USTHB **Faculté d'Electronique et d'Informatique**

- Septembre 2003 -

- Résumé de Thèse -

Thèse de Doctorat préparée par **Mr Sayoud Halim**, sous la direction de Mme Malika Boudraa,

Titre : **Reconnaissance Automatique du Locuteur**
-Approche connexionniste-

Résumé

De nos jours, un grand nombre d'applications nécessitent une phase d'authentification de l'utilisateur. Cette authentification peut être réalisée au moyen d'une carte à puce et d'un code confidentiel (retrait à un guichet automatique), au travers d'empreintes digitales (accès sécurisé à des locaux) ou d'empreintes génétiques (domaine juridique) ou encore grâce à la voix (serrure vocale). Dès lors qu'une application est accessible à distance (par le réseau téléphonique par exemple), la voix reste le seul élément disponible pour authentifier une personne. Cette thèse s'inscrit dans le cadre de la Reconnaissance Automatique du Locuteur dont l'objectif est de reconnaître une personne par l'analyse de sa voix. Il s'agit d'un domaine de recherche non encore maîtrisé.
Nous avons ainsi élaboré plusieurs types de méthodes émanant de deux visions différentes : une vision statistique et une vision connexionniste. A partir de ces deux types de visions, nous avons conçu plusieurs méthodes pour la reconnaissance automatique du locuteur. Nous en citons :

- Une méthode statistique basée sur la SOSM pour l'identification du locuteur
- Trois méthodes connexionnistes basée sur la LVQ pour l'identification du locuteur
- Une méthode connexionniste basée sur le MLP pour la vérification du locuteur

Des tests expérimentaux ont été faits sur ces méthodes et nous avons pu noté les constats suivants :

- La méthode statistique, basée sur les mesures de similarité du 2nd ordre, nécessite peu d'apprentissage. Cette méthode testée, en l'identification du locuteur indépendante du texte, a donné d'excellents résultats (précision de 100%). De plus cette méthode est évolutive et reste très simple à mettre en œuvre.
- Les méthodes connexionnistes basées sur la LVQ ont été testées en identification dépendante du texte. La complexité de ces méthodes est moyenne. Les résultats de reconnaissance sont assez bons sur la base de données testée. Mais les performances attendues sur des bases de données plus grandes devraient être moins appréciables.
- La méthode connexionniste basée sur le MLP a été testée en VAL (Vérification Automatique du Locuteurs). La méthode est assez complexe et son temps d'apprentissage est relativement élevé. Les résultats de vérification sont bons sur la base de données testée. Ces résultats, néanmoins, restent insuffisants pour des applications de vérification industrielles qui exigent des taux d'erreur proches de 0%.

Par ailleurs, nous avons entamé de nouvelles études expérimentales, pour l'optimisation de nos systèmes de RAL, qui ont bien enrichi le contenu de cette thèse. Nous en citons :

- Etude de robustesse de la méthode statistique : Des tests d'identification du locuteur ont été faits en milieu bruité par trois types de bruits différents et à différents RSBs.
- Recherche de la résolution spectrale optimale pour la méthode statistique : Plusieurs dimensions spectrales ont été testées sur la bande 0-8kHz et la bande téléphonique 300-3400Hz afin de trouver la meilleure résolution spectrale à adopter en RAL.
- Mise au point d'une nouvelle métrique adaptée aux caractéristiques hétérogènes : Nous avons appelé cette nouvelle métrique ODHEF. Celle-ci a permis d'associer des caractéristiques acoustiques différentes pouvant être utilisées simultanément en RAL.
- Conception d'une nouvelle méthode pour l'IAL à base de LVQ et qui est entièrement prosodique : Cette nouvelle méthode est originale car elle est basée exclusivement sur trois paramètres prosodiques simples tout en donnant de bons résultats en identification automatique du locuteur (IAL).
- Proposition d'un modèle de réseau de neurone mono-locuteur à base de MLP pour la VAL : Nous avons proposé un réseau MLP mono-locuteur (un réseau par locuteur) dédié aux taches de la vérification automatique du locuteur (VAL). Plusieurs dimensions et plusieurs vecteurs d'entrée ont été expérimentés ici.
- Amélioration de la précision du MLP mono-locuteur en jouant sur ses paramètres de sortie : Les résultats de la vérification du locuteur ont été nettement améliorés après une étude engagée sur l'effet même des paramètres de sortie du MLP.
- Mise en place d'un ensemble d'indications servant de conseils pour l'utilisation de ces méthodes en RAL : Cette thèse ne constitue pas seulement un ensemble de théories et d'observations expérimentales ; mais on a voulu aussi qu'elle rassemble un ensemble d'indications et de conseils pratiques aux chercheurs œuvrant dans ce domaine.

Enfin, pour plus de détail et d'éclaircissement, nous avons rajouté à la thèse un chapitre entier décrivant l'état de l'art en RAL ainsi que la tendance actuelle de la recherche.

I. Introduction

L'expression vocale est une caractéristique propre d'un locuteur : ainsi est-il possible, dans des conditions normales, de reconnaître son correspondant au cours d'une conversation téléphonique. Comment cela se passe-t-il ? Au fait, le cerveau, tel que créé par notre Créateur, comprend des modules extrêmement complexes de traitement du signal qui permettent de reconnaître simultanément ce qui est dit et celui qui parle…

Ces variations individuelles entre locuteurs ont deux origines essentielles. En premier lieu, les caractéristiques physiques de l'appareil de phonation influencent les formants et la valeur moyenne du pitch et cela indépendamment de la phrase prononcée. D'autre part, une même phrase n'est pas prononcée de la même façon par deux locuteurs ; on observe des différences dans les débits d'élocution, dans l'étendue des variations du Pitch,….

La reconnaissance automatique du locuteur est un terme générique pour discriminer parmi plusieurs personnes en fonction de leurs voix. Il convient dans ce domaine de recherche de reconnaître non pas ce qui a été dit, mais de reconnaître l'identité de la personne qui parle, à partir de ses caractéristiques vocales. On distingue en général l'identification du locuteur et la vérification du locuteur :

L'identification consiste à reconnaître un locuteur appartenant à une population de plusieurs locuteurs ; on compare pour cela son expression vocale à des références connues. La vérification consiste à accepter ou à refuser une identité proclamée par un locuteur ; dans ce but on compare à un certain seuil la distance entre son expression vocale et sa référence personnelle.

On distingue également, selon la phrase prononcée, deux types de reconnaissances : La reconnaissance indépendante du texte (free-text speaker recognition) ; les techniques d'identification actuelles tendent à s'intéresser à ce type de reconnaissance. Et la reconnaissance effectuée sur la base d'un texte imposé ; cette dernière procédure, par contre, est la plus courante pour la vérification du locuteur (fixed-text speaker vérification).

Dans le cadre de cette thèse, nous avons développé, deux types d'approches:

- Une approche statistique basée sur les mesures statistiques d'ordre 2.

et

- Une approche Connexionniste correspondant à deux méthodes différentes :
  une méthode basée sur la LVQ et une méthode basée sur le MLP.

Plusieurs études expérimentales ont été faites sur ces deux approches et de nombreuses conclusions en ont été déduites.

Ce résumé de thèse est organisé de la manière suivante :

La section II décrit la problématique et résume l'état de l'art en reconnaissance du locuteur.

La section III illustre la méthode statistique proposée pour l'identification automatique du locuteur.

La section IV illustre l'approche connexionniste proposée pour l'identification automatique du locuteur.

Les sections V, VI et VII présentent les résultats obtenus durant les différents tests effectués en reconnaissance du locuteur.

Plusieurs discussions ainsi que des conclusions sont apportées à la section VIII.

Enfin, des références bibliographiques sont mises à la disposition du lecteur (à la fin du manuscrit).

II. La Reconnaissance Automatique du Locuteur : Définition et applications

Cette section est une introduction au domaine de la RAL (Reconnaissance Automatique du Locuteur). elle présente tout d'abord les différentes tâches liées à la RAL telles que l'Identification et la Vérification Automatique du Locuteur ou des tâches plus récentes comme le suivi de locuteur ou l'Indexation par Locuteur de flux audio. Les principes ainsi que les techniques afférentes à ces différentes tâches sont décrits brièvement. Les divers problèmes de la RAL sont aussi exposés comme la variabilité intra-locuteur ou la variabilité due au matériel. Finalement, un point est donné sur les dernières tendances du domaine.

II.1 La Reconnaissance Automatique du Locuteur

II.1.1 Généralités

La caractérisation automatique du locuteur est un vaste domaine dans lequel la « machine » a pour tâche d'extraire du signal de parole les informations de nature à renseigner sur les spécificités d'un individu : identité, caractéristiques physiques, émotivité, état pathologique, particularités régionales, etc. Elle s'applique à différents thèmes de recherche traitant des informations véhiculées par la voix tels que la classification d'individus, ou l'étude psychique ou physiologique d'une personne.

La Reconnaissance Automatique du Locuteur - RAL - est un sous-problème de la caractérisation automatique du locuteur. Son objectif est de reconnaître l'identité d'une personne à l'aide de sa voix. La variabilité de la parole entre locuteurs (variabilité inter-locuteur) est l'essence même de la RAL. Sans cette variabilité, il serait impossible de reconnaître une voix parmi plusieurs voix possibles.

La RAL, contrairement à la Reconnaissance Automatique de la Parole (RAP) s'intéresse tout particulièrement aux informations extra-linguistiques véhiculées par un signal vocal (signal de parole). Pourtant, la RAL a très souvent bénéficié des avancées de la RAP. Ainsi, de nombreuses techniques ont été appliquées en RAP avant d'être adaptées au domaine de la RAL. Finalement, les applications de la RAL sont principalement liées aux problèmes d'authentification ou de confidentialité.

II.1.2 Différentes Tâches en RAL

L'Identification Automatique du Locuteur et la Vérification Automatique du Locuteur sont les tâches pionnières du domaine de la RAL [Atal, 1976], [Doddington, 1985], [O'Shaughnessy, 1986], [Rosenberg et al., 1991], [Naik, 1994], [Furui, 1994], [Furui, 1997], [Doddington, 1998].

Plus récemment, les besoins applicatifs ont fait naître de nouvelles tâches comme l'Indexation par Locuteur de flux audio [Johnson, 1999], [Delacourt, 2000] ou le Suivi de Locuteurs (ou speaker tracking) [Rosenberg et al., 1998a], [Sonmez et al., 1999], [Bonastre et al., 2000a], [Bonastre et al., 2000b], [Martin et al., 2000] ou de nouvelles variantes telles que la détection d'un locuteur dans une conversation [Przybocki et al., 1999], [Martin et al., 2000].

II.1.2.1 Identification Automatique du Locuteur

L'Identification Automatique du Locuteur (IAL) est le processus qui consiste à déterminer, parmi une population de locuteurs connus, la personne ayant prononcé un message donné. D'un point de vue schématique (voir figure 1), une séquence de parole est donnée en entrée du système d'IAL. Pour chaque locuteur connu du système, la séquence de parole est comparée à une référence caractéristique du locuteur : identité du locuteur dont la référence est la plus proche de la séquence de parole est donnée en sortie du système d'IAL.

Deux modes sont proposés en IAL :

- l'identification en ensemble fermé pour lequel on suppose que la séquence de parole est effectivement prononcée par un locuteur connu du système ;
- et l'identification en ensemble ouvert pour lequel le locuteur peut ne pas être connu.

