



La reconnaissance automatique du locuteur par la voix IP



Lieu de stage : **Laboratoire des Systèmes Intelligents & Applications**

Réalisé par : Houda KADI

Soutenu le : 18/06/2014

Encadré par : Devant le jury composé de :

Pr. Jamal KHARROUBI Pr. J.KHARROUBI

Pr. Fatiha MRABTI Pr. S.NAJAH

Pr. R.BENABBOU

Pr. L.LAMRINI

## 1.1 Introduction

La reconnaissance automatique du locuteur est une des branches de l’authentification biométrique, qui se réfère à la reconnaissance automatique de l’identité des personnes en utilisant certaines de leurs caractéristiques intrinsèques. Outre la voix, il y a beaucoup d’autres modèles physiques et comportementaux pour l’authentification biométrique, par exemple : l’iris, les réseaux veineux de la rétine, les réseaux veineux de la paume de la main, l’empreinte digitale, …etc. Pratiquement, la sélection d’un modèle biométrique adéquat devrait prendre en compte au moins les considérations suivantes : la robustesse, la précision, l’accessibilité et l’acceptabilité. Par rapport à ces critères de sélection, parmi toutes les technologies d’authentification biométriques, la reconnaissance du locuteur est probablement la plus naturelle et économique pour les systèmes de communication homme-machine parce que d’une part la collecte de données parole est beaucoup plus pratique que les autres motifs, et d’autre part, la parole est le mode dominant d’échange d’information pour les êtres humains et tend à être le mode dominant pour l’échange d’information pour les systèmes de communication homme-machine.

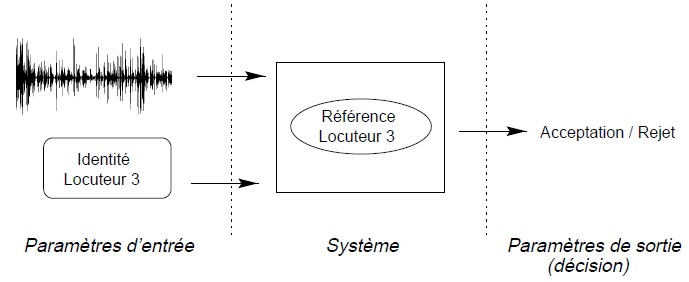
## 1.2 Terminologie

### 1.2.1 Identification & vérification automatique de locuteur

La reconnaissance automatique de locuteur consiste à obtenir des renseignements concernant l’identité d’une personne à partir d’un enregistrement de sa voix. Pour qualifier précisément les différentes tâches entrant dans le cadre d’un système de reconnaissance automatique de locuteur, on distingue entre deux tâches différentes

####  La Vérification Automatique de Locuteur (VAL)

Lorsqu’on cherche à décider si l’identité revendiquée par un locuteur est compatible avec sa voix. Dans ce type d’applications, il s’agit donc de trancher entre 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 (*fig 1.1*).



*Figure 1.1 : La vérification automatique de locuteur*

####  L’Identification Automatique de Locuteur ( IAL)

Il s’agit de déterminer, parmi un ensemble de N locuteurs potentiels, à quel locuteur correspond un enregistrement vocal. En identification, la réponse apportée n’est pas de type binaire (acceptation ou rejet) comme pour la vérification, puisqu’il est nécessaire de distinguer un locuteur parmi un groupe. On distingue encore deux sous problèmes d’identification selon que l’on est sûr ou non du fait que l’enregistrement provient bien d’un des membres du groupe de locuteurs :

* Si l’on a affaire à **un ensemble fermé** (*closed set*) (*fig.1.2*) : le système IAL décide de l’identité la plus probable parmi les utilisateurs connus (dont il possède une référence). Ce mode de fonctionnement tend à considérer que seules des personnes référencées peuvent accéder au système. Un tel système ne doit alors être utilisé que dans un environnement au sein duquel tous les individus sont connus.

L’identité IY retournée, correspondant à la référence Y est obtenue par :

**Y = argmax f(X|S).** f(X|S) est le score calculé lors de la comparaison des données S au modèle de

référence de l’individu IX.

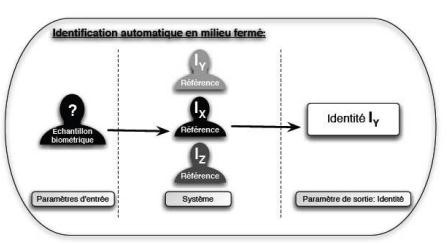
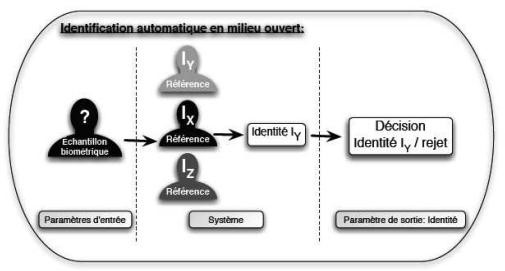


Figure 1.2 : L'identification automatique de locuteur *en un groupe fermé*

* Dans le cas d’**un ensemble ouvert** (*Open set*) (*fig.1.3*) : le système IAL a la possibilité de rejeter le locuteur dont il teste les données si elles ne correspondent à aucune des identités répertoriées. Ce locuteur est alors considéré comme inconnu du système. Pour ce faire, les données S sont comparées à chaque référence X connue par le système. Chaque comparaison fournit un score f(X|S). Le score le plus élevé est alors comparé à un seuil fixé préalablement.

Si le score est supérieur à ce seuil, le système décide qu’il s’agit de la personne correspondant à la référence sélectionnée. Si le score est inférieur à ce seuil, le système décide qu’il ne s’agit pas d’une personne « connue ».



*Figure1.3 : L'identification automatique de locuteur en un groupe ouvert*

Pour résumer, en identification en milieu ouvert, le système répond à deux interrogations : « *Quelle est l’identité la plus probable ?*» et « *Les données biométriques analysées correspondent-elles à cette identité ?*», alors qu’en milieu fermé il ne répond qu’à la première. L’identité la plus probable est obtenue, comme dans le cas de l’identification en milieu fermé, par **:** Y = argmax f(X|S). Après, cette probabilité maximale est comparée à un seuil Ω fixé au préalable, si la probabilité obtenue est supérieure à ce seuil, alors le locuteur est bien identifié, sinon, le système le déclare comme une identité non reconnue.

### 1.2.2 Dépendance et indépendance du texte

On parle de reconnaissance de locuteur en mode dépendant du texte (Text-dependant ou fixed-text) lorsque le texte prononcé par le locuteur est fixé et connu à l’avance. Par contre, si le texte prononcé n’est pas connu à priori, on parle du mode indépendant du texte (textindependant ou free-text). La distinction entre ces deux modes de fonctionnement des applications de reconnaissance du locuteur est très importante car les techniques utilisées ainsi que les performances obtenues, dans les deux cas sont très différentes.

