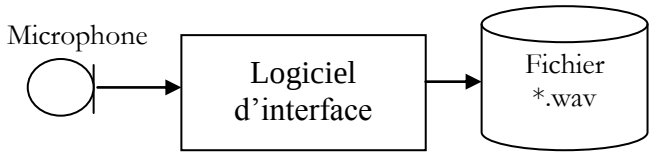


Figure IV.2 : Schéma descriptif de la phase d'acquisition.

IV.2.1.2. ANALYSE ACOUSTIQUE

Le signal parole étant acquit, on procède à l'analyse acoustique, son objectif est de réduire la redondance du signal de parole, en ne conservant qu'un ensemble de paramètres pertinents parmi toutes les données disponibles dans le but de diminuer la quantité de calcul et de stockage lors du traitement d'apprentissage et de reconnaissance.

Les paramètres qu'on a utilisés sont les coefficients issus des différentes modélisations *AR*, *ARMA*, *MFCC*, *LPCC*, *CARMA* dans le but de les comparer et révéler la meilleure méthode pour modéliser l'information acoustique servant à la vérification du locuteur. La *figure (IV.3)* représente l'organigramme de toute la phase d'analyse acoustique.

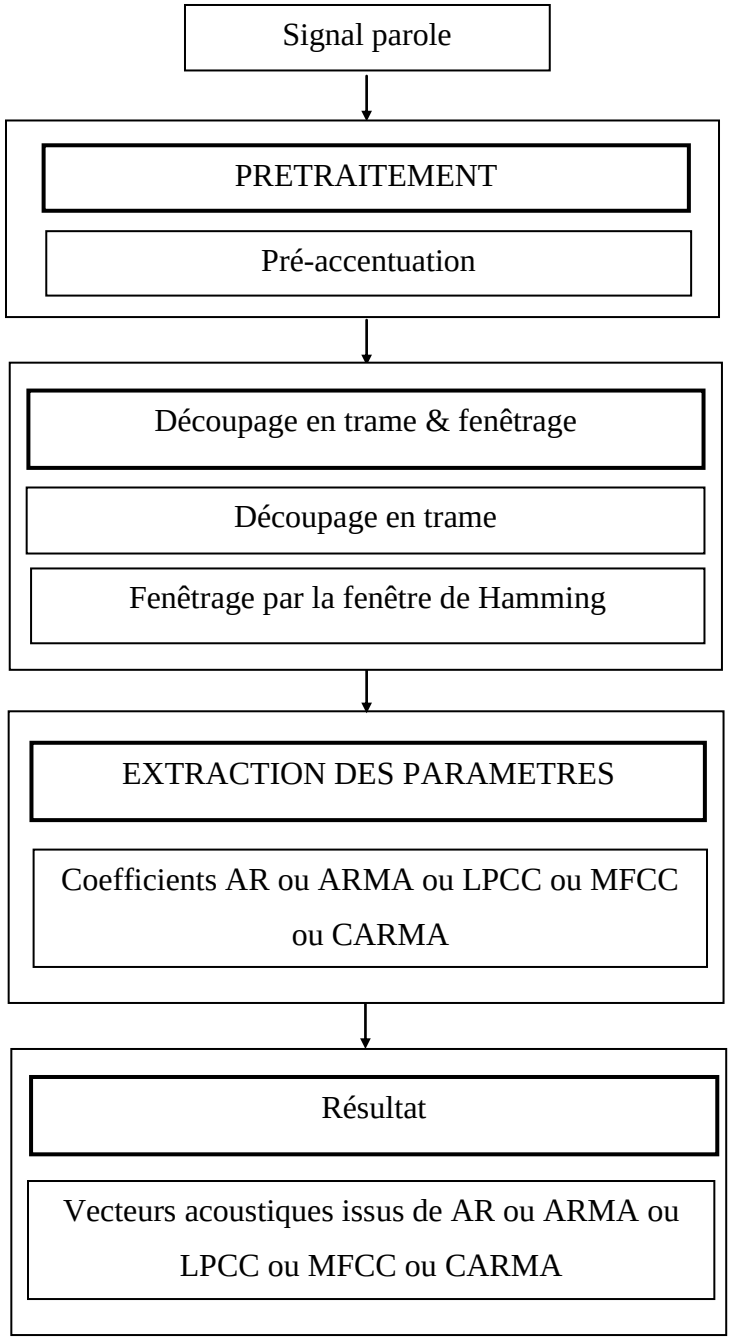


Figure IV.3 : Organigramme du module de l'analyse acoustique.

- **Pré-accentuation**

Le signal parole subit une opération de préaccentuation, qui consiste en un filtrage de type passe haut qui relève le niveau des aigus.

En pratique, on utilise simplement un filtre de réponse impultionnelle finie

$H(z) = 1 - az^{-1}$ *avec* $0.9 < a < 1.0$ Où *a* est généralement égal à 0.95.

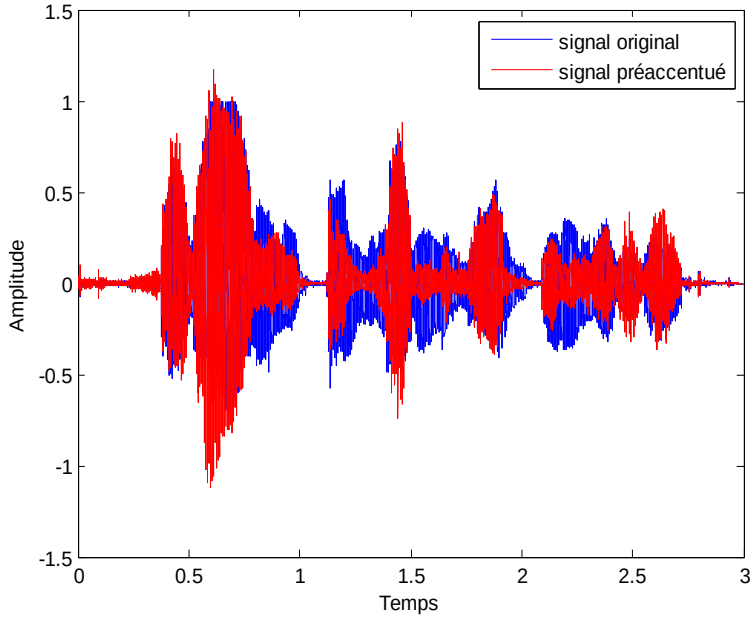


Figure IV.4 : Signal original et pré-accentué.