En mode «ensemble ouvert», le système d'IAL doit décider de la fiabilité de son jugement en acceptant ou rejetant identité qu'il a trouvée. De par son principe - déterminer une identité parmi les identités potentielles - les performances des systèmes d'IAL se dégradent généralement au fur et à mesure que la population de locuteurs augmente.

Applications

En IAL, les applications sont peu nombreuses. On peut retenir, par exemple, l'utilisation d'un système d'IAL en vue de faciliter l'adaptation au locuteur des systèmes de RAP. Par ailleurs, il peut être intéressant pour des applications commerciales d'associer un même mot de passe pour une petite population de locuteurs (membres d'une famille, d'une société). Dans une telle situation, un système d'IAL en ensemble ouvert et dépendant du texte peut être utilise pour contrôler l'accès à des données sensibles, a un réseau ou à un bâtiment [Rosenberg et al., 1998b].

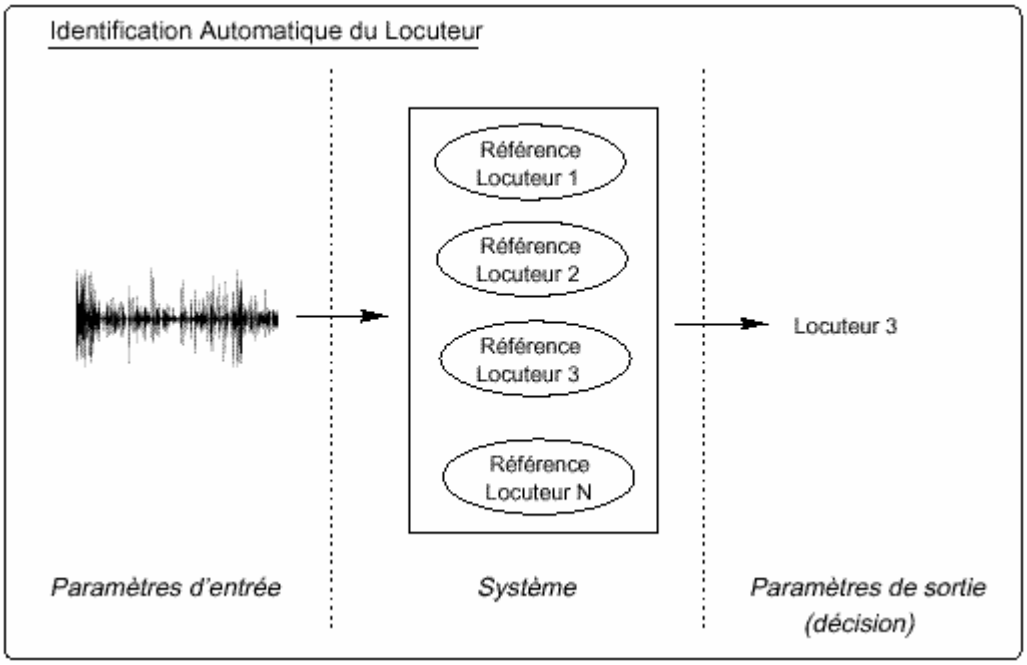

Figure 1 La tâche d'IAL. Principe de base de la tâche d'Identification Automatique du Locuteur.

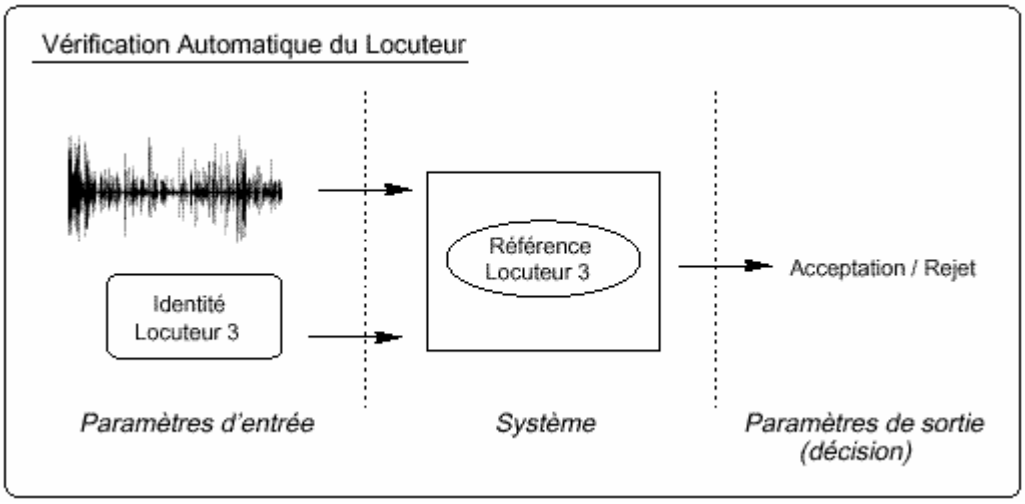

Figure 2 La tâche de VAL. Principe de base de la tâche de Vérification Automatique du Locuteur.

II.1.2.2 Vérification Automatique du Locuteur

La Vérification Automatique du Locuteur (VAL) est le processus décisionnel permettant de déterminer, au moyen d'un message vocal, la véracité de identité revendiquée par un individu (figure 2). L'identité ainsi que le message vocal constituent les deux entrées du système de VAL. L'identité, nécessairement connue du système, désigne automatiquement la référence caractéristique d'un locuteur. Une mesure de similarité est calculée entre cette référence et le message vocal puis comparée à un seuil de décision. Dans le cas où la mesure de similarité est supérieure au seuil, l'individu est accepté. Sinon., l'individu est considéré comme un imposteur et rejeté.

Applications

Les applications de VAL sont multiples et principalement commerciales [Boves, 1998] :
- serrures vocales pour le contrôle d'accès à des locaux ;
- authentification pour l'accès à distance à des données sensibles ou à des services spécifiques à travers le réseau téléphonique (consultations ou transactions bancaires, consultations de bases de données à caractère confidentiel, consultations de boites vocales, télé-achat, etc.) ;
- protection de matériel contre le vol (téléphones portables, voitures, etc.) ;
- incarcération à domicile nécessitant une authentification régulière du prévenu.

II.1.2.3 Détection de Locuteurs

La détection de locuteurs dans un flux audio est une variante de la VAL. Sa particularité est de considérer un flux audio composé de séquences de parole produites par plusieurs locuteurs (conversations, débats, conférences, etc.). Dans ce contexte, la tâche de détection consiste à déterminer si un locuteur donné intervient ou non dans le document audio. Dans le cas d'un flux audio mono-locuteur, la tâche de détection se résume à la tâche de vérification.

Applications

La tâche de détection est évidemment motivée par les instances militaires ou judiciaires. Néanmoins, elle demeure très intéressante dans le domaine de l'indexation de documents audio pour laquelle la détection d'un locuteur connu peut permettre de cibler plus facilement un document audio particulier (séquence d'un journal télévisé ou d'une émission radio).

II.1.2.4 Indexation par Locuteur et ses variantes

La tâche d'Indexation Automatique par Locuteur [Say, 03'''] consiste à cibler les interventions des locuteurs dans un flux audio. En d'autres termes, indexer un document audio en locuteurs revient à indiquer à quel moment un individu prend la parole et qui est cet individu. La seule entrée d'un système d'indexation est le document audio à indexer. Aucune information n'est donnée au système concernant le nombre de locuteurs présents dans le document ou leur identité.

Contrairement aux systèmes d'IAL ou de VAL, les systèmes d'indexation ne détiennent pas de référence pour les locuteurs présents dans un document audio. Leur principe repose généralement sur une phase de segmentation "aveugle" en locuteurs suivie d'une phase de regroupement. Un système d'IAL permet finalement d'identifier les différents locuteurs présents dans le document. La sortie d'un système d'indexation ressemble généralement à la séquence suivante : le locuteur A est intervenu aux instants t 1, t 4, t 6, le locuteur B aux instants t 2, t 5, le locuteur C à l'instant t 3.

La tâche de suivi de locuteurs peut être considérée comme une version simplifiée de l'Indexation par Locuteur d'un flux audio. Le principe reste le même : déterminer les interventions d'un ou plusieurs locuteurs, appelés locuteurs cibles, dans un flux audio. La simplification réside dans le fait que le système de suivi de locuteurs connaît nécessairement les locuteurs présents dans le document à indexer ou, du moins, ceux dont il doit détecter les interventions. Il possède une référence caractéristique pour chacun des locuteurs.
Malgré cette simplification, le suivi de locuteurs reste une tâche très complexe. Trois grandes approches sont recensées dans la littérature :

1. Une segmentation "aveugle" en locuteurs, identique à celle employée pour l'Indexation par Locuteur d'un flux audio, est appliquée sur le signal de test. Les segments - résultat de la segmentation - sont soumis à un système de VAL classique afin de déterminer les segments appartenant effectivement au locuteur cible [Bonastre et al., 2000b].
2. Le signal de test est découpé en une suite de blocs de trames, de taille fixe (ce découpage selon une taille fixe de blocs est entièrement indépendant des événements acoustiques observés sur le signal de parole.), sur lesquels sont appliques un système de VAL. Un processus de décision, à base de seuils, permet en phase finale d'accepter ou de rejeter les blocs appartenant au locuteur cible [Rosenberg et al., 1998a], [Bonastre et al., 2000a].
3. La troisième approche est similaire à la précédente excepté pour le processus de décision. Dans ce cas, la décision repose sur un HMM ergodique composé d'états correspondant au locuteur cible, à un modèle générique de parole et à un modèle générique de non parole (silence, bruit...) [Sonmez et al., 1999],[Meignier et al., 2000].

Applications
Les systèmes d'Indexation Automatique par Locuteur d'un flux audio sont principalement utilisés pour le traitement de bases de données audio (recherche de séquences d'émissions télévisées ou radiophoniques par le suivi du présentateur, estimation du temps de parole de chaque intervenant lors de débats, etc.). D'autres applications sont envisageables comme la recherche de messages par locuteur sur un répondeur téléphonique ou sur une boite vocale.

II.1.2.5 Applications criminalistiques
Un volet que nous n'avons pas encore évoque est l'utilisation de la RAL dans les domaines judiciaires ou criminalistiques [Hollien, 1990], [Kunzel, 1994], [Boe, 1998], [Champod et al., 1998]. Il s'agit par exemple de rechercher un individu parmi une population de suspects potentiels (tâche d'IAL) ou encore de comparer un enregistrement vocal issu d'une écoute téléphonique à la voix d'un suspect potentiel (tâche de VAL).
Dans ce contexte, il est important de souligner que la voix est très souvent assimilée, à tort, à une empreinte vocale au même titre que les empreintes digitales ou génétiques et peut constituer une preuve dans une procédure pénale. **Ce terme d'empreinte vocale est une aberration** sachant que la voix ne possède pas de caractéristiques qui peuvent la rendre unique [Boe, 1998], [Boe et al., 1999].

II.1.3 Mise en place d'un système de RAL
L'utilisation d'un système de RAL pour une application donnée (hormis pour la tâche d'Indexation Automatique par Locuteur d'un flux audio 2) se décompose en deux phases distinctes. La première phase est nécessaire à la construction des références ou modèles de chaque locuteur connu du système i.e. de chaque client de l'application. Elle consiste à collecter, auprès de ces clients, des signaux de parole dits d'apprentissage, lors de sessions d'enregistrement. La seconde phase est la phase de reconnaissance à proprement parler qui consiste, pour un client, à se présenter devant le système de RAL (phase de test).