### 1.2.3 Les variabilités du signal de la parole

Le signal de parole varie selon le locuteur, on parle de la variabilité **interlocuteur** lorsque les caractéristiques qui sont propres à chaque locuteur ne sont pas les mêmes chez d’autre locuteur.

La variabilité **intra-locuteur** concerne les changements de la voix du même locuteur et qui sont dus, en général, à la fatigue, le stress, le sommeil, l’horaire de la journée (matin ou soir), le débit de l’élocution, l’état émotionnel, ...

Pour la reconnaissance de locuteur, on cherche à extraire des caractéristiques du signal de parole qui présente une forte variabilité interlocuteur (pour pouvoir différencier les locuteurs entre eux) et une faible variabilité intra-locuteur (pour garantir la robustesse du système).

La variabilité **intersession** (entre sessions d’enregistrements) fait apparaître l’influence de facteurs extérieurs sur le signal de parole. A la sortie du conduit vocal humain, l’onde de parole est considérée comme idéale, car aucune déformation/distorsion de l’environnement extérieur ne l’a modifiée. L’environnement sonore lors de l’enregistrement, le matériel d’acquisition ou le canal de transmission utilisé vont ensuite déformer l’onde sonore originelle. Le canal de transmission, par exemple, agit comme un filtre en fréquence sur l’onde sonore. Ces facteurs rendent complexe la comparaison entre plusieurs échantillons d’un même individu. De nombreux travaux expérimentaux ont montré que des variations de matériel entre les phases d’apprentissage et de test sont à l’origine de graves dégradations des performances. Par exemple, l’acquisition d’un signal de parole sur le réseau GSM [10] introduit les dégradations suivantes sur le signal de parole :

* L’ajout du bruit de l’environnement,
* Le sous-échantillonnage à 8khz du signal,
* Le filtrage sur la bande de fréquence [300 – 3400] hz,
* Le codage à bas débit de la parole,
* L’ajout du bruit de quantification des paramètres émis,
* La transmission sur un lien sans-fil avec pertes.

## 1.3 Fonctionnement d’un système de reconnaissance automatique du locuteur

La réalisation d’un système de reconnaissance automatique de locuteur passe par trois phases : la paramétrisation, la modélisation, et la décision.

### 1.3.1 Paramétrisation

La variation de la nature du signal acoustique rend le traitement des données brutes issues de ce dernier très difficile. En effet, ces données contiennent des informations complexes, souvent redondantes et mélangées.

La phase de paramétrisation, qui traite le signal acoustique reçu, doit remplir plusieurs objectifs :

* Séparer le signal du bruit ;
* Extraire l’information utile à la reconnaissance ;
* Convertir les données brutes à un format directement exploitable par le système.

Afin de concevoir un bon système RAL, il faut choisir des paramètres qui sont fréquents, (ne pas correspondre à des évènements ne survenant que très rarement dans le signal), facilement mesurables, robuste face aux imitateurs, ne pas être affecté par le bruit ambiant ou par les variations due au canal de transmission.

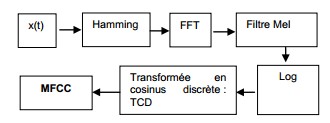
Pratiquement, il est très difficile de réunir tous ces éléments en même temps, la sélection des paramètres pose un problème très complexe, et influe fortement sur les résultats des systèmes RAL. D’après plusieurs recherches effectuées sur cette étape, les types de paramètres efficaces et utilisables sont les paramètres de l’analyse spectrale, les paramètres prosodiques, et les paramètres dynamiques.

####  Paramètres de l’analyse spectrale

Les principaux paramètres de l’analyse spectrale utilisés en RAL sont les coefficients de prédiction linéaire et leurs différentes transformations (LPC, LPCC,..) ; ainsi que les coefficients issus de l’analyse en banc de filtres et leurs différentes transformations (coefficients banc de filtres, MFCC...).

Plusieurs travaux ont été publiés pour comparer les différentes techniques en paramétrisation, l’enjeu de ces travaux était de cibler les meilleurs paramètres représentant de façon efficace les propriétés caractéristiques propres à chaque locuteur. Les meilleurs résultats ont été obtenus en utilisant la méthode **MFCC**.

Le codage MFCC (Mel-Frequency Cepstral Coefficients) est basé sur les variations des bandes critiques de l’oreille humaine avec la fréquence, les filtres sont espacés linéairement aux basses fréquences et logarithmiquement à hautes fréquences, ces filtres sont modélisés par une échelle non linéaire issue de connaissances sur la perception humaine : l’échelle de Mel. Pour les MFCC on utilise la fenêtre de Hamming durant la transformation du domaine temporel au domaine fréquentiel. Cette transformation est faite en utilisant la transformée de Fourier. Un filtrage est appliqué ensuite, par banc de filtres triangulaires espacés selon l’échelle de Mel. Cette échelle reproduit la sélectivité de l’oreille qui diminue avec l’accroissement de la fréquence. Après le calcul de log, une transformée en cosinus discrète est appliquée pour assurer un retour au domaine temporel (*figure 1.4*)



*Figure 1.4: Extraction des paramètres avec MFCC*

####  Paramètres prosodiques

Le terme "paramètres prosodiques" réunit l’énergie, la durée et la fréquence fondamentale (ou pitch). Ces paramètres caractérisent en grande partie le style d’élocution d’un locuteur. L'énergie contient l'information liée au niveau acoustique moyen du signal. Ces paramètres s’avèrent fragiles en pratique et ne permettent pas, à eux seuls, de discriminer de manière fiable les locuteurs. En conséquent, ils sont souvent associés aux paramètres de l’analyse spectrale.

####  Paramètres dynamiques

La prise en compte d’une information de type dynamique peut être un facteur d’amélioration des performances d’identification du locuteur.

Une première approche, employée pour utiliser cette information au niveau des paramètres, consiste à utiliser une concaténation de plusieurs trames successives de parole (méthodes prédictives). Cependant, cette approche nécessite plus de paramètres dans les modèles et est conduit à des problèmes d’estimation des modèles lors de l’apprentissage.

La seconde possibilité consiste à calculer les dérivées du premier et du second ordre appelé aussi coefficient de (∆) ou (∆∆) qui sont désormais très répandue en raison de leur simplicité de mise en œuvre.

### 1.3.2 Traitement post paramétrisation

Les paramètres MFCC sont sensibles au changement du canal ou d’environnement, pour cela il est nécessaire d’appliquer autres traitement pour remédier à ce changement, et pour réduire le bruit, dans cette partie on va présenter une étude sur les méthodes les plus connus pour ce traitement,

####  La méthode RASTA

La méthode Rasta est une méthode qui a été proposé en 1994, elle a pour but de supprimer les composantes spectrales dont l’évolution temporelle est plus rapide ou lente que celle du conduit vocal humain.