La *figure (IV.4)* nous montre le signal original de la parole et le signal pré-accentué.

- **Découpage en trame et fenêtrage**

Le signal de parole pré-accentué est ensuite segmenté en trames dont la durée est de *30 ms* décalés de *10 ms*. Chaque trame correspond à une portion sur laquelle le signal de parole peut être considéré comme stationnaire.

Ensuite, on applique une fenêtre de *Hamming* (*figure IV.5)* qui a pour fonction d'atténuer le signal au début et à la fin de chaque trame. Contrairement à d'autres fenêtres, la fenêtre de *Hamming* ne s'annule pas à ses extrémités, ne présente pas de coupure brusque et réduit les effets de bord. La *figure (IV.6)* le fenêtrage d'une seule trame.

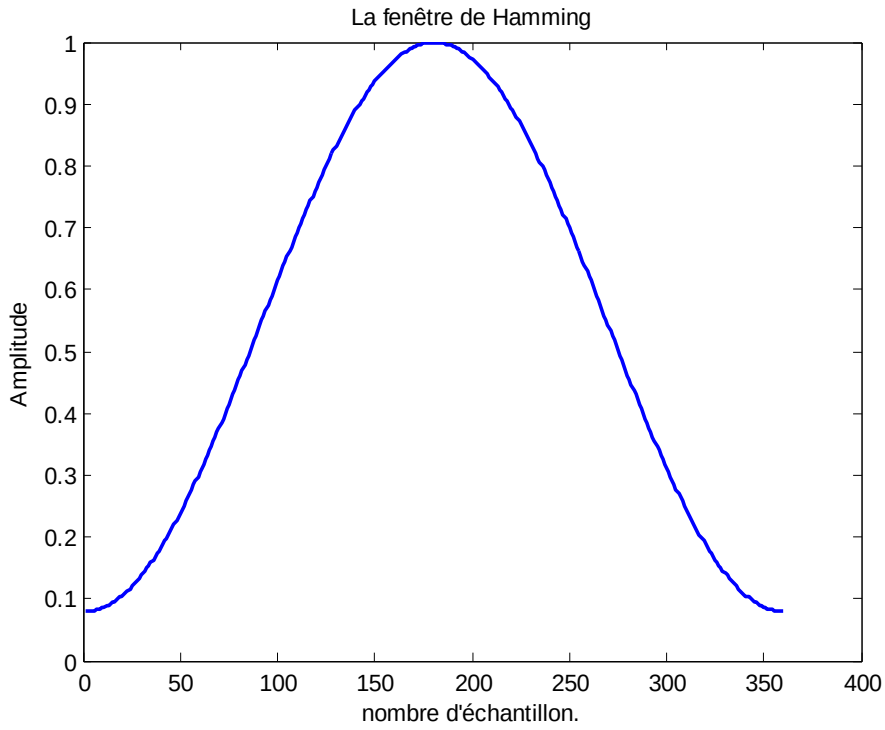


Figure IV.5 : Fenêtre de Hamming.

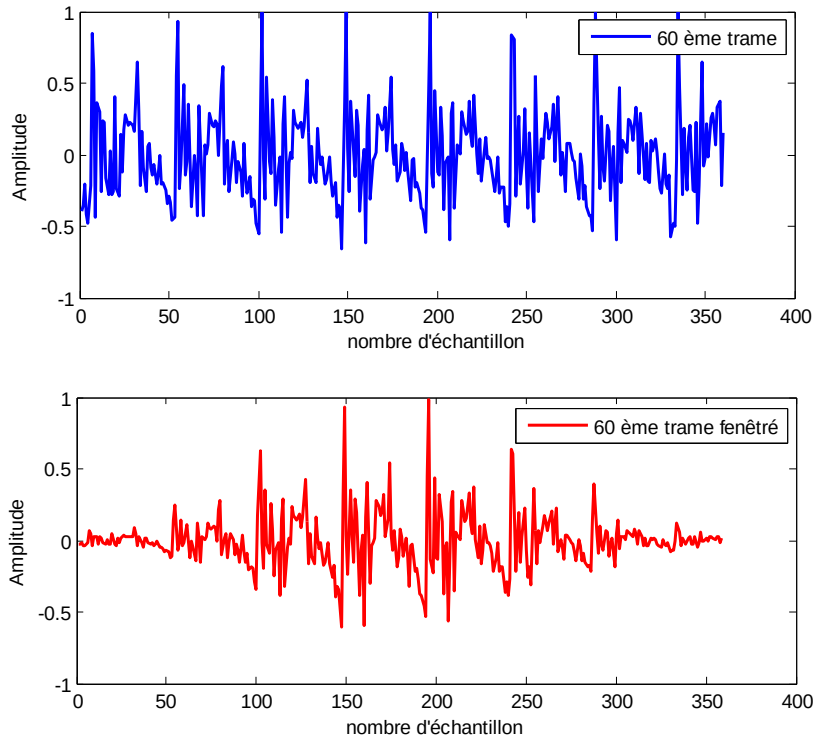


Figure IV.6: signal fenêtré.