II.1.4 Problèmes rencontrés en RAL
Le signal de parole est un signal très complexe où se mêlent informations linguistiques, informations caractéristiques du locuteur, informations relatives au matériel utilisé pour la transmission ou l'enregistrement du signal, etc. En outre, le signal de parole est très redondant. Cette caractéristique est d'ailleurs reconnue pour faciliter la communication entre deux personnes dans un environnement très bruyant ("cocktail party"). Par ces différents aspects, le signal de parole présente une très grande variabilité. La capacité des systèmes de RAL à différencier plusieurs individus repose essentiellement sur la variabilité interlocuteur i.e. la disposition du signal de parole à varier entre différents individus. Néanmoins, le signal de parole renferme d'autres types de variabilités qui rendent problématique la tâche de reconnaissance, telles que la variabilité intra-locuteur ou la variabilité due au matériel.
Par ailleurs, les systèmes de RAL doivent faire face à d'autres difficultés liées davantage au domaine applicatif, comme l'utilisation des systèmes dans des conditions difficiles, les tentatives d'imposture, etc.

Variabilité due au locuteur

Si le signal de parole est variable entre deux individus, il varie également pour un même individu. Cette variabilité intra-locuteur est induite par l'évolution naturelle ou volontaire de la voix d'une personne. Cette évolution peut être :

- Ponctuelle ou à très court terme.
  Etat pathologique (fatigue, rhume, etc.) ou émotionnel (stress) [Homayounpour, 1995], [Scherer et al., 1998], [Karlsson et al., 1998], [Banziger et al., 2000] d'une personne provoquent des altérations momentanées dans sa voix. Dans ce sens, la voix d'une personne peut évoluer entre le début et la fin de la journée (fatigue, irritation due à la pollution). D'autre part, il est impossible pour un individu de répéter consécutivement deux phrases identiques et de produire un même signal de parole pour ces deux phrases. Une légère variation est toujours observée. Finalement, une personne a la possibilité de modifier volontairement sa voix.
- A moyen terme.
  En RAL, le comportement d'un individu se modifie au fur et à mesure de son utilisation du système. L'individu devient de plus en plus confiant et sa voix évolue dans ce sens.
- A long terme. La voix change au fur et à mesure du vieillissement d'une personne.

La variabilité intra-locuteur pose le problème de la représentativité des signaux de parole collectés lors des sessions d'enrôlement (et des modèles des locuteurs correspondant) au sein des systèmes de RAL.
Des travaux ont montré que les performances d'un système sont très fortement corrélées au temps qui sépare les sessions d'enrôlement et les tests [Furui, 1977], [Rosenberg, 1976], [Setlur et al. 1994]. Plus ce temps augmente, plus les performances se dégradent. Néanmoins, même les variations à court terme (émotion, état pathologique) peuvent être très préjudiciables aux systèmes de RAL.

Variabilité due au matériel

Le signal de parole est porteur d'informations caractérisant le matériel utilisé lors de sa capture (ex : microphone, combine téléphonique), de sa transmission (ex : lignes téléphoniques, air ambiant) et de son enregistrement (ex : microphones, convertisseurs). Ces informations apparaissent le plus souvent sous la forme de déformations/dégradations du signal de parole. Ces déformations sont différentes selon le type de matériel utilisé. Si la bande téléphonique est reconnue pour dégrader les performances des systèmes de RAL, elle n'est pas la seule responsable. En effet, 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 [Van Vuuren, 1996].
Par exemple, dans [Reynolds, 1996] et [Auckenthaler et al., 2000], il est démontré que des différences de types de combines téléphoniques entre l'apprentissage et le test sont une des causes de ces dégradations.

Robustesse en environnements difficiles

Comme nous venons de l'évoquer, les environnements téléphoniques mettent à rude épreuve les systèmes de RAL. Néanmoins, d'autres environnements nécessitent de la part des systèmes de RAL une grande robustesse. Le réseau GSM est considéré ici comme un environnement à part entière, en marge des environnements téléphoniques.
Des travaux expérimentaux sur la comparaison du réseau téléphonique classique et d'un réseau GSM font état de différences significatives dans la qualité des signaux [Fissore et al., 1999]. En effet, les signaux transmis par réseau GSM montrent un niveau de bruit bien supérieur (les appels par téléphones mobiles sont souvent effectués dans des endroits plus bruyants que ceux d'un téléphone fixe : voiture, gare, rue), un niveau de voix plus élevé, souvent proche de la saturation entraînant des distorsions au sein du signal ainsi que de potentielles dégradations des signaux dues au codage de la parole. Jusqu'à présent, très peu de travaux ont porte sur la robustesse des systèmes de RAL à travers le réseau GSM [Besacier et al., 2000b], [Quatieri et al., 2000]. La raison est sans doute le manque de bases de données dédiées à cette problématique. Néanmoins, cette tendance est en train de s'inverser avec le développement toujours croissant de la téléphonie portable. Finalement, les systèmes de RAL doivent renforcer leur robustesse face au bruit ambiant. En effet, d'une manière similaire à la variabilité intra-locuteur ou à la variabilité due aux changements de matériel, la variabilité du niveau de bruit entre apprentissage et test peut susciter une baisse de performances des systèmes de RAL.

Tentatives d'imposture - locuteurs non coopératifs

Selon l'application visée, un système de RAL peut faire l'objet d'attaques d'individus usurpant l'identité de quelqu'un d'autre. Ces attaques (ou tentatives d'imposture) peuvent, par exemple, avoir pour dessein des transactions frauduleuses sur le compte bancaire d'un client ou accès à des données confidentielles. Un système de RAL doit par conséquent être robuste face à de telles attaques [Homayounpour, 1995].
Dans un contexte judiciaire, le système de RAL peut être soumis à des locuteurs non-coopératifs i.e. des locuteurs qui ne désirent pas être reconnus par le système. Dans ce cas de figure, les locuteurs tentent fréquemment de transformer leur voix.

II.2 Structure des systèmes de RAL et techniques associées

Un système de RAL, quelle que soit la tâche considérée, se résume à l'enchaînement de trois processus principaux qui sont : la paramétrisation, la reconnaissance et la décision. Contrairement au processus de paramétrisation (généralement basé sur des techniques communes à d'autres domaines comme la RAP), les principes mis en oeuvre pour la reconnaissance et la décision sont étroitement lies à la tâche visée.
Le processus de reconnaissance est différent selon qu'il repose sur une modélisation des caractéristiques des locuteurs connus du système (modèles clients pour les tâches d'IAL, de VAL ou de suivi de locuteurs) ou non (Indexation par Locuteur d'un flux audio).

Dans les sections suivantes, nous nous intéressons au premier cas de figure. Le lecteur pourra se reporter à [Delacourt, 2000] pour un état de l'art du processus de reconnaissance (Segmentation-Regroupement-Identification) employé pour la tâche d'Indexation par Locuteur d'un flux audio. D'une manière similaire, le processus de décision est présenté tâche par tâche.

II.2.1 Paramétrisation acoustique

Le processus de paramétrisation consiste à extraire du signal de parole les informations pertinentes en vue de la reconnaissance. Le signal de parole, de par sa complexité (multitudes d'informations et redondance), ne peut être exploité directement. Une représentation simplifiée du signal de parole est par conséquent nécessaire Cette représentation repose généralement sur des vecteurs de paramètres acoustiques, calculés périodiquement sur le signal de parole.
La première étape de la paramétrisation acoustique consiste à décomposer le signal de parole, à cadence régulière (ex : toutes les 10 millisecondes), en trames de signal (d'une longueur variant généralement de 20 à 31,5 millisecondes). Un traitement particulier est ensuite appliqué à ces trames afin de produire les vecteurs de paramètres acoustiques. La littérature propose un grand nombre de traitements selon la nature des informations à extraire du signal de parole. On considère généralement trois grandes classes de paramètres : les paramètres de l'analyse spectrale, les paramètres prosodiques et les paramètres dynamiques. Néanmoins, d'autres classifications sont envisageables.

II.2.1.1 Paramètres de l'analyse spectrale

L'analyse spectrale est l'analyse la plus employée en RAL. Les paramètres qui en découlent sont généralement représentatifs des caractéristiques physiques de l'appareil phonatoire (forme du conduit vocal) de chaque individu. De multiples paramètres ont été étudiés dans la littérature (le lecteur se reportera aux travaux suivants [Reynolds, 1994], [Homayounpour et al., 1994] et [Charlet, 1997] pour une description rapide et une comparaison de différents paramètres). Nous citons ici les plus pertinents en RAL :

- coefficients issus d'une analyse par prédiction linéaire [Grenier, 1977] : LPCC (Linear Predictive Cepstral Coefficients) ou LPC (Linear Predictive Coefficients) ;
- coefficients spectraux issus d'une analyse en banc de filtres 3 : LFSC (Linear Frequency Spectral Coefficients) ou MFSC (Mel Frequency Spectral Coefficients) ;
- coefficients cepstraux issus d'une analyse en banc de filtres : LFCC (Linear Frequency Cepstral Coefficients) ou MFCC (Mel Frequency Cepstral Coefficients).

II.2.1.2 Paramètres prosodiques

Les paramètres prosodiques illustrent en grande partie le style d'élocution d'un locuteur : vitesse élocution (débit), durée et fréquence des pauses ainsi que les caractéristiques de la source glottale (fréquence fondamentale, énergie, taux de voisement,...). Néanmoins, ces paramètres caractéristiques du locuteur, notamment la fréquence fondamentale et ses variations [Atal, 1976], ne sont pas suffisamment discriminants pour être utilises seuls dans un système de RAL. Ils sont généralement associés aux paramètres de l'analyse spectrale pour améliorer les performances des systèmes de RAL.

II.2.1.3 Paramètres dynamiques

Comme nous le soulignerons dans la première partie de cette thèse, l'information dynamique véhiculée par le signal de parole est une source potentielle d'informations pour la caractérisation du locuteur, qui reste encore mal exploitée par les systèmes de RAL. Les paramètres dynamiques les plus répandus demeurent les coefficients dérivés des vecteurs de paramètres instantanés, appelés coefficients Delta (première dérivée) et Delta- Delta (seconde dérivée) [Furui, 1981], [Soong et al., 1988], [Bernasconi, 1990]. D'autres paramétrisations sont proposées dans la littérature pour exploiter les informations dynamiques du signal telles que l'utilisation des Composantes Principales Temps-Frequence (TFPC : Time Frequency Principal Components) [Magrin Chagnolleau et al., 1999], la concatenation de trames successives de signal [Hattori, 1992], [Konig et al., 1998], [Fredouille et al., 1998], [Fredouille et al., 2000a]. Un panorama et un bref descriptif de ces différentes approches sont fournis dans [Fredouille 2000].

II.2.2 Reconnaissance - Modélisation et Mesure

Le processus de reconnaissance s'appuie généralement, pour les tâches d'IAL, de VAL et de suivi de locuteurs, sur une modélisation des caractéristiques de chaque locuteur connu du système (modèles de locuteurs ou modèles clients). Cette modélisation est réalisée à partir des données d'apprentissage collectées au cours des sessions d'enrôlement. Une mesure de similarité est ensuite calculée entre un modèle client et un signal de parole, puis transmise au processus de décision. On peut distinguer quatre grandes approches pour la construction des modèles clients : les approches vectorielles, statistiques, prédictives et connexionnistes. Nous présentons ici brièvement les fondements de chacune de ces approches, les techniques qui leur sont associées ainsi que les mesures de similarité utilisées.

II.2.2.1 L'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.

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 de figure, 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. De par son principe, la programmation dynamique est utilisée exclusivement en mode dépendant du texte [Furui, 1981], [Booth et al., 1993], [Yu et al., 1995]. Très rapide et montrant des performances relativement bonnes, la programmation dynamique est toutefois très sensible à la qualité d'alignement et notamment au choix du point de départ.