La méthode RASTA est intégrée à une analyse PLP. En effet, après avoir effectué la transformée de Fourier discrète à court terme, on calcule le spectre d’amplitude en bandes critiques. On applique le logarithme pour récupérer l’enveloppe spectrale du signal comme pour une analyse cepstrale. On effectue ensuite un filtre passe bande qui a pour conséquence de supprimer les composantes constantes ou lentes du signal. On réalise après une compression de l’amplitude par l’application d’une racine cubique. Enfin, on calcule les coefficients selon la méthode LPC classique (*figure 1.5*).

Signal sonore



Analyse Spectrale



Banc de compression statique non linéaire



Banc de filtre passe bande linéaire



Banc de dilatation statique non linéaire



Processus optionnel



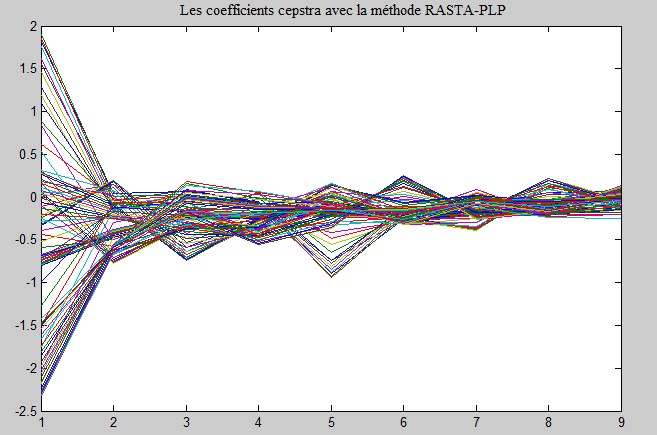
Coefficients RASTA PLP



RASTA

*Figure 1.5: Chaine de calcul de coefficient RASTA-PLP*

Un exemple de calcul de quelques paramètres d’un signal de parole utilisant cette technique est illustré par la figure 1.6 :

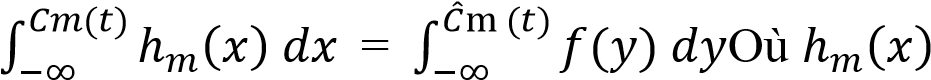


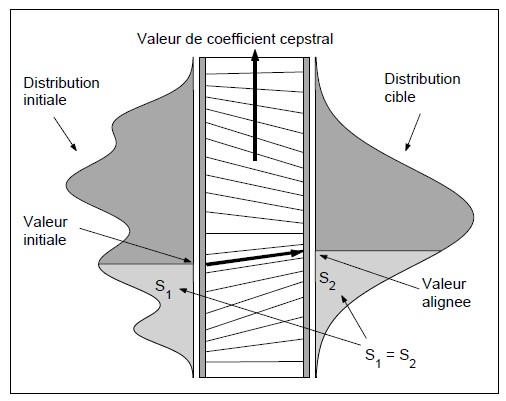
*Figure 1.6: Les coefficients Cepstra avec la méthode RASTA-PLP*

Plusieurs études réalisées pour mesurer la performance de cette technique, ont permis de confirmer ses bonnes qualités relativement aux distorsions et ses moindres qualités face aux bruits additifs, signe de la présence de plusieurs sources sonores dans un même environnement.

####  Feature warping

Le feature warping est une technique, apparue en 2001 [20], qui a pour but de conditionner la distribution de chaque trajectoire cepstrale en alignant les coefficients cepstraux Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

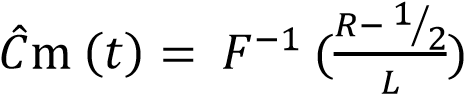
d’une telle façon que sa distribution soit égale à une distribution donnée. Pour un coefficient cepstral Cm(t), on cherche donc 𝐶̂m  satisfaisant : est la distribution de la 𝑚è𝑚𝑒trajectoire cepstrale et f(y) est la distribution cible. Généralement on prend comme distribution cible la distribution de la loi normale. Dans ce cas, cette méthode s’appelle aussi gaussianisation. Un schéma explicatif est représenté dans la figure 1.7:



*Figure 1.7: Alignement des vecteurs acoustiques*

La mise en œuvre de cette méthode passe par ces quatre étapes :

* Pour chaque coefficient Cm(t) fenêtrer sa trajectoire par une fenêtre de taille L de telle façon que ce coefficient soit au milieu de la fenêtre. Il est conseillé de prendre L= 3 sec.
* Ranger les coefficients obtenus dans l’ordre croissant de ces valeurs.
* Calculer le rang de chaque coefficient comme sa position dans ce rangement (Le plus petit coefficient obtient le rang 1 bien que le plus grand L). Cela revient à calculer la fonction de répartition de la distribution locale de la trajectoire cepstrale.  Soit R, le rang de Cm(t), calculer la nouvelle valeur 𝐶̂m (𝑡) comme :



Où F est la fonction de répartition de la distribution cible

La technique feature warping peut être combinée avec d’autres techniques pour améliorer la robustesse d’un système de reconnaissance sous certaines conditions.

####  CMVN (Normalisation moyenne et variance des paramètres cepstraux)

La normalisation moyenne et variance des paramètres cepstraux est une technique très simple et très répondue en reconnaissance automatique de locuteur. Elle consiste à retirer Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

la moyenne de la distribution de chacun des paramètres cepstraux (la composante continue), et à ramener la variance à une variance unitaire en les divisant par l’écart type global des paramètres acoustiques [26]. Quand seule la moyenne est normalisée, on parle alors de la

Cepstral Mean Subtraction (CMS) ou *Cepstral Mean Normalization* (CMN)

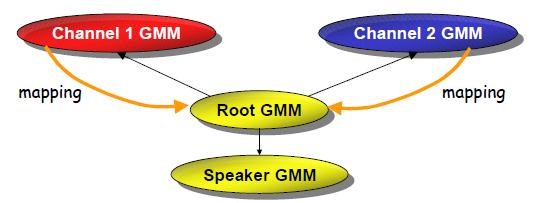
####  Le feature mapping

La CMVN, le filtrage RASTA, et le feature warping sont toutes des techniques non supervisées qui n’utilisent aucune connaissance sur le canal, par contre le feature mapping, proposé par Reynolds en 2003 [22], est une technique supervisé de normalisation qui projette les paramètres acoustiques liées à une condition d’enregistrement donnée dans un nouvel espace de caractéristiques indépendant du canal. Cette transformation permet de réduire les effets de la variabilité du canal.

La technique de feature mapping est inspirée de l’approche Speaker Model Synthesis [24], elle se base sur la projection des caractéristiques des différents canaux, en un seul espace de caractéristiques indépendant du canal.

La figure 1.8 montre la structure de la technique de feature mapping, comme en SMS défini dans [24], le FM apprend un modèle GMM pour chaque canal connue par adaptation MAP d’un modèle indépendant du canal. Le modèle dépendant du canal modélise en réalité un sous espace de l’espace acoustique global. Il en est déduit une transformation représentant la relation entre le modèle indépendant du canal et le modèle dépendant du canal.