- **Extraction des paramètres**
  - ✓ *Extraction des coefficients AR (coefficients des pôles)*

La détermination des coefficients de prédiction linéaire est faite par l'algorithme de *Levinson-Durbin*. On choisit premièrement l'ordre de prédiction *P* lorsque on a *Fe*=10*Khz,* donc p est entre 12 et 14. On prend P=12.

Les paramètres estimés d'une trame sont montrés dans le tableau IV.1

Les paramètres AR estimés d'une trame	
a_1	**0,63994**
a_2	**-1,1415**
a_3	**1,0650**
a_4	**-1,4658**
a_5	**0,8869**
a_6	**-1,1492**
a_7	**1,1431**
a_8	**-1,2043**
a_9	**0,8195**
a_{10}	**-0,8823**
a_{11}	**0,5235**
a_{12}	**-0,2018**

Tableau IV.1 : Coefficients des pôles**.**

  - ✓ *Extraction des coefficients cepstraux LPCC*

Les coefficients cepstraux LPCC sont tirés des coefficients AR (ou LPC) à partir de la formule suivante [16,33,34] :

$$c_1 = a_1$$
$$c_n = \sum_{k=1}^{n-1} (1 - k/n) a_k c_{n-k} + a_n \qquad 1 < n \leq p \tag{IV.1}$$

Les paramètres estimés d'une trame sont montrés dans le tableau IV.2

Les paramètres LPCC estimés d'une trame	
c_1	**0,6399**
c_2	**-0,9367**
c_3	**0,8215**
c_4	**-1,2954**
c_5	**0,6992**
c_6	**-1,0546**
c_7	**1,0380**
c_8	**-1,1129**
c_9	**0,7339**
c_{10}	**-0,8298**
c_{11}	**0,4722**
c_{12}	**-0,1739**

Tableau IV.2 : Les coefficients LPCC**.**

✓ *Extraction des coefficients ARMA (coefficient des pôles et des zéros)*

Pour chaque trame la détermination des coefficients *a(i)* et *b(i)* est faite par l'algorithme de *Steiglitz Mc Bride*. On choisit l'ordre *p et q* égal à 12 pour pouvoir effectuer des comparaisons entre les résultats de *AR* et *ARMA*.

Les paramètres estimés d'une trame sont montrés dans le tableau (IV.3)

Les paramètres ARMA estimés d'une trame			
a_1	-0,3846	b_1	-0,0913
a_2	1,577	b_2	0,2129
a_3	-0,8397	b_3	-0,2618
a_4	2,0330	b_4	0,4664
a_5	-0,6028	b_5	-0,3098
a_6	1,9568	b_6	0,5472
a_7	-0,8604	b_7	-0,3391
a_8	1,8976	b_8	0,4955
a_9	-0,6384	b_9	-0,3955
a_{10}	1,5560	b_{10}	0,4898
a_{11}	-0,3783	b_{11}	-0,2954
a_{12}	0,5536	b_{12}	0,2362

Tableau IV.3 : Coefficients des pôles et des zéros.

✓ *Extraction des paramètres CARMA*

Les coefficients cepstraux *CARMA* sont tirés comme leur nom l'indique à partir des coefficients *ARMA* comme montré sur la figure (IV.7).

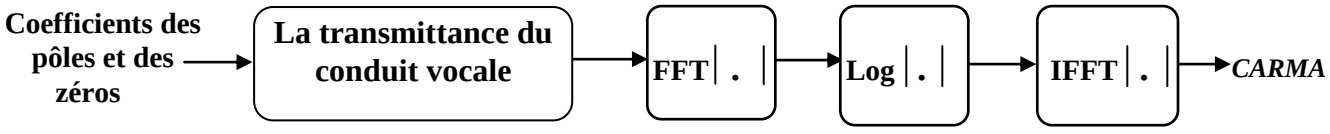


Figure IV.7 : Chaine de calcul des paramètres *CARMA*.

Les paramètres *CARMA* estimés d'une trame sont montrés dans le tableau (IV.4).

Les paramètres CARMA estimés d'une trame	
c_1	**-0,2252**
c_2	**-0,1935**
c_3	**0,2171**
c_4	**-0,0909**
c_5	**-0,0233**
c_6	**0,1817**
c_7	**0,0557**
c_8	**-0,1134**
c_9	**-0,0208**
c_{10}	**0,1277**
c_{11}	**0,0412**
c_{12}	**-0,0602**

Tableau IV.4 : Les coefficients *CARMA*.

✓ *Extraction des coefficients ceptraux MFCC*

La procédure de calcul des coefficients *MFCC* est présentée dans la figure (IV.8)

Figure IV.8 : Chaine de calcul des coefficients MFCC [32,35,36].

Les coefficients *MFCC* estimés d'une trame sont montrés dans le tableau (IV.5).

Les paramètres MFCC estimés d'une trame	
c_1	**-38,8487**
c_2	**-18,2322**
c_3	**-10,6885**
c_4	**-8,2634**
c_5	**-1,7561**
c_6	**-13,3270**
c_7	**-1,0122**
c_8	**-8,7932**
c_9	**-6,5563**
c_{10}	**-0,8136**
c_{11}	**0,3585**
c_{12}	**0,2791**

Tableau IV.5: Les coefficients *MFCC*.

IV.2.1.3. MODULE DE PRODUCTION DES MODELES HMM

Il a pour but de générer les différents modèles *HMM* adoptés pour chaque locuteur du corpus.