Quantification vectorielle

La quantification vectorielle (Vector Quantisation : 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, et est 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 [Soong et al., 1992], [Mason et al., 1989], [Matsui et al., 1992]. 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 ; néanmoins, le processus devient d'autant plus lent.

II.2.2.2 L'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 seul modèle des locuteurs [Pruzansky, 1963]. 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 (SOSM) est de représenter une séquence de vecteurs acoustiques par une distribution gaussienne multi-dimensionnelle [Bimbot et al., 1995]. Le modèle d'un locuteur se résume alors par le triplet (x_m, X, M) où x_m est un vecteur moyenne, X est une matrice de covariance, tous deux estimes à partir de la séquence de M vecteurs acoustiques. Les SOSM sont généralement associées à des mesures de similarité particulières en vue de la reconnaissance. Ces mesures ont pour particularité de faire intervenir le triplet (y_m, Y, N). Ce dernier est estimé sur la séquence de vecteurs de test de manière analogue au triplet (x_m, X, M). Les mesures reposent ainsi essentiellement sur une ressemblance entre les matrices X et Y [Gish et al., 1986], [Bimbot et al., 1995], [Magrin Chagnolleau et al., 1996].

L'avantage majeur des SOSM est leur simplicité de mise en oeuvre. Performantes sur de courtes durées (3 secondes) [Bimbot et al., 1995], elles 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. Ces spécificités des SOSM se justifient par le fait que les mesures de ressemblance associées à ces dernières sont calculées à partir d'estimations réalisées sur l'ensemble du signal de parole, que ce soit au niveau des signaux d'apprentissage ou de test.

Mélange de gaussiennes
Un moyen de pallier ce problème (variations locales moyennées par les SOSM) est de considérer les modèles à mélanges de gaussiennes multi-dimensionnelles (Gaussian Mixture Model : GMM) [Reynolds, 1992], [Reynolds, 1995], [Reynolds et al., 2000]. 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 multi-dimensionnelles, chacune caractérisée par un vecteur moyen et une matrice de covariance. Lors de l'apprentissage, les paramètres des modèles clients (vecteur moyen $\bar{x}_i$, matrice de covariance Σ_i, pondération p_i de chaque distribution gaussienne) sont généralement estimés à l'aide de l'algorithme EM (Expectation-Maximization) [Dempster et al., 1977] couplé à l'approche par Estimation du Maximum de Vraisemblance (EMV). Lors de la reconnaissance, la mesure de similarité entre un modèle client et une séquence de vecteurs de test repose à nouveau sur l'approche 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
Empruntés à la RAP [Rabiner, 1989], les modèles de Markov cachés (Hidden Markov Models : HMM) permettent de caractériser les variations temporelles du signal de parole. Ils reposent sur une machine à états, i.e. 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 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.

II.2.2.3 L'approche connexionniste
L'approche connexionniste, telle que nous l'entendons ici, 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 clients 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 client se présente sous la forme d'un ou plusieurs réseaux de neurones pour lequel la séquence de vecteurs d'apprentissage du client concerné ainsi que celles des autres clients du système sont fournies en entrée.
Différents types de modèles de réseaux sont proposés dans la littérature ; tels que les RBF ou Radial Basis Functions [Oglesby et al., 1991], [Frederickson et al., 1994]. Lors de la reconnaissance, la vraisemblance pour qu'une séquence de vecteurs de test soit produite par un réseau de neurones est calculée.
Le principal inconvénient de l'approche connexionniste est la complexité d'apprentissage. En outre, elle pose le problème de l'ajout d'un nouveau client qui nécessite dans la majorité des cas le réapprentissage de tous les modèles. En effet, une nouvelle phase de classification est nécessaire afin de prendre en compte le nouveau client au sein du processus de discrimination entre locuteurs.

II.2.2.4 L'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. L'erreur de prédiction moyenne constitue alors la mesure de similarité entre le signal de test et le modèle de locuteur (fonction de prédiction). Une autre solution envisagée est d'estimer une fonction de prédiction sur la séquence de test et de la comparer, à l'aide d'une distance, à la fonction de prédiction estimée lors de l'apprentissage. Deux grandes techniques sont rattachées à l'approche prédictive : les modèles ARV [Grenier, 1980], [Bimbot et al., 1992], [Montacie et al., 1992], [Grifin et al., 1994], [Magrin Chagnolleau et al., 1996] et les réseaux prédictifs [Hattori, 1992], [Artieres et al., 1993], [Bennani et al., 1994], [Paoloni et al., 1996]. Ces deux techniques sont détaillées dans [Fredouille 2000].

II.3 Les tendances
Aucune grande révolution n'a pu être observée ces cinq dernières années, notamment au niveau des techniques de paramétrisation ou de modélisation. Toutefois, on peut remarquer des améliorations pertinentes des techniques actuelles ou l'émergence de nouvelles tendances.

Adaptation d'un modèle générique
La quantité de signaux d'apprentissage pour la construction des modèles clients reste une problématique majeure des systèmes de RAL. Dans cette optique, de nombreux travaux de recherche ont porté sur l'utilisation d'un modèle générique de locuteurs pour pallier le problème des données manquantes [Reynolds, 1997], [Reynolds et al., 2000]. Dans ce contexte, un modèle client est dérivé du modèle générique par adaptation des paramètres de ce dernier. Cette adaptation des paramètres est réalisée à partir des signaux d'apprentissage du client par une technique d'adaptation de type MAP (Maximum A Posteriori) [Gauvain et al., 1994] ou MLLR (Maximum Likelihood Linear Regression) [Leggetter et al., 1995].

Apprentissage incrémental
Une seconde alternative est proposée dans la littérature pour pallier l'insuffisance des signaux d'apprentissage. Cette solution, appelée apprentissage incrémental, consiste à adapter en ligne les modèles clients en utilisant des signaux de parole collectés lors de l'utilisation du système de RAL en mode non supervise [Fredouille et al., 2000b].

Téléphonie mobile
Face au développement de la téléphonie mobile à travers le monde, l'adaptation des systèmes de RAL à ce nouvel environnement devient une préoccupation omni-présente au sein de notre communauté scientifique. Néanmoins, le manque de bases de données accessibles reste, à l'heure d'aujourd'hui, le principal frein à l'ouverture d'une nouvelle voie d'investigation.

III. Approche Statistique

III.1. Introduction
Nous avons choisi d'élaborer, en premier lieu, un système de RAL basé sur les mesures statistiques de vraisemblance du type SOSM *(Second Order Statistical Measures)* [Oua,01]. Les performances de cette approche ont été clairement constatées durant les tests de reconnaissance : En plus de sa simplicité, on notera la très bonne précision d'identification atteinte avec cette approche et ceci même en milieu bruité.

III.2. Les mesures statistiques du deuxième ordre « SOSM »
III.2.1. Propriétés du modèle gaussien
Soit $\{x_t\}_{1\le t\le M}$ une suite de *M* vecteurs résultant de l'analyse acoustique de dimension p d'un signal de parole prononcé par le locuteur ***x.*** Les coefficients composant ces vecteurs sont obtenus soit par bancs de filtres, par prédiction linéaire ou par cepstre.
Sous l'hypothèse d'un modèle Gaussien du locuteur [Bim,95][Bon,97], la suite des vecteurs $\{x_t\}$ peut être résumée par son vecteur moyenne $\bar{x}$ et sa matrice de covariance *X*, tels que :

$$\bar{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \qquad (1.a) \quad \text{et} \qquad X = \frac{1}{M}\sum_{t=1}^{M} (x_t - \bar{x})(x_t - \bar{x})^T \qquad (1.b)$$

De même, pour un autre locuteur ***y***, la suite $\{y_t\}$ de *N* vecteurs peut être modélisée par $\bar{y}$ et Y, avec :

$$\bar{y} = \frac{1}{N}\sum_{t=1}^{N} y_t \qquad (2.a) \quad \text{et} \qquad Y = \frac{1}{N}\sum_{t=1}^{N} (y_t - \bar{y})(y_t - \bar{y})^T \qquad (2.b)$$

Les vecteurs moyenne $\bar{x}$ et $\bar{y}$ sont de dimension p, tandis que les covariances X et Y sont des matrices symétriques de dimension p*p.
Ainsi, le locuteur ***x*** (respectivement ***y***) sera représenté par $\bar{x}$, X et *M*, (respectivement $\bar{y}$, Y et *N*).

III.2.2. Notion de mesure de similarité
La mesure de similarité $\mu(\boldsymbol{x}, \boldsymbol{y})$ entre les locuteurs ***x*** et ***y*** peut être exprimée comme la fonction Φ suivante.

$$\mu(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\bar{x}, X, M, \bar{y}; Y, N) \qquad (3)$$

Elle est non-négative, c'est à dire :

$$\forall \boldsymbol{x},\ \forall \boldsymbol{y},\ 0 \le \mu(\boldsymbol{x}, \boldsymbol{y}) \qquad (4) \quad ; \quad \text{et elle satisfait la propriété :} \quad \forall \boldsymbol{x},\ \mu(\boldsymbol{x}, \boldsymbol{x}) = 0 \qquad (5)$$

Dans leur forme fondamentale, ces types de mesures ne sont pas symétriques, mais il y a plusieurs méthodes pour les rendre symétriques, si bien que $\forall \boldsymbol{x},\ \forall \boldsymbol{y},\ \mu(\boldsymbol{x}, \boldsymbol{y}) = \mu(\boldsymbol{y}, \boldsymbol{x})$ (6)

III.2.3. Les différentes mesures statistiques du 2ème Ordre

Les mesures statistiques les plus courantes sont les suivantes.

- La Mesure de Vraisemblance Gaussienne notée 'μ_G'.
- La Mesure de Vraisemblance Gaussienne à Covariance notée 'μ_{GC}'.
- La Mesure Arithmétique- Géométrique Sphérique notée 'μ_{SC}'.
- La Mesure de Déviation Absolue notée 'μ_{DC}'.