Lors de la phase de l’apprentissage des modèles, on cherche en premier lieu à identifier le canal le plus vraisemblable pour l’enregistrement traité, et on applique par la suite la transformation qui lui est associé sur les vecteurs paramétriques, cette transformation permet de réduire les effets de la variabilité du canal, entre conditions de test et d’apprentissage.



*Figure 1.8: Feature mapping*

Cette nouvelle approche a montré de bonnes performances au niveau de la réduction des effets du canal, elle peut être appliquée pour d’autres applications de reconnaissance de parole.

####  VTLN (Vocal Tract Length Normalization)

La variabilité interlocuteur est l’un des problèmes majeurs dans un système de reconnaissance de locuteur, elle influence les performances globales du système à cause de la variabilité des locuteurs, de leurs accents, leurs styles de vocabulaires.

La normalisation de la longueur de conduit vocal (VTLN) est une technique de normalisation de locuteur intervenant au niveau des paramètres acoustiques. Elle est largement répandue dans les systèmes de reconnaissance à grands vocabulaires [28]. Cette normalisation repose sur une modification linéaire de l’échelle des fréquences afin de compenser les différences de longueur de conduit vocal entre les locuteurs.

En général, deux questions sont posées dans VTLN [23] : Premièrement, à partir d’un signal de parole donné, comment obtenir le facteur de normalisation ? La deuxième question est comment fait la normalisation en sachant le facteur de normalisation ?

Le facteur de normalisation (warping factor) reflète la différence du conduit vocal entre plusieurs locuteurs, il existe plusieurs techniques pour calculer l’estimer, en générale il est sélectionné parmi un ensemble de valeurs de 0,8 à 1,25.

Pour la deuxième question, l’utilisation de l’échelle de bark/mel [23]

La majorité des articles publiés sur la normalisation de conduit vocale s’intéresse sur l’une des sujets suivants :

* Type de l’échelle de fréquence (linéaire ou non linéaire), et son implémentation sur quel domaine (Domaine temporel, domaine fréquentiel, domaine cepstral).
* Estimation de facteur de normalisation (warpingfactors) pour l’apprentissage.
* Estimation d’un facteur de normalisation efficace pour le test
* Gain en taux de reconnaissance en utilisant la méthode VTLN et dans plusieurs situations (Environnement bruité ou propre, large corpus d’apprentissage ou limité, un grand vocabulaire ou petit).
* La comparaison de la technique VTLN avec des techniques d’adaptation

####  HLDA (HETEROSCEDASTIC LINEAR DISCRIMINANT ANALYSIS)

De façon générale, l’objectif de la transformation discriminante des observations consiste à trouver un espace de projection de faible dimension mais conservant l’information discriminante, pour cela, l’approche de LDA et sa généralisation HLDA sont les plus utilisés.

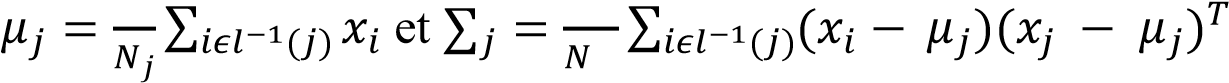
De point de vue mathématique, on peut exprimer cela en appliquant la transformation linéaire suivante : 

Avec y représente l’ensemble des vecteurs du nouveau espace, x représente l’espace de vecteurs d’entrée, et  est la matrice de transformation, de dimension *n x p*.

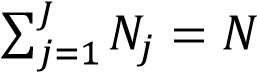
Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

#####  LDA (Analyse discriminante linéaire)

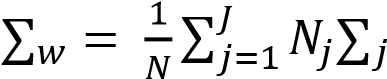
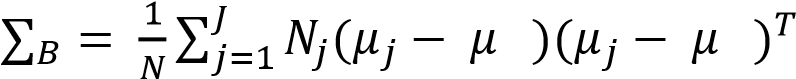
C’est une méthode utilisé pour les statistiques, la reconnaissance de formes, et l’apprentissage automatique pour trouver une combinaison linéaire qui caractérise ou sépare deux ou plusieurs classes d’objets ou d’événement, la combinaison résultante peut être utilisée comme un classificateur linéaire, ou plus couramment la réduction de la dimensionnalité plus tard avant la classification [17].La LDA consiste tout d’abord à calculer les vecteurs moyens et les matrices de covariances pour chaque classe j :

1 1

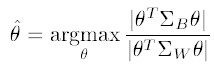
Où :

- Nj le nombre de vecteurs dans la classe j, 

Ensuite on calcule les deux matrices de covariance : inter-classe ∑ 𝑤et intra-classe ∑ 𝐵

 et

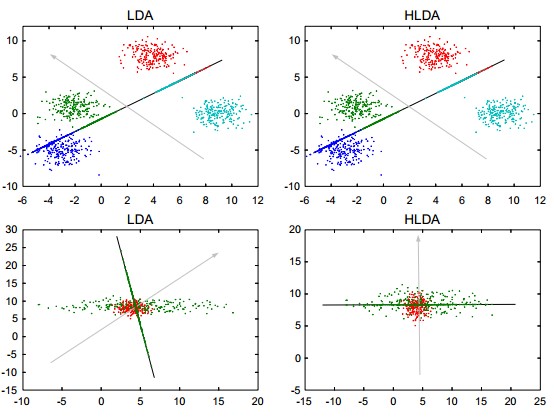
Enfin on cherche la transformation 𝜃̂ satisfaisant le critère suivant :



Le but de ce critère est de maximiser la variabilité entre les classes tout en minimisant la variabilité entre les classes.

Le problème de LDA, c’est qu’elle est essaie de séparer la moyennes des classes sans prendre en compte l’information discriminante présente dans la différence de matrice de covariance, elle est incapable de traiter les données dans le cas hétéroscédastique, càd le cas dans lequel les classes n’ont pas de matrices de covariance égale, cette limitation devient très évidente dans le cas de deux classes avec deux matrice de covariance différente (figure 1.9). D’où l’apparition de la méthode HLDA [17].

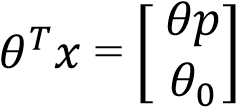
Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur



*Figure 1.9: Limitation de LDA*

#####  HLDA (HeteroscedasticLinear Discriminant Analysis)

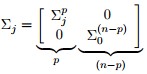
L’idée de HLDA repose sur l’hypothèse qu’après la transformation linéaire de la matrice 𝜃, seulement les p-colonnes portent l’information discriminante et couvrent le sous espace de dimension p dans lequel les moyens et les variances différent, alors que les autre seront homogènes.