La *figure (IV.7)* représente l'organigramme global de la phase d'apprentissage

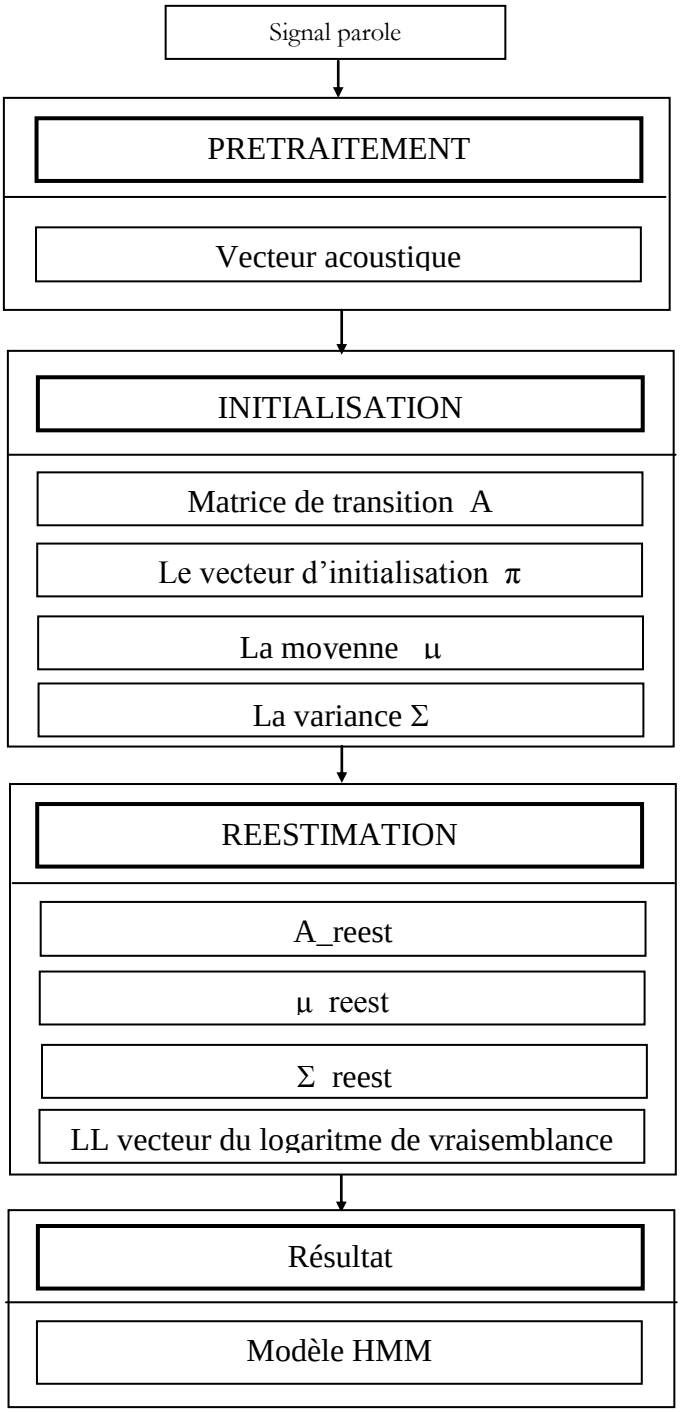


Figure IV.9: Organigramme du module de production des modèles *HMM*.

L'étape finale de la phase d'apprentissage est l'extraction des paramètres du modèle.

Chaque modèle est présenté par les variables suivantes :

- Probabilité de la matrice initiale
- La matrice de Transition A_réest

- La matrice de La moyenne μ_réest ;
- Le vecteur du logarithme de vraisemblance.

IV.2.1.4. RESULTATS D'APPRENTISSAGE

Nous allons présenter dans les résultats suivant seulement les variables de modélisation *AR* et *ARMA*

✓ *Cas de modélisation AR*

- **Probabilité de la matrice initiale**

$$\pi = \begin{bmatrix} 0,2 & 0,2 & 0,2 & 0,2 & 0,2 \end{bmatrix}$$

- **Matrice de transition entre les états**

$$A = a_{ij} \begin{vmatrix} 0,93026221 & 0,05964112 & 8,15e-05 & 0,01001519 & 1,19e-20 \\ 0,02364027 & 0,83836165 & 0,09908462 & 0,03891346 & 4,22e-46 \\ 0,07613294 & 0,0409509 & 0,85427082 & 0,02864534 & 1,64e-45 \\ 0,0571647 & 5,92e-05 & 3,00e-31 & 0,92857443 & 0,01420171 \\ 4,71e-26 & 1,67e-24 & 3,23e-42 & 0,07797853 & 0,92202147 \end{vmatrix}$$

- **Matrice des moyennes pour la première gaussienne**

Nombre de coefficients AR	Etat 1	Etat 2	Etat 3	Etat 4	Etat 5
1	***0.2692***	***-0.7554***	***-0.2818***	***-0.0514***	***-0.1154***
2	***-0.4198***	***-1.5716***	***-0.5982***	***-0.2777***	***-0.4771***
3	***0.6982***	***-0.3256***	***0.8938***	***0.4303***	***0.0601***
4	***-0.5245***	***-1.0647***	***0.0065***	***-0.3987***	***-0.7984***
5	***0.5851***	***0.2312***	***0.8039***	***0.2492***	***0.1799***
6	***-0.5849***	***-0.9476***	***-0.8003***	***-0.2971***	***-0.3006***
7	***0.5782***	***0.3795***	***0.0168***	***0.3255***	***0.2970***
8	***-0.5407***	***-0.8275***	***-0.8448***	***-0.1701***	***-0.3109***
9	***0.4044***	***0.3958***	***0.6581***	***0.1914***	***0.2183***
10	***-0.5325***	***-0.5832***	***0.0260***	***-0.1673***	***-0.1752***
11	***0.4815***	***0.3306***	***0.6787***	***0.1188***	***0.2278***
12	***-0.4115***	***-0.4128***	***-0.3681***	***-0.1587***	***-0.0515***

Tableau IV.6 : Matrice des moyennes pour la première gaussienne (cas de *AR*)