III.2.3.1 La mesure de vraisemblance gaussienne – Définition –

En supposant que tous les vecteurs acoustiques extraits du signal de parole prononcé par le locuteur ***x*** sont distribués selon une distribution gaussienne, la vraisemblance d'un vecteur acoustique seul y_t prononcé par le locuteur ***y*** est donnée par la fonction $G(y_t/\boldsymbol{x})$ suivante.

$$G(y_t/\boldsymbol{x}) = \frac{1}{(2\Pi)^{p/2}(\det X)^{1/2}} \times \exp\left(-\frac{1}{2}(y_t - \bar{x})^T X^{-1}(y_t - \bar{x})\right) \qquad (7)$$

Et si nous supposons que tous les vecteurs y_t sont indépendamment observables, la moyenne du log-vraisemblance de $\{y_t\}_{1\le t\le N}$ peut être décrite par :

$$\bar{G}_x(y_1^N) = \frac{1}{N}\log G(y_1 \ldots y_N/\boldsymbol{x}) = \frac{1}{N}\sum_{t=1}^{N}\log G(y_t/\boldsymbol{x}) =$$

$$= -\frac{1}{2}\left[p\log 2\Pi + \log(\det X) + \frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{x})^T X^{-1}(y_t - \bar{x})\right] \qquad (8)$$

Par ailleurs, en remplaçant $y_t - \bar{x}$ par $y_t - \bar{y} + \bar{y} - \bar{x}$ et en utilisant la propriété mathématique (9)

$$\frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{y})^T X^{-1}(y_t - \bar{y}) = tr(YX^{-1}) \qquad (9)$$

Nous aurons alors

$$\bar{G}_x(y_1^N) + \frac{p}{2}\log 2\Pi = -\frac{1}{2}\left[\log(\det X) + tr(YX^{-1}) + (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x})\right] \qquad (10)$$

de même aussi l'expression suivante :

$$\frac{2}{p}\bar{G}_x(y_1^N) + \log 2\Pi + \frac{1}{p}\log(\det Y) + 1 \text{ sera égale à}$$

$$= \frac{1}{p}\left[\log\left(\frac{\det Y}{\det X}\right) - tr(YX^{-1}) - (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x})\right] + 1 \qquad (11)$$

Donc, si nous définissons la mesure de vraisemblance gaussienne μ_G comme :

$$\mu_G(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{p}\left[tr(YX^{-1}) - \log\left(\frac{\det Y}{\det X}\right) + (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x})\right] - 1 \qquad (12)$$

$$= \frac{1}{p}\left[tr(\Gamma) - \log(\det\Gamma) + \delta^T X^{-1}\delta\right] - 1 \qquad (13)$$

$$= a - \log g + \frac{1}{p}\delta^T X^{-1}\delta - 1 \qquad (14)$$

alors nous aurons **Argmax**$_x$ $\{\bar{G}_x(y_1^N)\}$ = **Argmin**$_x$ $\{\mu_G(\boldsymbol{x}, \boldsymbol{y})\}$ (15)

Sachant que $a(\lambda_1, \lambda_2, \ldots, \lambda_p) = \frac{1}{p}\sum_{i=1}^{p}\lambda_i$ représente la moyenne arithmétique, (16)

avec $\Gamma = X^{-\frac{1}{2}} Y X^{\frac{1}{2}}$, (17)

les λ_i représentent les valeurs propres de Γ, X représente la matrice de covariance représentant ***x***, Y représente la matrice de covariance représentant ***y***,

$g(\lambda_1, \lambda_2, ..., \lambda_p) = \left(\prod_{i=1}^{p} \lambda_i \right)^{1/p}$ représente la moyenne géométrique, (18) et $\delta = \bar{y} - \bar{x}$ (19)

III.3. Apprentissage et Test

III.3.1. Phase d'apprentissage

L'apprentissage est une phase d'entraînement pour le système afin d'apprendre à distinguer les locuteurs, en ne gardant que la partie représentative de leurs paroles qui caractérise réellement la caractéristique vocalique de chacun d'eux, cette partie représentative est le vecteur moyenne et la matrice de covariance calculés à partir des coefficients MFSC. L'ensemble des vecteurs moyennes et des matrices covariances, ainsi calculés, constituera les références.

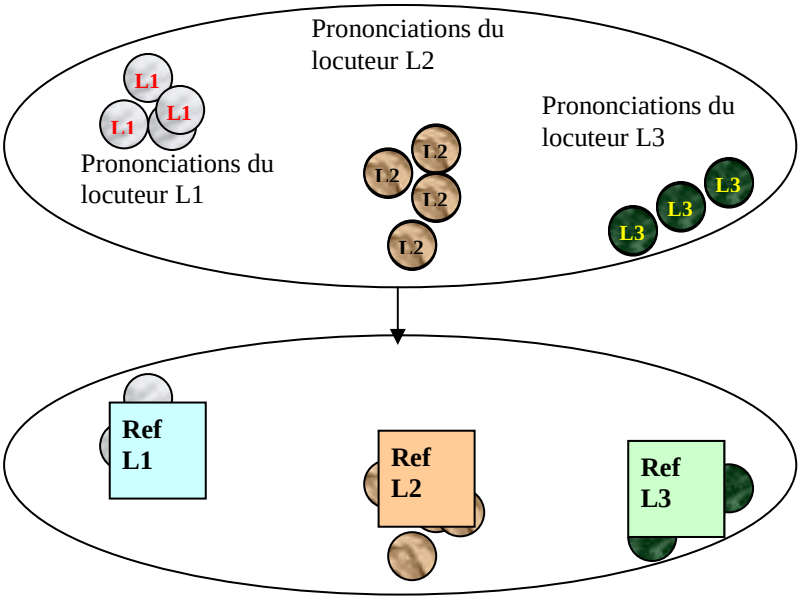

Fig. 3 Construction des références : chaque ensemble de phrases est représenté par une référence.

III.3.2 Phase de test en reconnaissance

Dans cette phase, on calcule à partir des coefficients MFSC d'un locuteur inconnu (après qu'il ait prononcé une phrase), son vecteur moyenne et sa matrice de covariance, qu'on comparera à ceux de l'ensemble des références afin d'identifier le locuteur en question.

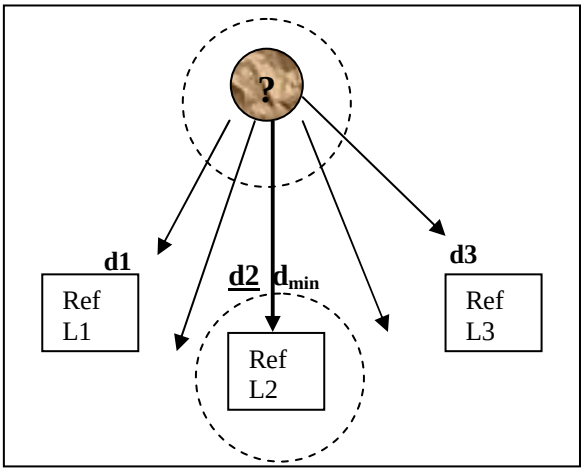

Fig. 4 Recherche de la distance minimale. Ici le locuteur inconnu est identifié comme étant la référence L2.

La règle du plus proche voisin est ici appliquée : recherche de la plus petite distance entre le locuteur inconnu et les différentes références (figure 4). La distance utilisée est la µG dans ses différentes variantes.

IV. Approche Connexionniste

IV.1 Introduction

Les réseaux de neurones ne sont pas des dispositifs biologiques mais plutôt des circuits électroniques dont chaque élément est sensé simuler le fonctionnement de la cellule élémentaire du cerveau humain qu'est le neurone.
Bien souvent en pratique, les chercheurs ne font pas appel à de véritables neurones électriques mais simulent une nouvelle fois les réseaux de neurones à l'aide d'un simple programme.

IV.2 Quelques Modèles de réseaux

IV.2.1 le perceptron

Un perceptron est un réseau de cellules composé de plusieurs modules, disposé en couches :

- **La rétine :** première couche. Elle contient les cellules d'entrée. Chaque cellule se contente de recopier la valeur qu'elle reçoit de l'extérieur sur sa sortie.
- **La couche d'association :** ou deuxième couche. Elle est composée de cellules associatives. Chaque cellule a des connexions entrantes pouvant provenir de toutes ou d'une partie des cellules de la rétine. Les fonctions de transfert de ces cellules f_i sont fixées à priori et sont en général différentes d'une cellule à une autre.
- **La couche de cellules de décision :** elle est composée d'automates à seuil. Chaque automate est connecté à toutes les sorties de la couche précédente. Les coefficients linéaires (les poids) de ces cellules sont déterminés par apprentissage. Chaque cellule de décision calcule donc sa sortie selon l'expression suivante :

  $$Y_j(X) = H\left[\sum_i W_i f_i(X)\right] \qquad (20)$$

  où $Y_j(X)$ représente la sortie de la $j^{ième}$ cellule.
  X désigne la forme présentée en entrée sur la rétine R et H désigne la fonction de seuillage. Seule cette seconde couche est donc adaptative et soumise à l'apprentissage.

Le perceptron est un réseau qui a pour tâche la classification des formes $X_1, X_2,\ldots,X_m$ présentées sur la rétine en p classes $C_1, C_2,\ldots, C_p$.

IV.2.2 Le perceptron multicouches MLP (Multi-Layer perceptron)

Le perceptron multicouche est un réseau de groupes de neurones ou couches, chaque couche est connectée à la suivante. Il comporte en général une couche d'entrée, une couche de sortie et une ou plusieurs couches dites cachées.
La fonction d'activation des neurones de ce réseau est la fonction sigmoïde. Dans ce type de réseau, le superviseur fournit à l'entrée un ensemble de couples (entrée, sortie désirée). L'information circule et se propage de l'entrée vers la sortie, et le réseau calcule sur sa couche de sortie un résultat qui doit être le plus proche possible de la sortie désirée pour toutes entrées.
L'apprentissage est réalisé avec l'algorithme de rétropropagation du gradient de l'erreur entre la sortie calculée et la sortie désirée. Le principe de cet algorithme est que, de même que l'on est capable de propager un signal provenant des cellules d'entrées vers la couche de sortie, on peut, en suivant le chemin inverse, rétropropager l'erreur commise en sortie vers les couches internes afin d'ajuster les poids synaptiques du réseau en commençant par les dernières couches jusqu'aux premières, cela pour permettre au réseau de converger vers un état qui permettra à tous les modèles d'apprentissage d'être codés [Zaa,00].

Algorithme de rétropropagation du gradient RPG

La rétropropagation du gradient est certainement l'un des plus simples et des plus efficaces algorithmes d'apprentissage pour les réseaux multicouches [Ben,92].
Mathématiquement, cet algorithme utilise simplement les règles de dérivations composées et ne présente aucune difficulté particulière. Le principe de cet algorithme est que, de même que l'on est capable de propager un signal provenant des cellules d'entrées vers la couche de sortie, on peut, en suivant le chemin inverse, rétropropager l'erreur commise en sortie vers les couches internes.
Le réseau utilisé est un réseau à couches, comportant une couche d'entrée, une couche de sortie et un certain nombre de couches dites cachées. Chaque neurone est connecté à l'ensemble des neurones de la couche suivante par des connexions dont les poids sont des nombres réels quelconques. En ce qui concerne la fonction d'activation des neurones, on utilise en général une fonction sigmoïde qui s'écrit :

$$f(x) = \frac{1}{1+e^{-x}} \qquad (21)$$

Remarque : le système neuronal artificiel ainsi constitué est appelé réseau RPG (ou BPN : Back Propagation Network).

Apprentissage :
L'apprentissage de RPG fonctionne de la façon suivante :
On dispose d'un ensemble d'exemples qui sont les couples (entrées, sorties désirées). A chaque étape un exemple est présenté en entrée du réseau. Un signal est propagé de proche en proche à travers chaque couche supérieure à la couche d'entrée jusqu'à ce qu'une sortie soit générée. Cette sortie est alors comparée à la sortie désirée et un signal d'erreur (somme quadratique des erreurs sur chaque cellule de sortie) est calculé. Ce signal d'erreur est ensuite rétro-propagé de la couche de sortie vers chaque cellule de la couche intermédiaire qui contribue directement à la sortie.
Cependant, chaque unité dans la couche intermédiaire reçoit seulement une portion de l'erreur totale basée sur la contribution relative de cette unité à la génération de sortie.
Ce processus est répété, couche par couche, jusqu'à ce que chaque cellule du réseau ait reçu le signal d'erreur.
En parallèle, les poids des connexions sont alors mis à jour pour chaque cellule, et cela pour permettre au réseau de converger vers un état qui permettra à tous les modèles d'apprentissage d'être codés. Ce processus est répété, en présentant successivement chaque exemple.
Si pour tous les exemples l'erreur est inférieure à un seuil choisi, on dit alors que le réseau a convergé.
Le sens de ce processus est que durant l'apprentissage du réseau, les cellules des couches intermédiaire s'organisent de manière à ce qu'elles apprennent à reconnaître les différentes caractéristiques de tout l'espace d'entrée [Free,92], [Dav,90], [Hér,94].