On écrit donc : 𝜃 = [ 𝜃𝑝 𝜃0] avec y =  x

L’ensemble de données d’entrée est supposé une distribution gaussienne, après la transformation linéaire, l’ensemble de données de sortie est aussi une distribution gaussienne, la moyenne est donc :



De même, l’expression de covariance pour une classe j est :



En se basant sur ces deux définitions, l’expression de la densité de probabilité gaussienne devient :



Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

Avec g(i) = j représente la projection des vecteurs d’observation vers des classes, pour obtenir le meilleur estimateur pour 𝜃, il faut calculer la dérivée de la fonction de max de vraisemblance

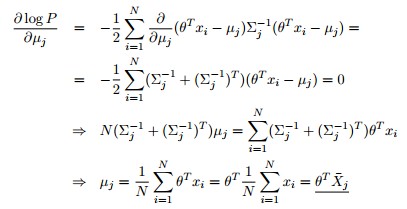
par rapport à 𝜃 et la mettre égale à 0



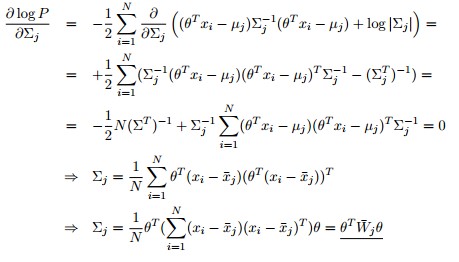
On ne peut pas calculer directement la dérivée de cette expression, parce que les paramètres de covariance et de la moyenne dépende encore de 𝜃,

D’abord, on calcul d’abord :

* la dérivée par rapport à 𝜇𝑗

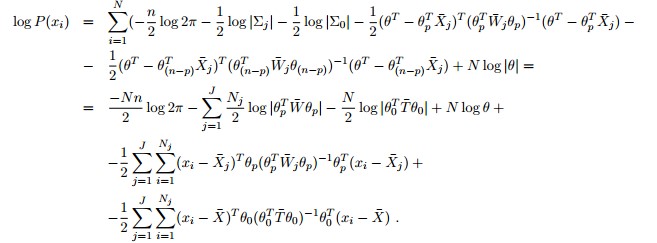


* la dérivéepar rapport à ∑𝑗

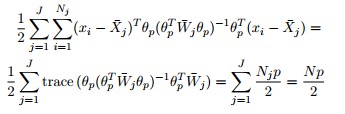


Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

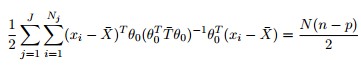
En remplaçant ces valeurs par leurs expressions, on trouve :



Cette expression peut être simplifiée en remplaçant



Et



Et finalement on obtient :



Les méthodes de la transformation linéaire comme le LDA ou HLDA ou leurs variantes (SHLDA proposé dans [6], MAP-SHLDA) sont largement utilisés pour réduire la dimension dans toute sorte de problème de classification. Les avantages de ces méthodes sont : elles peuvent être calculées sur des ensembles de données énormes, ainsi qu’elles ont un nombre de paramètres relativement faible, pourtant elles donnent des estimations robustes pour le modèle sous-jacent.

####  Speech enhancement

Le « speech enhancement » est un ensemble de techniques de traitement du signal de parole qui filtre le bruit et met la parole en valeur. L’amélioration de la qualité de signal (Speech Enhancement) est un problème très difficile, les sources de complexité sont de natures différentes, d’une part la nature et les caractéristiques du signal bruité varie d’une application à une autre, et donc il n’est pas évident de trouver un modèle ou un algorithme qui fonctionne pour toutes les applications, et quel que soit l’environnement du travail, d’autre part il y a deux critères pour mesurer les performances d’une application : la qualité d’un signal de parole Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

(mesure subjective) et l’intelligibilité de la parole[[1]](#footnote-1) (mesure objective), et donc il est plus difficile de satisfaire à ces deux critères en même temps.

Dans ce qui suit, on va présenter quelques méthodes qui ont pour objectif de restaurer un signal utile à partir des observations corrompues par un bruit supposé souvent additif, cette hypothèse est souvent utilisée, à la fois pour sa simplicité, mais aussi elle permet de modéliser un grand nombre de situations pratiques. Le signal observé est donc considéré comme la somme du signal de parole et du bruit ambiant. Ce modèle omet tout bruit convolutif, électrique ou de quantification.

#####  La soustraction spectrale

La soustraction spectrale est la méthode de débruitage la plus ancienne [7]. Elle opère dans le domaine fréquentiel et a pour principe de soustraire une estimé du bruit à partir du signal observé. Le bruit est supposé additif, stationnaire ou légèrement variant ce qui nous permet de l’estimer pendant les périodes de silence. Il existe deux versions de base de la soustraction spectrale se différenciant l’une de l’autre par l’utilisation soit de la puissance soit de l’amplitude.

* Si |𝑆̂(𝑣)| = |𝑌(𝑣)| − |𝐵̂(𝑣)| Il s’agit de la soustraction spectrale d’amplitude
* Si par contre, le bruit estimé est donné par son spectre de puissance, on aura la soustraction spectrale de puissance :



Vu que le second terme de l’équation précédente peut être négatif, on peut le rendre positif en changeant de signe ou bien en l’annulant comme l’équation suivante. Ceci fait partie des premières améliorations apportées à la soustraction spectrale

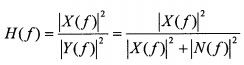


Le passage dans le domaine temporel est réalisé par la transformée de Fourrier inverse en gardant la phase du signal bruité. On se permet de procéder ainsi, d’une part, parce que notre oreille est peu sensible aux variations de la phase et, d’autre part, parce qu’une estimation de la phase est une tâche très compliquée.

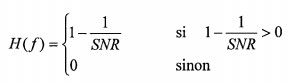
avec |𝑆̂(𝑣)| le spectre d’amplitude du signal rehaussé et arg Y(v) la phase du signal bruité Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

#####  Le filtre de Wiener

Une autre approche de débruitage consiste à appliquer au signal bruité un filtre optimal appelé filtre de Wiener [7]. Cette approche utilise également le spectre d’énergie et par conséquent, un estimé de signal dans le domaine spectral est requis. Comme dans le cas de la soustraction spectrale, cet estimé peut être obtenu par l’équation précédente. Ce type de filtrage consiste à obtenir une estimation du signal original suivant le critère minimum de l’erreur quadratique moyenne. Ce critère conduit au filtre de Wiener dont la réponse fréquentielle est donnée par :



Pour cette méthode, il est essentiel de connaître l’estimation du bruit pour évaluer|𝑁(𝑓)|2, concernant|𝑋(𝑓)|2, cela est obtenu par soustraction de l’estimation de |𝑁(𝑓)|2du spectre d’énergie du signal bruité et du conduit au résultat :



Avec SNR est le rapport signal-bruit en entrée.

Ce filtre convient non seulement au son, mais aux images, car le spectre de fréquence de la plupart des images visuelles est souvent bien conditionné et peut être estimé facilement.