- **Matrice des moyennes pour la deuxième gaussienne**

Nombre de coefficients AR	Etat 1	Etat 2	Etat 3	Etat 4	Etat 5
1	***0.4102***	***0.7806***	***0.7213***	***-1.3622***	***-0.8929***
2	***-0.8434***	***-0.9017***	***-1.0895***	***-1.1249***	***-0.7720***
3	***0.8813***	***1.1977***	***1.3316***	***-0.6708***	***-0.2550***
4	***-1.0542***	***-1.2532***	***-1.6861***	***-0.8902***	***-0.6545***
5	***0.8989***	***0.8227***	***1.5409***	***-0.8434***	***-0.2355***
6	***-1.0284***	***-1.3561***	***-1.8783***	***-0.5854***	***-0.1625***
7	***1.0102***	***1.1909***	***1.8794***	***-0.3229***	***0.1524***
8	***-1.0669***	***-1.0036***	***-1.8498***	***-0.2448***	***-0.1165***
9	***0.8311***	***0.9938***	***1.6382***	***-0.3897***	***-0.0405***
10	***-0.9550***	***-0.7826***	***-1.4633***	***-0.3901***	***-0.0706***
11	***0.6455***	***0.4823***	***1.1155***	***-0.2220***	***0.1485***
12	***-0.6115***	***-0.6032***	***-1.0592***	***-0.1394***	***0.0551***

Tableau IV.7 : Matrice des moyennes pour la deuxième gaussienne (cas de *AR*)

✓ *Cas de modélisation ARMA :*

- **Probabilité de la matrice initiale**

$$\pi = \begin{bmatrix} 0,2 & 0,2 & 0,2 & 0,2 & 0,2 \end{bmatrix}$$

- **Matrice de transition entre les états**

$$A = a_{ij} \begin{vmatrix} 0,67347528 & 0,32606978 & 5,91e-24 & 5,80e-196 & 0,00045494 \\ 0,23580478 & 0,57957923 & 4,02e-19 & 2,77e-215 & 0,18461599 \\ 0,00022035 & 0,00031306 & 0,5943492 & 0,1118416 & 0,2932758 \\ 2,86e-196 & 3,23e-105 & 0,2140596 & 0,68181818 & 0,10413586 \\ 0,01255029 & 0,06399662 & 0,07300749 & 0,01465119 & 0,83579442 \end{vmatrix}$$

- **Matrice des moyennes pour la première gaussienne**

Nombre de coefficients AR	Etat 1	Etat 2	Etat 3	Etat 4	Etat 5
1	-0.0080	-0.0014	-0.0003	0.0001	0.0028
2	0.0218	-0.0003	-0.0003	0.0006	0.0145
3	-0.0499	0.0025	-0.0096	0.0017	0.0534
4	0.0399	-0.0155	-0.0112	0.0058	0.1338
5	-0.1217	0.0170	-0.0211	0.0140	0.2687
6	0.0616	-0.0262	-0.0061	0.0258	0.4257
7	-0.1475	0.0329	-0.0120	0.0325	0.5670
8	0.1138	-0.0295	0.0013	0.0312	0.6244
9	-0.1243	0.0236	-0.0169	0.0218	0.5955
10	0.1160	-0.0316	0.0065	0.0109	0.4802
11	-0.1181	0.0200	0.0093	-0.0004	0.3437
12	0.0607	-0.0283	0.0416	-0.0087	0.2114
13	-0.1261	0.0189	0.0308	-0.0121	0.1214
14	0.0077	-0.0333	0.0327	-0.0095	0.0617
15	-0.0783	0.0060	0.0103	-0.0040	0.0317
16	0.0135	-0.0154	0.0065	-0.0006	0.0127
17	-0.0079	0.0070	-0.0034	0.0005	0.0048
18	0.3648	-0.8853	1.9454	5.1029	5.6664
19	2.9028	1.7249	3.2635	14.8626	19.3215
20	0.1630	-2.0800	2.8609	29.5567	46.2752
21	4.5082	3.0721	3.2958	46.7495	88.1614
22	-0.3468	-2.7354	1.7015	62.9762	139.3180
23	5.1883	3.1786	1.4303	76.5896	189.7105
24	-1.0659	-3.4771	-0.0865	83.5385	224.3816

Tableau IV.8 : Matrice des moyennes pour la première gaussienne (cas de *ARMA*)

- **Matrice des moyennes pour la deuxième gaussienne**

Nombre de coefficients AR	Etat 1	Etat 2	Etat 3	Etat 4	Etat 5
1	- 0.0006	0.0007	-0.0002	-0.0019	0.0009
2	- 0.0031	-0.0046	-0.0022	-0.0038	0.0010
3	0.0114	0.0160	-0.0036	-0.0075	0.0016
4	- 0.0155	-0.0375	-0.0085	0.0020	0.0021
5	0.0172	0.0666	-0.0092	0.0256	0.0013
6	-0.0209	-0.0922	-0.0183	0.0838	0.0014
7	0.0222	0.1125	-0.0204	0.1598	-0.0030
8	-0.0187	-0.1215	-0.0334	0.2560	0.0001
9	0.0189	0.1120	-0.0335	0.3331	-0.0083
10	-0.0156	-0.0972	-0.0426	0.3844	-0.0021
11	0.0025	0.0780	-0.0346	0.3765	-0.0054
12	-0.0146	-0.0540	-0.0327	0.3238	0.0045
13	0.0077	0.0315	-0.0176	0.2301	-0.0009
14	-0.0042	-0.0165	-0.0118	0.1375	0.0013
15	0.0049	0.0046	-0.0031	0.0630	-0.0014
16	-0.0112	0.0003	-0.0013	0.0217	0.0009
17	0.0015	-0.0008	0.0007	0.0042	0.0018
18	-1.8404	-3.4329	3.0621	5.4218	0.4180
19	3.3358	7.0508	6.7106	17.1825	0.9896
20	-4.9951	-12.2697	9.9895	37.0327	-0.8901
21	7.3639	18.4713	13.9334	63.0452	0.7474
22	-8.0193	-23.3417	16.5473	89.7576	-0.8621
23	9.6349	27.1675	19.2675	113.9059	1.3542
24	-10.1010	-29.4023	19.3841	129.6698	-1.0230

Tableau IV.9 : Matrice des moyennes pour la deuxième gaussienne (cas de *ARMA*).