Test d'arrêt de l'apprentissage
Théoriquement, on cherche à arrêter l'apprentissage dès que le minimum de l'erreur quadratique est atteint. Ce qui correspond à un gradient nul. En pratique, ce minimum n'atteint jamais zéro. La méthode consiste à arrêter l'apprentissage dès que **Ep** est inférieure à un seuil **Emin.** Notons que **Emin** ne doit pas être trop petit (un grand nombre d'itération).

IV.2.3 La LVQ -Learning Vector Quantization-
Les algorithmes de quantification vectorielle avec apprentissage LVQ ont été proposés par T. KOHONEN. Ils constituent une adaptation de méthodes de quantification vectorielle au problème de la classification.
Ces techniques déterminent lors d'une phase d'apprentissage un ensemble de vecteurs de référence qu'elles utilisent par la suite pour classer tout nouveau point.
L'apprentissage vise donc à partitionner l'espace des formes en groupes ou « clusters », chacun étant étiqueté par une classe. En fin d'apprentissage, les groupes sont complètement définis par un vecteur de référence associé à chacun d'entre eux et une distance dans l'espace des formes.
Une fois l'apprentissage réalisé, une forme dont la classe est inconnue sera affectée au groupe dont elle est le plus proche au sens de la distance choisie. L'étiquette de ce groupe fournira alors la classe de notre forme. On reconnaît pour cette phase de décision une classification du type « plus proche voisin » sur un ensemble de base constitué des vecteurs de référence des différents groupes déterminés lors de l'apprentissage. [Ben,92]

Algorithme de la LVQ
Les algorithmes de quantification vectorielle avec apprentissage sont de méthodes développées avant tout dans un but pratique et pouvant être mises en œuvre sur des problèmes réels. Elles n'ont pas encore donné lieu à des études théoriques. Des études comparatives ont cependant montré que ces algorithmes ont de bonnes performances tout en étant rapides. Elles permettent de réaliser des modules de décision efficaces.
Il faut noter que bien que présentés ici sous un formalisme réseau, ces techniques sont plus proches du domaine de la reconnaissance des formes classique que de celui des réseaux de neurones. Toutefois nous les présentons ici car elles ont été conçues comme un cas particulier et une variation de l'algorithme des cartes topologiques [Koh,88]. De plus elles peuvent être intégrées dans des architectures de réseaux complexes à plusieurs modules. Il est donc intéressant de les présenter dans ce cadre.

Présentation
Les algorithmes LVQ peuvent être considérés comme une adaptation d'algorithmes du type k-moyennes ou quantification vectorielle, qui sont non supervisés, au cas de la classification.

Ces techniques déterminent lors d'une phase d'apprentissage un ensemble de vecteurs de référence qu'elles utilisent par la suite pour classer tout nouveau point. Celui-ci se voit attribuer la classe du vecteur de référence le plus proche pour une distance donnée: le critère de décision est donc un critère de plus proche voisin.
Dans un formalisme connexionniste, cela revient à considérer des architectures à deux couches: la couche d'entrée sur laquelle on présente les formes et celle de sortie qui donne la classe d'affectation de l'exemple. Ces deux couches sont totalement connectées. Les vecteurs de poids des cellules de décision seront les coordonnées des vecteurs de référence cités plus haut. Le problème peut alors se formuler comme celui de l'apprentissage des poids des cellules de décision. Sur présentation d'une forme, le réseau calcule la distance de cette forme à chaque vecteur de poids.
L'apprentissage vise donc à partitionner l'espace des formes en groupes ou "clusters", chacun étant étiqueté par une classe. En général plusieurs groupes seront ainsi associés à une même classe. Dans les versions proposées par **KOHONEN** [Koh,84], le nombre de ces groupes est prédéterminé, il constitue un des paramètres de l'algorithme. En fin d'apprentissage, les groupes sont complètement définis par un vecteur de référence associé à chacun d'entre eux et une distance dans l'espace des formes.
Une fois l'apprentissage réalisé, une forme dont la classe est inconnue sera affectée au groupe dont elle est le plus proche au sens de la distance choisie. L'étiquette de ce groupe fournira alors la classe de notre forme. On reconnaît pour cette phase de décision une classification du type plus proche voisin sur un ensemble de base constitué des vecteurs de référence des différents groupes déterminés lors de l'apprentissage.

IV.2.3.1. La LVQ 1

La règle d'adaptation pour LVQl est la suivante. On présente successivement les vecteurs de l'ensemble d'apprentissage qui sont donc étiquetés. Parmi tous les vecteurs de référence, on sélectionne le plus proche de l'exemple présenté. Si ce vecteur de référence appartient à la même classe que l'exemple, on le rapproche de ce dernier d'une quantité proportionnelle à la différence entre les deux vecteurs. Dans le cas contraire, l'exemple est mal classé et on écarte la référence de la même quantité.
On voit que tous les vecteurs sont mis à contribution pour modifier les vecteurs de référence. La règle de modification ressemble à celle du perceptron avec la différence qu'il s'agit d'une règle de type *récompense/punition.* Elle va rapprocher le vecteur de référence sélectionné w* de la forme présentée x s'ils sont dans la même classe et l'éloigner sinon.
Pour la reconnaissance, on présente un vecteur de classe inconnue et on lui affecte la classe du vecteur de référence le plus proche.

IV.2.3.2. La LVQ2

Dans la version LVQ2, pendant l'apprentissage, on va réduire le nombre de cas où l'adaptation est effectuée. Les erreurs de classification se produisent naturellement aux frontières entre les classes: LVQ2 effectue les modifications des vecteurs de référence uniquement dans ces régions. LVQ2 semble avoir des performances légèrement supérieures à celles de LVQl et nécessiter moins de vecteurs de référence par classe, ce qui est très important en pratique, les calculs de distance étant très longs.
Plus précisément on n'adapte les vecteurs de référence que dans le cas où les conditions suivantes sont réunies:

- l'exemple est mal classé: la classe de la référence la plus proche est différente de la classe de l'exemple.
- la référence suivante la plus proche est de la même classe que l'exemple.
- l'exemple tombe dans une fenêtre définie symétriquement autour du milieu de ces deux vecteurs de référence.

L'algorithme consiste alors à éloigner la référence de la "mauvaise" classe et de rapprocher celle de la "bonne" classe dans le cas où l'exemple tombe à l'intérieur de la fenêtre définie précédemment. La reconnaissance est identique à celle de LVQl

Algorithmes

Nous donnons tout d'abord une version générale de l'algorithme, commune aux deux variantes.

- Données: couples $(x,y) \in R^n$, y indice de classe
- But: classification des vecteurs X par modification des valeurs des connexions
- Paramètres: le nombre de vecteurs de référence, le pas de modification, la taille de la fenêtre dans LVQ2.
- Initia1isation :

  - Déterminer un nombre donné de vecteurs de référence initiaux à l'aide d'un algorithme non supervisé. On peut par exemple utiliser un algorithme de type k-moyennes.

  - Étiqueter le groupe associé à chaque vecteur de référence par la classe majoritaire dans le groupe. On fournit ainsi une étiquette pour le vecteur de référence.

- Présentation aléatoire des formes de l'ensemble d'apprentissage.

1) A la date *t*, présenter un vecteur exemple *x*, dont la classe sera notée C.
2) déterminer le vecteur de référence le plus proche *w*(t)* et sa classe *C**.
3) Modification éventuelle de *w*(t)* (voir plus bas).
4) Faire *t=t+1*.

Critère d'arrêt : nombre d'itérations fixé à priori ou taux de classification satisfaisant.

Les deux algorithmes diffèrent uniquement par l'étape 3.

- **Dans LVQ l elle devient**

Modifier le vecteur de référence *w*(t)* par:
$w^*(t+1) = w^*(t) + \alpha\,(t)\,(x - w^*(t))$ si C = C* (22)
(la référence et l'exemple sont de même classe)
$w^*(t+1) = w^*(t) - \alpha\,(t)\,(x - w^*(t))$ si C ≠ C* (23)
(la référence et l'exemple sont de classes différentes)
Dans tous les autres cas, faire:
$w^k(t+1) = w^k(t)$ (24)

$\alpha(t)$ est un nombre suffisamment petit que l'on fait décroître au cours du temps, par exemple on peut utiliser une fonction linéaire décroissante dans le temps comme:
$\alpha(t) = 0.1\,(1 - t/M)$ (25)
et *M* est le nombre maximum d'itérations après lequel l'apprentissage est arrêté.

- **Et dans L VQ2 :**

Si *x* et *w*(t)* sont de classes différentes *(C≠ C*)*, déterminer le vecteur de référence le plus proche suivant, *w**(t)*, et sa classe *C***

Si X et w**(t) sont de même classe (C = C**), déterminer une fenêtre symétrique autour du milieu de [w*(t), w**(t)] de largeur L.
Si l'exemple x est dans cette fenêtre, déplacer les vecteurs de référence w*(t) et w**(t) :
$w^*(t+1) = w^*(t) - \alpha(t)\,(x - w^*(t))$ (26)
$w^{**}(t+1) = w^{**}(t) + \alpha(t)\,(x - w^{**}(t))$ (27)
Dans tous les autres cas, faire:
$w^k(t+1) = w^k(t)$ (28)

x: vecteur de la base d'apprentissage. w et w**: vecteurs de référence. di(x) et dj(x) : distributions des classes Ci et Cj*

V. Identification du Locuteur par la Méthode SOSM

Dans cette section, nous décrivons les expériences faites en identification du locuteur par la méthode SOSM, nous exposons les différents résultats obtenus et enfin nous essayerons de discuter les résultats d'un point de vue objectif. Ici, les tests de reconnaissance sont faits en milieu sain et en milieu bruité.
Par ailleurs, la méthode SOSM sera étudiée selon plusieurs points de vue : la résolution spectrale variant de 12 canaux à 60 canaux, le RSB variant de 0 dB à son maximum correspondant à un environnement non bruité et enfin la bande passante du signal de parole pouvant être soit une 'bande téléphonique' de 3 kHz soit une 'bande large' de 8 kHz (c'est-à-dire la bande d'origine de la base de données). Tous ces paramètres sont pris comme facteurs variables de l'environnement d'identification.
Les nombreux résultats obtenus, illustrés sous forme de graphiques, montrent que cette méthode se comporte différemment selon le cas étudié. Ceux-ci permettent de suivre le comportement et la robustesse de la méthode d'identification en fonction de ces trois paramètres de variabilité.
En plus des constats qualitatifs de la SOSM vis-à-vis des types de bruit utilisés (bruit blanc, bruit de foule et bruit de voiture), nous avons observé l'importance de la haute résolution spectrale 60 canaux / 8 kHz pour la bande spectrale 0-8kHz et celle de 48 canaux / 8kHz sur la bande réduite, alors que les recherches actuelles favorisaient toujours des bancs de 24 canaux seulement. A titre d'exemple, on peut citer une amélioration d'identification de 11%, sur la bande téléphonique, dès qu'on passe de 24 canaux à 48 canaux. Enfin nous proposons des solutions diverses pour améliorer la qualité de reconnaissance du locuteur qui se résument à la fin de cette section.

V.1. Cas non bruité

L'étude suivante consiste à identifier 37 locuteurs de la base de données parlée TIMIT [Fis,86] par la SOSM, dans un environnement non bruité. Deux cas sont prévus : L'identification sur la bande [0-8 kHz] et l'identification sur la bande téléphonique. Les dimensions des MFSC varient, alors de 12 à 60 coefficients.