#####  Filtrage de Kalman

Le filtre de Kalman est une méthode visant à estimer des paramètres d’un système évoluant dans le temps à partir de mesures bruités, on retrouve ce filtre dans un bon nombre de domaines relatifs au traitement du signal, radar, traitement d’images,…

Le fonctionnement du filtre de Kalman est expliqué en détail en [3] peut se diviser en deux étapes :

* Une première étape de prédiction de l’estimation selon le modèle du système. Pour ce faire, le filtre de kalman reprend l’estimation précédente des paramètres et de l’erreur et prédit les nouveaux paramètres et la nouvelle erreur en fonction de la modélisation du système.
* La seconde étape va faire la mise à jour de cette prédiction grâce aux nouvelles mesures. Ces mesures (par définition bruités) vont permettre d’obtenir une estimation des paramètres et de l’erreur à partir de la prédiction faite. Si jamais le modèle comporte des erreurs, cette étape de mise à jour permettra de les rectifier.

Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

De façon générale, l’extraction des paramètres acoustiques est une étape très importante dans les systèmes de reconnaissances automatique de locuteurs. Son but essentiel est d’extraire les données pertinentes à l’étape de modélisation statistique, et minimise ainsi les données redondantes et le bruit qui se présent dans un signal de parole. Il est à noter qu’il existe d’autre méthodes et variantes qui ne sont pas présenté dans cette partie.

### 1.3.3 Modélisation

L’étape de modélisation exploite les données fournis dans l’étape de la paramétrisation afin de créer la représentation d’un individu qui servira, par la suite, à l’authentifier. Le modèle utilisé est généralement une représentation statistique des données acquises.

On peut distinguer quatre grandes approches pour la construction des modèles de locuteur : l’approche vectorielle, statistique, prédictive et connexionniste.

####  Approche vectorielle

Dans l'approche vectorielle, un modèle de locuteur est un ensemble de vecteurs de paramètres représentatifs de l'espace acoustique construit lors de la phase de paramétrisation des signaux d'apprentissage. Lors de la reconnaissance, une distance entre cet ensemble de vecteurs et les vecteurs de paramètres issus des signaux de test est calculée.

L'approche vectorielle compte deux grandes techniques : la programmation dynamique et la quantification vectorielle.

#####  La programmation dynamique

La programmation dynamique (Dynamic Time Warping : DTW) consiste à aligner temporellement une séquence de vecteurs de paramètres de test avec une séquence de vecteurs d'apprentissage. Dans ce cas, le modèle de locuteur est tout simplement l'ensemble des vecteurs de paramètres obtenus après paramétrisation des signaux d'apprentissage. Une distance est calculée entre vecteurs d'apprentissage et de test et moyennée sur l'ensemble de la séquence.

La programmation dynamique est utilisée exclusivement en mode dépendant du texte, c’est une approche très rapide et fournie des résultats relativement bonne, mais elle est très sensible à la qualité d'alignement et notamment au choix du point de départ.

#####  La quantification vectorielle

La quantification vectorielle (Vector Quantization : VQ) repose sur un partitionnement de l'espace acoustique en sous-espaces. Chaque sous-espace est associé à leur vecteur centroïde (i.e. à un vecteur de paramètres représentant l'ensemble des vecteurs composant le sous-espace). Dans ces conditions, un modèle de locuteur est composé d'un ensemble de vecteurs centroïdes, appelé dictionnaire de quantification (codebook).

Lors de la phase de reconnaissance, une distance est calculée entre un vecteur de test et chaque vecteur centroïde du dictionnaire. La distance minimale est assignée au vecteur de test. La distance d'une séquence de vecteurs de test est obtenue par moyenne des distances minimales attribuées à chacun des vecteurs de test.

La quantification vectorielle s'applique en mode dépendant ou indépendant du texte. La rapidité et les performances de cette technique dépendent fortement de la taille du dictionnaire : plus la taille du dictionnaire augmente, meilleures sont les performances sinon, le processus devient plus lent.

####  Approche statistique

L'approche statistique consiste à représenter une séquence de vecteurs acoustiques issus de la paramétrisation par des statistiques à long terme. Les premiers travaux suggèrent d'utiliser les paramètres du spectre moyen à long terme comme un seul modèle des locuteurs. Lors de la reconnaissance, le spectre moyen estimé sur les vecteurs de test est comparé, à l'aide d'une distance spectrale, au spectre moyen issu de l'apprentissage.

Par la suite, l'approche statistique a été enrichie par l'introduction de statistiques d'ordre supérieur (statistiques d'ordre 2) qui permettent notamment de caractériser la variation des paramètres acoustiques (matrice de covariance).

#####  Méthodes statistiques du second ordre

Le principe des Méthodes Statistiques du Second Ordre (MSSO) est de représenter une séquence de vecteurs acoustiques par une distribution gaussienne multidimensionnelle. Le modèle d'un locuteur se résume alors par le triplet {𝑥̅, *X*0, M} où 𝑥̅ est un vecteur moyen, *X*0 est une matrice de covariance, tous deux estimés à partir de la séquence de M vecteurs acoustiques.

L'avantage majeur des MSSO est leur simplicité de mise en œuvre, elles sont performantes sur de courtes durées (3 secondes), et ne capturent que les caractéristiques stables le long du signal de parole. Les variations locales sont, quant à elles, moyennées et ne sont pas prises en compte par les modèles.

#####  Mélange de gaussiennes

Un moyen de pallier ce problème (variations locales moyennées par les MSSO) est de considérer les modèles à mélanges de gaussiennes multidimensionnelles (Gaussian Mixture Model : GMM) [21]. Dans ce contexte, une séquence de vecteurs acoustiques d'apprentissage est représentée par un mélange de gaussiennes i.e. une somme pondérée de M distributions gaussiennes multidimensionnelles, chacune caractérisée par un vecteur moyen et une matrice de covariance.

Lors de l'apprentissage, les paramètres des modèles de locuteur (vecteur moyen 𝑥̅i, matrice de covariance Σ𝑖, pondération pi de chaque distribution gaussienne) sont généralement Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

estimés à l'aide de l'algorithme EM (Expectation-Maximization) couplé à l'approche par Estimation du Maximum de Vraisemblance (EMV).

Par les performances qu'ils obtiennent, les mélanges de gaussiennes sont considérés comme la modélisation « état de l'art » des systèmes de RAL en mode indépendant du texte. L'inconvénient majeur de cette technique est la quantité de signaux d'apprentissage requise pour une bonne estimation des paramètres des modèles.

#####  Modèles de Markov cachés

Les modèles de Markov cachés (Hidden Markov Models : HMM) permettent de caractériser les variations temporelles du signal de parole. Ils reposent sur une succession d'états associés à des probabilités de transition d'un état à l'autre. Une ou plusieurs distributions de probabilité associées à chaque état caractérisent les probabilités d'émission des vecteurs acoustiques par un état.

Lors de la reconnaissance, la vraisemblance pour qu'une séquence de vecteurs de test soit issue de la chaîne de Markov est calculée.

***NB :*** Les mélanges de gaussiennes peuvent être considérés comme un modèle de Markov caché à un seul état. De même, la quantification vectorielle décrite précédemment est souvent interprétée comme une dégénérescence des modèles de Markov cachés à un seul état pour lequel les probabilités d'émission sont remplacées par des mesures de distance.