IV.2.2. PHASE DE RECONNAISSANCE (TEST)

La reconnaissance est basée sur le principe illustré dans la *figure (IV.I)*

- un module de paramétrisation.
- un module de reconnaissance.
- un module de décision.

IV.2.2.1. LE MODULE DE PARAMETRISATION

Le principe de paramétrisation est identique à celui utilisé durant la phase d'apprentissage. Conserver la même paramétrisation lors des phases d'entrainement et de test

est primordial afin de fournir au système automatique des informations comparables et de même nature.

IV.2.2.2. LE MODULE DE RECONNAISSANCE

Le module de reconnaissance a un rôle essentiel dans le système de vérification du locuteur. Il compare les paramètres d'un individu aux paramètres de référence extraits du signal client lors de la phase d'apprentissage. Cette comparaison se fait en calculant le maximum de la probabilité de vraisemblance de chaque locuteur issu de l'apprentissage appelé score. Ce score est ensuite transmit au module de décision.

IV.2.2.3. LE MODULE DE DECISION

A partir de ce score, le module de décision fournit une décision qui constituera la réponse finale du système de vérification du locuteur.

Un système de vérification du locuteur doit être confronté à deux types de tests :

- Les tests clients lors desquels la parole présentée au système correspond à l'identité clamée.
- Les tests imposteurs lors desquels la parole présentée au système provient d'un individu inconnu du système.

Nous avons effectué 20 tests : 10 clients et 10 imposteurs :

Le système automatique doit répondre à chaque tentative de vérification auquel il fait face par une décision binaire. Il peut donc engendrer trois types d'erreurs :

- Faux rejet **(*FR*)** erreur commise lorsque le système rejette, à tort, un client légitime (erreur commise lors d'un test client) ;
- Fausse acceptation **(*FA*)** erreur commise lorsqu'un imposteur est accepté en tant qu'utilisateur légitime (erreur commise lors d'un test imposteur).
- Erreur moyenne ***(EM)*** erreur moyenne des deux erreurs.

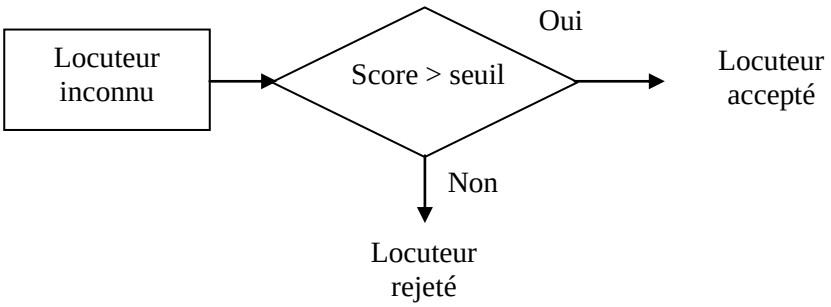


Figure IV.10 : Décision de vérification du locuteur.

Un score élevé indiquera que la probabilité pour que l'utilisateur testé corresponde à l'identité qu'il annonce est élevée, et un score faible indiquera que cette probabilité est faible. La décision qui constitue la sortie du module résulte de la comparaison de ce score avec un seuil défini à l'avance. Si le score est supérieur au seuil, l'utilisateur est accepté et s'il est inférieur au seuil, l'utilisateur est rejeté. *Figure IV.10*

Le choix d'un seuil a une incidence directe sur les performances du système. Pour un système idéal, les scores obtenus par les clients seront tous plus élevés que les scores obtenus par les imposteurs, assurant ainsi une vérification parfaite.

IV.2.2.4. RESULTATS DE RECONNAISSANCE

Le *Tableau VI.10* résume les performances de la vérification du locuteur effectuée en milieu non bruité.

Dans le cas où le signal parole est non bruité, La meilleure vérification est celle où on a utilisé les coefficients du modèle *ARMA* avec *EM=20%*, le modèle *LPCC avec EM=25%, CARMA* et *MFCC* avec *EM=30%* et enfin *AR* avec *EM=40%*.

Signal parole non bruité	***FA***	***FR***	***EM***
ARMA	**10 %**	**30 %**	**20 %**
LPCC	**30 %**	**20 %**	**25 %**
CARMA	**40 %**	**20 %**	**30 %**
MFCC	**30 %**	**30 %**	**30 %**
AR	**30 %**	**50 %**	**40 %**

Tableau IV.10 : Comparaison de performance entre *AR, LPCC, ARMA, CARMA, MFCC* dans le cas où le signal parole n'est pas bruité.

Le *Tableau VI.11* résume les performances de la vérification du locuteur effectuée en milieu bruité.

Signal parole bruité	*FA*	*FR*	*EM*
ARMA	**30 %**	**30 %**	**30 %**
CARMA	**30 %**	**40 %**	**35 %**
LPCC	**30 %**	**40 %**	**35 %**
AR	**30 %**	**50 %**	**40 %**
MFCC	**90 %**	**60 %**	**75 %**

Tableau IV.11 : Comparaison de performance entre *AR, LPCC, ARMA, CARMA, MFCC* dans le cas où le signal parole est bruité.