Deux constats intéressants sont à noter :

- **Pour la bande 0-8000Hz,** le taux de fausses identifications est de 5.4% avec 12 MFSC et il est de 0% pour tous les autres cas (24, 36, 48 et 60 MFSC) ce qui signifie une bonne identification à partir de 24 canaux.
- **Pour la bande téléphonique,** le taux de fausses identifications est de 56.7% avec 12 MFSC, il est de 13.5% avec 24 MFSC, il est de 8.1% avec 36 MFSC, il est de 2.7% avec 48 MFSC et il est de 8.1% avec 60 MFSC. Par conséquent la dimension de 48 canaux s'avère être la meilleure dimension sur la bande téléphonique.

V.2. Cas bruité

La deuxième étude consiste à identifier ces 37 locuteurs en environnement bruité. Les bruits utilisés sont décrits ci-dessous, avec une pondération variable allant de 0 dB à 18 dB. Deux cas sont prévus: l'identification sur la bande [0-8 kHz] et l'identification sur la bande téléphonique. Les dimensions des MFSC varient, alors, de 12 à 60.
Nous avons utilisé trois types de bruits :

- le bruit blanc, gaussien, centré et stationnaire au deuxième ordre noté BBG [Cou,87].
- le bruit de chahut ou de foule synthétisé à partir d'une sommation de 7 signaux de dialogue enregistrés à partir de différentes chaînes de télévision.
- le bruit de voiture qui consiste en la somme de 3 bruits réels de voiture enregistrés par un microphone de haute qualité dans une route à moyenne circulation.

Le bruitage est pondéré par le RSB choisi, allant de 0 dB jusqu'à 18 dB. Notons que le signal original non bruité sera symbolisé par un RSB de 24 dB pour simplifier la représentation graphique des courbes de résultat.
Les résultats de cette deuxième étude (environnement bruité) sont exposés sur les figures 5 à 7. Nous pouvons alors noter les points suivants :

- Sur la bande audible 0-8 kHz, les meilleurs taux sont obtenus avec 60 canaux, ce qui signifie que la haute résolution spectrale renforce le système d'identification contre les bruits.
- Sur la bande téléphonique, les meilleurs taux sont obtenus tantôt avec 48 canaux et tantôt avec 60 canaux. Nous voyons une oscillation de la courbe de 48 canaux autour de celle de 60 canaux ; ce qui s'explique par un taux optimal intermédiaire.
- Sur 0-8 kHz, le bruit le plus gênant est le chahut suivi du BBG et enfin le bruit de voiture qui parait non gênant, relativement aux deux précédents.
- Sur la bande téléphonique le bruit le plus gênant est le chahut suivi des deux autres types.
- Sur la bande téléphonique on remarque la paradoxale diminution du taux de reconnaissance au moment où le RSB évolue de 18 à 24 dB (sans bruit) dans le cas de la courbe à 12 canaux.
- Sur la bande 0-8 kHz, un bruit très fort à 0 dB de RSB dévalue le système de reconnaissance de plus de 20%, sauf pour le cas du bruit de voiture où le taux persiste à 97.3% même à 0 dB, en considérant 60 canaux toujours.
- Sur la bande réduite, le BBG et le bruit de voiture à 0 dB dévaluent le système de reconnaissance de plus de 20%. Pour ce qui est du bruit de chahut la dévaluation dépasse les 40%, ce qui implique une défaillance du système d'identification en présence du chahut.

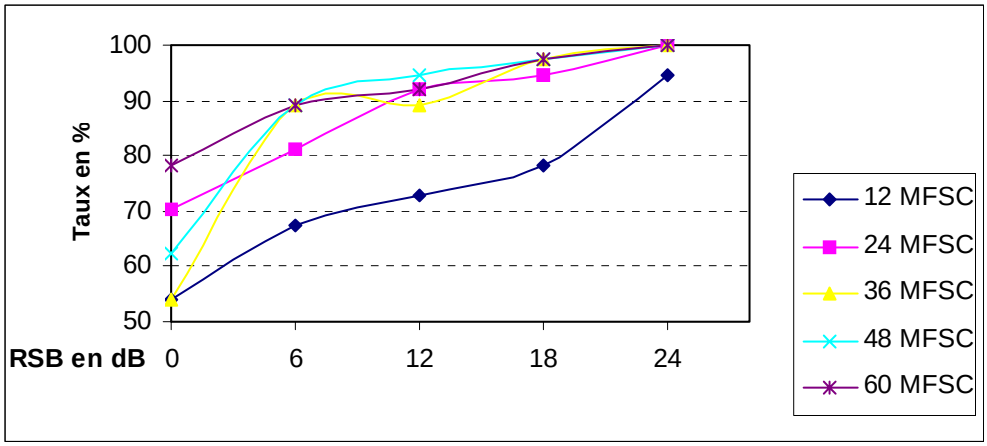

Fig. 5 Taux de reconnaissance en environnement bruité par le BBG, sur la bande 0-8 kHz

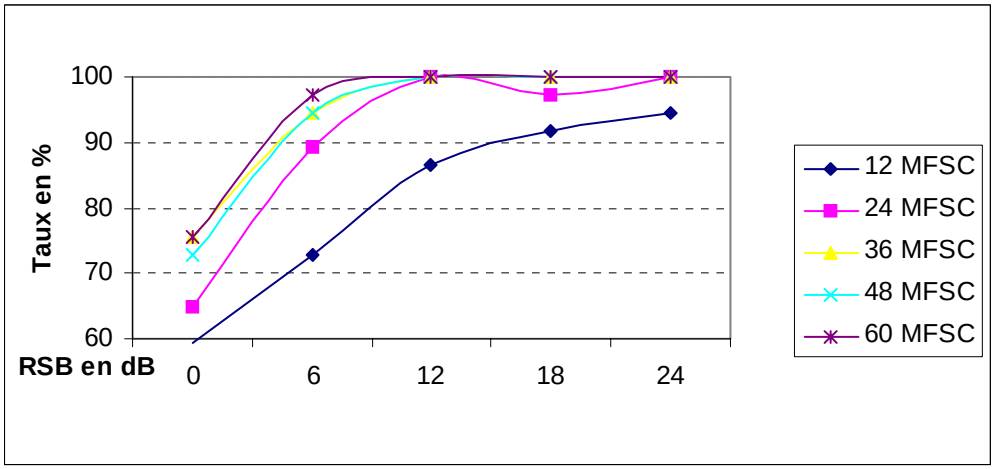

Fig. 6 Taux de reconnaissance en environnement bruité par le chahut, sur la bande 0-8 kHz

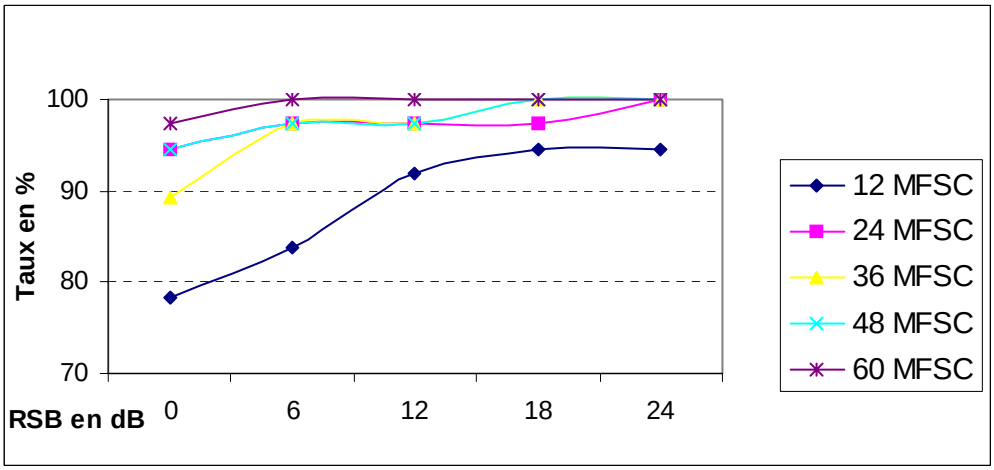

Fig. 7 Taux de reconnaissance en environnement bruité par le bruit de voiture, sur la bande 0-8 kHz

V.3. Synthèse générale

Dans cette partie du travail, nous avons développé un système statistique pour l'identification automatique du locuteur, basé sur l'algorithme SOSM ou Second Order Statistical Measures.

La méthode SOSM est simple à mettre en œuvre, rapide en exécution et très précise en identification.

Les résultats en milieu sain ont dévoilé un taux de bonne reconnaissance de 100% sur la bande normale (0-8 kHz) et de 97.3% sur la bande téléphonique (300-3400 Hz). Ce qui représente un résultat excellent sur une population de 37 locuteurs.

Cependant, en milieu bruité cette méthode présente quelques problèmes ; voilà pourquoi nous avons essayé d'améliorer ses performances en jouant sur la résolution spectrale (taille des MFSC).

Durant cette étude nous avons pu montrer l'importance de la haute résolution spectrale en identification, en trouvant le nombre optimal de coefficients spectraux à adopter dans le cas de la méthode SOSM.

Ainsi, les tests faits en milieu bruité (BBG, chahut et bruit de voiture) ont montré que la haute résolution spectrale apporte une grande quantité d'informations propre au locuteur et aide à mieux le reconnaître en milieu bruité. Ainsi la dimension « 60 coefficients / 8 kHz » apparaît comme une dimension très favorable aux systèmes de reconnaissance du locuteur qui sont utilisés en milieu bruité. Cependant, sur la bande téléphonique l'optimal serait « 48 coefficients / 8 kHz » en milieu non bruité et ce serait un compromis entre ces deux dernières dimensions (48 et 60) quand la parole est bruitée. Nous pensons que cette légère réduction est due à la limitation majorée par 3400 Hz qui sous-entend une perte importante des informations contenues dans les hautes fréquences.

Les résultats obtenus pour le bruit de voiture prouvent que le bruit de voiture n'est pas gênant en reconnaissance du locuteur, contrairement à ce qu'on pouvait penser vu la gêne auditive générée par ce dernier. Par conséquent, un système de vérification de locuteur peut aisément être implanté dans des endroits à proximité des grandes routes (station service, autoroutes, stations de bus etc.). Alors que le chahut, si bien filtré par notre système nerveux, s'avère être extrêmement gênant pour identifier un locuteur. Probablement, cette défaillance est due au fait que le chahut est

de même nature que le signal de parole et ainsi donc les caractéristiques spectrales subissent une forte altération au niveau des coefficients les plus pertinents pour le locuteur. Il s'en suit qu'avant toute procédure de vérification de locuteur, il faut s'assurer que la zone d'enregistrement microphonique ne soit pas une zone à chahut ou une zone riche en bruit de foule (marché, bureau de poste, gare etc.).

Enfin, nous tenons à faire remarquer que les résultats de cette partie de notre travail sont typiques à une seule méthode de reconnaissance du locuteur, en l'occurrence : la SOSM, et nous ne pouvons étendre ces résultats pour d'autre méthodes utilisées dans les mêmes fins, sans à priori refaire les tests d'identifications décrits dans ce chapitre.