Les modèles de Markov cachés s'appliquent parfaitement au mode dépendant du texte, obtenant d'excellents résultats. En revanche, l'utilisation des modèles HMM en mode indépendant du texte n'améliore pas les performances obtenues par des modèles plus simples à base de GMM.

####  Approche prédictive

L'approche prédictive repose sur le principe qu'une trame de signal peut être prédite par la seule observation des trames précédentes. De par ce concept, cette approche est considérée dans la littérature comme une approche dynamique i.e. une approche tenant compte des informations dynamiques véhiculées par le signal de parole. Elle s'appuie principalement sur l'estimation d'une fonction de prédiction, propre à chaque locuteur et apprise sur les signaux d'apprentissage. Lors de la reconnaissance, une erreur de prédiction peut être calculée entre une trame prédite (par la fonction de prédiction) et la trame réellement observée dans la séquence de test.

Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

####  Approche connexionniste

L'approche connexionniste repose sur la discrimination entre locuteurs. Elle consiste à fournir à un réseau de neurones un ensemble de signaux de parole issus d'une population de locuteurs afin que ce dernier apprenne comment discriminer un locuteur des autres. L'approche connexionniste se résume, par conséquent, à une tâche de classification. Un modèle se présente sous la forme d'un ou plusieurs réseaux de neurones pour lequel la séquence de vecteurs d'apprentissage du locuteur concerné ainsi que celles des autres locuteurs du système sont fournies en entrée. Différents types de modèles de réseaux sont proposés dans la littérature : Les réseaux multicouches (MLP) utilisés au départ ont rapidement présenté des problèmes lors de l’apprentissage, qui devient long et complexe quand le nombre de locuteurs est grand. Pour éviter ce problème, la tâche de classification est divisée en plusieurs sous-tâches de complexité moindre. On peut aller jusqu'à construire un classificateur pour chaque paire de locuteurs. Un apprentissage plus rapide peut également être obtenu en remplaçant les réseaux multicouches par des réseaux RBF (Radial Basis Function). Les réseaux TDNN (Time Delay Neural Networks) permettent quant à eux de prendre en compte l’information dynamique en réalisant la classification sur des segments de plusieurs trames concaténées. Enfin, l’approche LVQ (Learning Vector Quantization) est une méthode de type quantification vectorielle avec apprentissage discriminant des vecteurs de référence à l’aide d’un réseau de neurones.

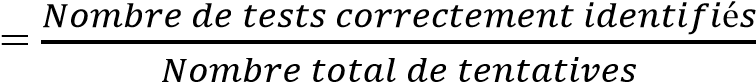
Le principal inconvénient de l'approche connexionniste en reconnaissance est la modularité. En effet, dans le cas d’un apprentissage discriminant, les modèles de tous les locuteurs doivent être réappris quand une nouvelle personne est ajoutée dans la base.

### 1.3.4 Décision et mesures de performances

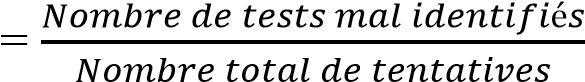
La stratégie mise en jeu dans cette partie dépend essentiellement des deux processus : la vérification et l’identification automatique de locuteur.

####  Identification automatique de locuteur

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

Ic 

Et

Ii 

Avec :

Ic + Ii = 100%

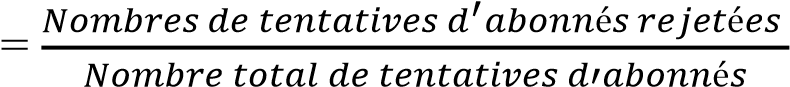
Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

####  Vérification automatique de locuteur

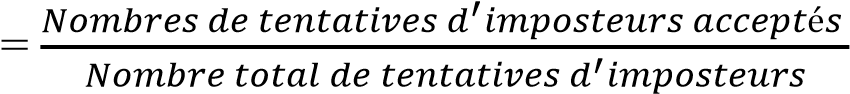
Consiste à vérifier l’adéquation du message vocale avec la référence acoustique du locuteur qu’il prétend être. C’est une décision en tout ou rien. Les performances de vérification de locuteur sont données en termes des faux rejets fr, et de fausses acceptations fa.

***Faux rejet*** : erreur commise lorsque le système rejette, à tort, un locuteur légitime (i.e.

erreur commise lors d’un test de locuteur) ;

fr 

***Fausse acceptation*** : erreur commise lorsqu’un imposteur est malencontreusement accepté en tant qu’utilisateur légitime (i.e. erreur commise lors d’un test imposteur) ;

fa 

## 1.4 Evolution d’un système de reconnaissance automatique de locuteur

Les recherches sur la reconnaissance du locuteur ont été entreprises depuis plus de 50 ans, et continues d’être un domaine actif de traitement de la communication parlée. Le développement de la technologie de la reconnaissance du locuteur est étroitement concomitant avec l’avancement dans la connaissance de la parole, le traitement du signal et la technologie des ordinateurs.

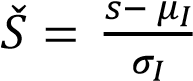
La reconnaissance du locuteur par les humains a été largement étudiée dans les années 1960. La motivation de ces études était d’apprendre comment l’homme reconnait les locuteurs et la fiabilité d’un humain à reconnaitre un locuteur. Le travail le plus important qui a stimulé la recherche sur la reconnaissance du locuteur par la machine a été réalisé par Kersta qui a introduit le spectrogramme (ou il l’a noté comme empreinte vocale) en tant que moyen d’identification personnelle.

3 Dans les années 1970, l’attention a été tournée vers la reconnaissance du locuteur par ordinateur et devient la reconnaissance automatique du locuteur. A cette époque, les systèmes de reconnaissance du locuteur, en général, ne portait que sur une petite population (moins de 20 locuteurs). La transformée de Fourier, les techniques de prédiction linéaire et d’analyse cepstrale ont été appliquées pour générer des paramètres du locuteur. Les moyennes long-terme de ces paramètres ont été utilisées comme références des locuteurs.

Dans les années 1980, des méthodes statistiques de reconnaissance des formes plus compliquées ont été investiguées, par exemple, l’alignement temporel dynamique(DTW) et la quantification vectorielle (VQ), pour des systèmes de reconnaissance du locuteur à grande échelle (>100 locuteurs). La contribution des caractéristiques statiques et dynamiques pour la reconnaissance du locuteur a également été étudiée.

Depuis les années 1990, la mise à disposition de bases de données de parole plus importantes (par exemple, corpus YOHO) a boosté les études sur des modèles plus compliqués Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

pour la représentation des locuteurs. Ces modèles comprennent les modèles stochastiques (par exemple, les modèles de Markov cachés (HMM)), le modèle de mélange de Gaussiennes (GMM), les réseaux de neurones (par exemple, Perceptron Multicouches MLP), fonctions a base radiale (RBF)) et les machines a vecteurs de support (SVM)), …etc. Parmi ces techniques de modelisation, la GMM a été reconnu comme la plus efficace à caractériser la distribution de la densité des données de la parole et a été considérée comme la technique de modélisation dominante pour les systèmes de reconnaissance du locuteur. En ce qui concerne l’extraction des caractéristiques, les coefficients cepstraux incorporant le modèle auditif, connus sous le nom de Coefficients Cepstraux à Fréquence Mel (MFCC) et leurs coefficients dynamiques ont été les caractéristiques ou paramètres dominants. Un système avec les paramètres MFCC, une modélisation GMM est considérée comme un système de référence en mode indépendant du texte pour comparer les nouvelles technologies [25].