Dans ce cas aussi, La meilleure vérification est celle où on a utilisé les coefficients du modèle *ARMA* avec *EM=30%* , *CARMA* et *LPCC avec EM=35%, AR* avec *EM=40%* et enfin *MFCC* avec *EM=75%.*

Les performances des coefficients *MFCC* devront être meilleures en milieu bruité, et ce résultat est dû au non élimination du silence (bruit de repos) du signal parole avant le traitement de ce dernier.

IV.3. CONCLUSION

L'utilisation des coefficients *ARMA* dans La vérification du locuteur en mode dépendant du texte s'avère la meilleur méthode de modélisation de la parole et ceux dans un milieu bruité et non bruité.

Alors que l'utilisation des coefficients *AR* s'avère la méthode la plus médiocre dans la vérification du locuteur et a été remplacée par les *LPCC.*

CONCLUSION GENERALE ET PERSPECTIVES

Cette étude a été consacrée à la modélisation *AR* et *ARMA* du signal de la parole pour une vérification du locuteur en mode dépendant du texte. Le but de ce travail est de montrer que la modélisation *ARMA* est meilleure que la modélisation *AR* et autre.

Pour cela, nous avons réalisé une étude sur les différentes méthodes de modélisation *AR, LPCC, MFCC, ARMA* et *CARMA* de la parole, choisit les deux meilleure méthodes : une de corrélation (*AR)* et l'autre de Steiglitz McBride (*ARMA*), ensuite nous avons introduit les paramètres pertinents de chaque modèle dans la vérification du locuteur.

Pour la vérification automatique du locuteur, nous avons fait une étude des différents modèles existants, nous avons trouvé dans nos recherche que le modèle de Markov caché *HMM* était le plus performant dans le mode dépendant du texte.

Pour mener à bien le traitement prévu, nous avons dû stocker des voix de clients et d'imposteurs, puis procéder à l'analyse afin d'extraire les paramètres qui les caractérisent (pôles et zéros de la transmitance). Nous avons développé ensuite un programme qui permet la vérification automatique de ces locuteurs.

Après Comparaison des performances entre *AR, LPCC, MFCC, ARMA* et *CARMA*, nous avons trouvé que malgré que le temps de calcul était plus élevé dans le cas de *ARMA* par rapport au autres cas, il y a une amélioration de performance que ce soit dans le milieu bruité ou non bruité.

En ce sens nous espérons que ce travail trouvera une suite en implantant le modèle *ARMA* sur un système réel avec élimination du silence et optimisation des calculs.

Références Bibliographiques

[1] R. Boite. Traitement de la parole. Collection Electricité. Presses Polytechniques et Universitaires Romandes, 2000.

[2] M . Kunt and R. Boite. Traitement de la parole. Presses Polytechniques Romandes, press polytechnique romandes edition, 1987.

[3] John G. Webster. Wiley encyclopedia of electrical and electronics engineering, Volume 16. 1999.

[4] Thomas Hueber. Reconstitution de la parole par imagerie ultrasonore et vidéo de l'appareil vocal : vers une communication parlée silencieuse. Thèse de doctorat de l'université Pierre Marie Curie. 2009

[5] J. Makhoul. Linear prediction: A tutorial review. Proc. IEEE, 63(4):561–580, april 1975.

[6] S. Yim, D. Sen, and W.H. Holmes. Comparison of arma modelling methods for low bit rate speech coding. IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 1, page 273–276, 1994.

[7] J.L. Shanks. Recursion filters for digital processing. Geophysics, XXXII: 33–35, 1967

[8] G. E. Kopec, A. V. Oppenheim, and J. M. Tribolet, Speech analysis by homomorphic prediction, IEEE Trans. Acoust., Speech, Signal Processing, volume ASSP-25, page 40-49, Feb. 1977.

[9] K. Steiglitz and L.E. McBride. A technique for the identification of linear systems. IEEE Trans. Auto. Control, AC-10:461–465, October 1965.

[10] S. Park and J. Cordaro. Improved estimation of sem parameters from multiple observations. IEEE Trans.Electromagn. Compat., EMC-30:145–153, May 1988.

[11] K. Steiglitz, On the simultanous estimation of poles and zeros in speech analysis, IEEE Trans. Acoust. Speech, Signal Processing. Volume 25, page 194–202 June 1977.

[12] Alexandre. PRETI. Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur. , thèse de doctorat de l'école Doctorale 166 I2S, Décembre 2008.

[13] Yassine. Mami. Reconnaissance De Locuteur Par Localistion Dans un Espace de Locuteurs de Référence. Thése de Doctorat de l'École National de Télécommunications, Paris, Octobre 2003.

[14] T. Artières and P. Gallinari. Approches prédictives neuronales pour l'identification. XXème Journées d'Etudes sur la parole (JEP),Trégastel, France, pages 275–280, 1994.

[15] M. Homayounpour et G. Chollet, Neural net approaches to speaker verification: Comparison with second order statistic mesures. Dans Internatioonal Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages 353-356, 1995.

[16] Furui, S. Cepstral analysis technique for automatic speaker verification. Dans IEEE Transactions acoustics, Speech, and Signal Processing, volume 29, pages 254-272, 1981.

[17] Booth, I., Barlow, M., et Watson, B. Enhancements to DTW and VQ decision algorithms for speaker recognition. Speech Communication, 13(3-4): 427-433, 1993.

[18] L. Rabiner, B. H. J. Fundamentals of speech recognition. Prentice Hall Signal Processing Series, 1993.

[19] Reynolds, D. Speaker identification and verification using Gaussian mixture speaker models. Speech Communication, 17(1): 91-108, 1995.

[20] Reynolds, D. A., Quatieri, T. F., et Dunn, R. B. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10(1-3) :19-41, 2000.

[21] Juang B.-H. Rabiner L, Fundamentals of speech recognition, Prentice Hall, 1997.