Les années 2000 ont connues l’apparition d’une nouvelle famille de techniques de normalisation de scores [12], dans laquelle des scores sont normalisé par la soustraction de la moyenne, divisés par l’écart type des scores imposteurs :(avec 𝑆̌ est le score normalisé, s le score original, 𝜇𝐼 𝑒𝑡 𝜎𝐼 sont la moyenne et l’écart type) Les techniques les plus couramment utilisées sont : Znorm, Hnorm, Tnorm, Htnorm, Cnorm, et Dnorm, en outre, les paramètres de haut niveau (phonèmes, paramètres idiolectaux, sémantique, accent, prononciation…) proposé en 2001 par Doddington, ont été largement utilisées en vérification de locuteur en mode indépendant de texte [11].

## 1.5 Les approches classiques utilisées dans le système de reconnaissance automatique de locuteur

Différentes méthodologies sont utilisées en RAL pour réaliser les références de locuteurs. Les approches génératives regroupent des méthodes qui utilisent les données d’apprentissage pour modéliser les densités de probabilité de chaque classe, par une famille de fonctions paramétriques. L’approche générative dominante pour représenter la référence du locuteur, en RAL indépendante du texte, est le modèle de mélanges de Gaussiennes (GMM,

Gaussian Mixture Model) qui constitue l’état de l’art des systèmes de RAL.

Il existe d’autres approches génératives comme les modèles de Markov cachés (HMM, Hidden Markov Model). Les HMM sont très employés en RAL dépendante du texte car ils sont capables de capturer les dépendances temporelles entre différentes variables aléatoires.

Les approches à base de quantification vectorielle ont été utilisées en RAL. Elles proposent une représentation minimale d’une classe de paramètres observes : un représentant (dans un dictionnaire) pour chaque classe. Chaque classe de paramètres est déterminée par un algorithme de classification du type K-moyennes. Cette représentation est choisie en minimisant la distance entre le centroïde et les paramètres de la population observée. Ces approches ne sont plus très employées depuis l’apparition des GMM en RAL.

Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

L’approche discriminante : les Support Vector Machine (SVM) est largement utilisé en

RAL. A l’origine, ils ont été conçus comme une fonction discriminante permettant de séparer au mieux des régions complexes dans des problèmes de classification à 2 classes. Ils démontrent aujourd’hui des performances similaires à l’approche GMM. Ces deux méthodes sont aussi combinées dans un nouveau formalisme, le GMM/SVM Super-Vecteur qui profite des capacités génératives du GMM et discriminantes du SVM.

## 1.6 Domaines d’application

Nos voix ne sont pas seulement un moyen de communiquer. Elles offrent également un moyen fiable de nous reconnaître, et font partie intégrante de notre identité. C'est la raison pour laquelle les banques et d'autres grandes entreprises se tournent aujourd'hui vers l'authentification vocale.

La voix humaine est unique. Elle est avec nous tout le temps contrairement à nos clés de voitures, et aux mots de passes ou codes PIN qu'on peut très souvent oublier. C'est à la fois cette sécurité et cette simplicité d'usage offerte par l'authentification biométrique vocale qui poussent les banques, les opérateurs de télécommunications et autres grandes organisations à choisir ce mode d'authentification.

La biométrie vocale, tout comme la reconnaissance et la synthèse vocale, s'est d'abord propagée dans les serveurs vocaux automatiques des centres d'appels. Mais aujourd'hui, elle est également utilisée dans des domaines aussi variés que l'authentification mobile et le paiement par cartes de crédit.

* **Sécurisation des applications mobiles**

Les grandes entreprises voient désormais leurs clients utiliser massivement les canaux mobiles pour prendre contact et effectuer les opérations courantes. C'est même devenu une attente forte des clients et des consommateurs. Mais la multiplication des applications et services en ligne fait qu'il devient difficile de gérer tous ces mots de passes, de forme et de tailles différentes. L’authentification vocale devient dès lors le mode d'authentification mobile idéal. Il suffit simplement de donner une simple phrase clé à prononcer à un client pour vérifier son identité.

En plus d'éliminer la frustration née des mots de passe difficiles à mémoriser ou à saisir, le 'login vocal' réinvente véritablement l'authentification mobile. Le mobile devenant de plus en plus le point de contact principal entre un consommateur et un fournisseur de services, améliorer l'expérience utilisateur et la sécurité deviennent une priorité.

* **Sécurisation des transactions à risque par carte de crédit**

La reconnaissance de locuteur constitue aussi une solution sûre et pratique pour vérifier les transactions à risque par carte de crédit (par exemple celles en dehors des habitudes de consommation du client ou de son emplacement géographique habituel). Quand une opération à risque est détectée, une demande de vérification de la transaction peut être envoyée au titulaire de la carte de crédit, via un appel sortant automatique, sur son téléphone portable. Le détenteur est alors invité à prononce une phrase clé : "J'autorise cette transaction par ma signature vocale." Chapitre 1 : Etat de l’art sur les systèmes de reconnaissance automatique du locuteur

A l'inverse, si la transaction est suspecte, il peut tout aussi facilement rejeter celle-ci, ce qui permet alors à l'institution financière d'investiguer sur les transactions marquées comme suspectes.

* **Paiement en ligne**

La reconnaissance de la voix peut être utilisée pour sécuriser des paiements en ligne, typiquement des paiements à risque tels que le premier paiement en ligne sur un site d'ecommerce, par exemple le transfert de l’argent ou des opérations importantes. Lorsque ces opérations sont effectuées, un appel sortant automatique est émis vers le téléphone portable du titulaire du compte effectuant l'opération. Si cette opération est valide, l'utilisateur est invité à confirmer le paiement de la même façon qu'il peut confirmer l'achat par carte de crédit.

* **Aide aux handicapés**

La reconnaissance de locuteur est très utile dans ce cas, elle offre la possibilité de saisir les données à la voix, commandes vocales (ouverture porte, contrôle des équipements au domicile).

## Conclusion

Dans ce chapitre, nous avons présenté de façon générale l’état de l’art d’un système de reconnaissance automatique de locuteur, nous avons présenté, également, la structure générale d’un système RAL et ses composants modulaires. Pour chaque module, nous avons décrit les différentes techniques utilisées en citant leurs avantages et leurs faiblesses, et nous avons terminé par la présentation des domaines d’application de cette discipline

1. Intelligibilité de la parole : signifie la capacité de comprendre un message linguistique contenu dans un signal de parole.

   [↑](#footnote-ref-1)