[22] Rissanen, E. L. et Webb, J. J. Speaker identification experiments using HMMs. Dans Internatioonal Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 2, pages 387-390, 1993.

[23] Rosenberg, E., Lee, C.-H., et Gokcen, S. Connected word talker verification using whole word hidden Markov models. Dans International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages 381-384, 1991.

[24] Savic, M. et Gupta, S. K. Variable parameter speaker verification system based on hidden Markov modeling. Dans International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 1, pages 281-284, 1990.

[25] J.M. Naik, Speaker Verification: A Tutorial, IEEE Communication Magazine, volume 28, page 42-48, January 1990

[26] Ying Liu, Martin Russell, and Michael Carey. The role of dynamic features in text-dependent and -independent speaker verification. In Proceedings of the IEEE (ICASSP). volume 1: 669– 672, 2006.

[27] L. P. Wong and M. Russell. Text-dependent speaker verification under noisy conditions using parallel model combination. Acoustics, Speech and Signal Processing Proceedings of the IEEE ICASSP International Conference on, volume 1, page 457–460, 2001.

[28] Johnny Mari Ethoz, Johnny Mari, and Samy Bengio. A comparative study of adaptation methods for speaker verification, Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP). 2002.

[29] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In Proceedings of the IEEE, 77(2): 257–286, 1989.

[30] Changshou Deng and Pie Zheng. A New Hidden Markov Model with Application to Classification. In Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on, volume 2, pages 5882–5886, 2006.

[31] Mohd Zaizu Ilyas,, Salina Abdul Samad, Aini Hussain, and Khairul Anuar Ishak, Members, IEEE, Speaker Verification using Vector quantization and Hidden Markov Model, In Research and Development, 2007. SCOReD 2007. 5th Student Conference on, pages 1–5, 2007.

[32] Asmaa Amehraye, Débruitage perceptuel de la parole, thèse de doctorat à l'Ecole Nationale Supérieure des Télécommunications de Bretagne, 2009.

[33] B.S ATAL, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, J. Acoust. Soc. Am., Volume 55, page : 1304-1312, June, 1974.

[34] S. Furui. Spaeaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. Acoustics, Speech and Signal Processing, IEEE Transactions on, 34(1):52–59, 1986.

[35] Ben Milner and Xu Shao. Speech reconstruction from Mel frequency cepstral coefficient using a source filter model. In John H. L. Hansen and Bryan L. Pellom, editors, *INTERSPEECH*. ISCA, 2002.

[36] Kenneth Thomas Schutte, Parts-based Models and Local Features for Automatic Speech Recognition, thèse de doctorat à MASSACHUSETTS INSTITUTE OF TECHNOLOGY, Department of Electrical Engineering and Computer Science, 2009.

Résumé

Extraire d'un signal parole les paramètres pertinents les plus performants pour une vérification automatique du locuteur définit la motivation principale de cette thèse. Ces paramètres représentent les caractéristiques du conduit vocal caractérisant la voix d'une personne.

Pour cela, nous nous intéressant à la modélisation pôles zéros *ARMA* qui est une méthode performante et peu utilisée vu sa non linéarité donc sa complexité.

Dans un premier temps, différentes méthodes d'analyses sont appliquées, *AR, LPCC, ARMA, CARMA et MFCC* dans un milieu bruité et non bruité. Ce travail est poursuivi par la modélisation du locuteur en utilisant l'approche statistique *HMM*. Pour chaque locuteur chaque modèle est entrainé sur la phrase prononcé (mot de passe), il est basé sur le calcul d'un score de vraissemblance.

En l'occurrence du test, nous avons obtenue un taux de reconnaissance de 60% pour le modèle *AR* et *80%* pour le modèle *ARMA* dans un milieu non bruité et un taux de 60% pour le modèle *AR et 70%* pour le modèle *ARMA* dans un milieu bruité.

Abstract

Extracting from a speech the pertinent parameters the most efficient for automatic speaker verification that defines the main motivation of this thesis. These parameters represent the vocal tracts' characteristics that characterize a person's voice.

For this, we are interested in modeling poles zeros *ARMA* which is an efficient method and less used due to its nonlinearity therefore its complexity.

At the beginning, different methods of analysis are applied *AR, LPCC, ARMA, CARMA* and *MFCC* in a noisy environment and a non-noisy one. This work was pursued by modeling the speaker using the statistical approach *HMM*. For each speaker, every model is trained on the pronounced sentence (password), it is based on the computation of a score Likelihood.

In the test occurrence, we have obtained a recognition rate of *60%* for the *AR* model and *80%* for the *ARMA* model in a non-noisy environment and a rate of *60%* for the *AR* model and *70%* for the *ARMA* model in a noisy environment.

تلخيص

إن الدافع الرئيسي لهذه الأطروحة هو استخراج البارامترات الأكثر كفاءة من الحديث من أجل التحقق التلقائي من المتحدث. تمثل هذه البارامترات خصائص الجهاز الصوتي لشخص معين.

و لذلك نحن مهتمون بنموذج الأقطاب أصفار *AMRA* و الذي هو وسيلة فعالة لكنه لا يستخدم إلا قليلا، ذلك لكونه ليس خطيا و بالتالي معقد.

أولا يتم تطبيق أساليب مختلفة للتحليل *AR، LPCC، ARMA، CARMA، MFCC* في بيئة صاخبة و أخرى غير صاخبة و من ثم نمذجة المتحدث باستعمال النموذج الإحصائي. حيث لكل متحدث نموذجا يقوم على كلمة السر، و يعتمد هذا النموذج على حساب احتمال العتبة.

على ضوء التجربة حصلنا على معدل المعرفة بنسبة 60 % في *RA* و 80 % في *AMRA* في بيئة غير صاخبة وبنسبة 60 % في *RA* و 70% في *AMRA* في بيئة صاخبة.