VI. Identification du Locuteur par la LVQ - Nouvelle Métrique Proposée -

Nous utilisons ici la méthode LVQ sous différentes formes (LVQ1 et MLVQ1) et avec différents types de caractéristiques acoustiques.
Cependant, souvent dans cette méthode nous nous heurtons au problème d'incompatibilité des paramètres, dans le cas où ces paramètres (caractéristiques acoustiques) ne sont pas homogènes, tel que l'association des MFSC avec Fo, par exemple.
Pour contourner ce problème, nous avons proposé une nouvelle métrique associative (que l'on a appelé ODHEF) dans le cas des paramètres acoustiques hétérogènes. Celle-ci nous a permis d'associer des paramètre hétérogènes dans le calcul d'une distance interlocuteur. Ainsi des caractéristiques très significatifs (tels que Fo) ont pu être associés aux paramètres spectraux (ou cepstraux) classiques, en rehaussant le taux de reconnaissance.

Nous décrivons ci-dessous le principe de la distance ODHEF proposée.

VI.1. Principe de la distance ODHEF (cas de 3 paramètres différents)

Nous avons élaboré une nouvelle métrique, adaptée à notre application d'identification de locuteur, quand les caractéristiques acoustiques sont de natures différentes (on considère 3 paramètres seulement pour simplifier).

Description :
Dans les conditions de travail propres à la reconnaissance du locuteur, il nous a paru judicieux de définir la distance entre deux vecteurs X et Xr (à 3 dimensions), de la manière suivante :

$$Dist^2(X, Xr) = \sum_{i=1}^{3} (Fiab_i)(Norm_i)(Xr_i - X_i)^2 \qquad (29)$$

X étant un vecteurs représentant une prononciation et Xr représente une référence par exemple. Les composantes, de caractéristiques différentes, du vecteur X sont X1, X2 et X3 (de même Xr1, Xr2, Xr3 pour le vecteur Xr).

- Le 1er terme, après la somme, est le coefficient de fiabilité de Wolf, dans le but de privilégier les paramètres les plus pertinents. Car l'association d'un paramètre pertinent avec un autre paramètre de faible fiabilité, dans une métrique uniforme, ne fera que diminuer la fiabilité du 1er paramètre.

- Le 2ème terme est un facteur de normalisation, par la moyenne, pour ramener tous les paramètres à un même ordre de grandeur. Par exemple si un paramètre est de l'ordre de 10^5 il sera mal associé à un paramètre de l'ordre de 10^1, si on ne fait pas de normalisation pour ramener les 2 paramètres au même ordre de grandeur.

- Le 3ème terme est la distance euclidienne classique.

L'expression générale de cette nouvelle distance est, alors:

$$Dist^2(X, Xr) = \sum_{i=1}^{3} \left(\frac{\sigma^2_{Inter}(X_i)}{\sigma^2_{Intra}(X_i)}\right)\left(\frac{1}{E(E(X_i))}\right)^2 (Xr_i - X_i)^2 \qquad (30)$$

Ce qui peut se simplifier en:

$$Dist^2(X, Xr) = \sum_{i=1}^{3} \left[\left(\frac{\sigma_{Inter}(X_i)}{\sigma_{Intra}(X_i)}\right)\left(\frac{1}{E(E(X_i))}\right)(Xr_i - X_i)\right]^2 \qquad (31)$$

ou mieux encore:

$$\sum_{i=1}^{3} \left[\left(\frac{\sigma_{Inter}(X_i)}{\sigma_{Intra}(X_i)}\right)\left(\frac{1}{E(E(X_i))}\right)Xr_i - \left(\frac{\sigma_{Inter}(X_i)}{\sigma_{Intra}(X_i)}\right)\left(\frac{1}{E(E(X_i))}\right)X_i\right]^2 \qquad (32)$$

soit: $$Dist^2(X, Xr) = \sum_{i=1}^{3} [Pond(i)Xr_i - Pond(i)X_i]^2 \qquad (33)$$

où Pond(i) représente le coefficient de pondération normalisé du paramètre "i".
En introduisant la variable pondérée Y, l'expression précédente devient:

$$Dist^2(X, Xr) = \sum_{i=1}^{3} (Yr_i - Y_i)^2 \qquad (34)$$

avec Yi=Pond(i).Xi., X=(X1,X2,X3)^t et Xr=(Xr1,Xr2,Xr3)^t.

Donc on peut utiliser la métrique L^2 classique à condition de transformer les vecteurs X en des vecteurs adaptés Y par une transformation de pondération. Nous avons appelé cette métrique : **ODHEF** (Optimized Distance for HEterogeneous Features)**.**

VI.2. Méthodes de reconnaissance du locuteur à base de LVQ1 et MLVQ1
Nous avons proposé 3 méthodes différents, toutes basées sur la LVQ, pour tenter d'identifier chacun des locuteurs du test.
VI.2.1. Méthode Prosodique à base de LVQ1 (nouvelle méthode)
Nous avons réalisé un système d'identification de locuteur purement prosodique, basé sur Fo, la durée et l'énergie.
VI.2.1.1. Caractéristiques
Cette méthode est basée sur la valeur moyenne du pitch (Fo_{moy}), la durée originale (D_{orig}) et l'énergie basse fréquence (E_{bf}), dans cette méthode chaque phrase est représentée par un vecteur de 3 dimensions (Fo_{moy} , D_{orig} , E_{bf}). On peut voir sur la figure 8 la représentation de tous les locuteurs avec ces 3 paramètres.

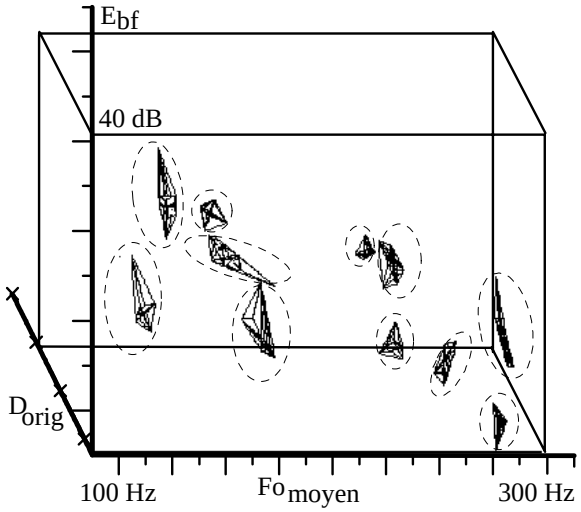

Fig. 8 Représentation des locuteurs avec les 3 dimensions prosodiques Fo_{moy}, D_{orig} et E_{bf}.

On peut encore rajouter à ces 3 paramètres un paramètre dit de nasalité estimé par la différence entre la sortie du 8ème canal et du 13ème canal du banc de filtres. Celui-ci représente, généralement, la pente du 1er anti-formant.

VI.2.1.2. Algorithme
On utilise dans nos expériences la LVQ1 et la MLVQ1 (Modified Learning Vector Quantization 1). Le temps d'apprentissage est de 30 secondes et la distance utilisée est la distance euclidienne multivariable (voir métrique).
Résultat des expériences : On a obtenu un taux de reconnaissance du locuteur de 100% sur un ensemble fermé de 11 locuteurs et sur des phrases constantes (identification dépendante du texte). En enlevant le paramètre Fo_{moy} ce taux devient 66.7% seulement. Voir tableau 1.

Tableau 1 Taux de Reconnaissance pour la méthode prosodique.

Methode	Taux de Rec. %	
	Ensemble fermé (11 loc.)	Ensemble ouvert (15 loc.)
LVQ1 avec Fo	98	93.3
MLVQ1 avec Fo	100	93.3
MLVQ1 sans Fo	66.7	62.2
Amélioration due à Fo	+33.3	+31.1

VI.2.2 Méthode Cepstrale à base de LVQ1

Nous utilisons les caractéristiques cepstrales du signal de parole issu des locuteurs (12 caractéristiques MFCC sont utilisées). A partir de ces caractéristiques MFCC, on extrait le vecteur moyenne $MFCC_{moy}$ et la matrice covariance $MFCC_{cov}$. A partir de la matrice de covariance on tire le 1er vecteur propre et nous gardons, alors, seulement les caractéristiques suivantes: **$MFCC_{moy}$,** et **$MFCC_{vp1}$.**

Ainsi, chaque locuteur sera représenté par le couple (**$MFCC_{moy}$, $MFCC_{vp1}$**).

Phase d'apprentissage:

D'abord, on enregistre une phrase issue d'un locuteur de référence, puis on détermine les vecteurs caractéristiques sur chaque fenêtre de 25 ms. Finalement, on calcule les 2 vecteurs **$MFCC_{moy}$** et **$MFCC_{vp1}$** à partir des MFCC trouvés. Ces 2 vecteurs représentent, alors, une référence. Durant l'apprentissage, des exemples seront comparés au dictionnaire de référence et selon la classe trouvée, par la méthode du plus proche voisin, une correction LVQ1 sera affectée à la position des éléments de référence dans le dictionnaire : rapprochement ou éloignement.

Phase de reconnaissance :

Le locuteur à reconnaître prononce une phrase différente, on en extrait les 2 vecteurs caractéristiques de la même manière que précédemment et on le classe dans la classe de la référence la plus proche :

Si dist(X,Xref) =min Alors X⇔Xref.

Le calcul de distance se fait comme dans le cas des paramètres prosodiques (Distance ODHEF).

Dans notre cas nous avons utilisé cette distance selon la formule ci-dessous:

$$Dist_{MEMORES} = \sqrt{\alpha_1 d^2{}_1 + \alpha_2 d^2{}_2} \qquad (35)$$

$d^2{}_1$=distance relative à $MFCC_{moy}$ et $d^2{}_{2=}$ distance relative à $MFCC_{vp1}$, les coefficients α_i valent $\alpha_1=\alpha 2 \approx 0.5$.

Résultats

Les tests effectués sur un ensemble de 11 locuteurs ont donné un taux de reconnaissance dépendant du texte de 79% de bonne reconnaissance.

La base de données expérimentale, ici, est constituée de 153 fichiers représentant les 11 locuteurs, chaque fichier est constitué du vecteur moyenne et du 1er vecteur propre. Voir tableau 2.

Tableau 2 Taux de Reconnaissance par la méthode cepstrale.

Méthode	**Taux de bonne Reconnaissance**
LVQ1	79%
MLVQ1	79%

VI.2.3 Méthode Spectrale à base de LVQ1

Nous utilisons les caractéristiques spectrales du signal de parole issu des locuteurs (24 caractéristiques MFSC sont utilisées). A partir de ces caractéristiques MFSC, on extrait le vecteur moyenne $MFSC_{moy}$.

Nous préservons alors seulement le vecteur moyenne: **$MFSC_{moy}$** de dimension 24x1**.**

Ainsi, chaque locuteur sera représenté par ce vecteur (**$MFSC_{moy}$**).

Phase d'apprentissage:

D'abord, on enregistre une phrase issue d'un locuteur de référence, puis on détermine les vecteurs caractéristiques sur chaque fenêtre de 25 ms. Finalement, on calcule le vecteur **$MFSC_{moy}$** à partir des MFSC trouvés. Ce vecteur représente, alors, une référence. Durant l'apprentissage, des exemples seront comparés au dictionnaire de référence et selon la classe trouvée, par la méthode du plus proche voisin, une correction LVQ1 sera affectée à la position des éléments de référence dans le dictionnaire : rapprochement ou éloignement. Voir « Algorithme de la LVQ1 ».

Phase de reconnaissance:

Le locuteur à reconnaître prononce une phrase différente, on en extrait le vecteur caractéristique de la même manière que précédemment et on le classe dans la classe de la référence la plus proche :

Si **dist(X,Xref) = min alors X⇔Xref**.

Résultats

Les tests effectués sur la base de donnée décrite précédemment (ensemble de 11 locuteurs référencés et 4 locuteurs étrangers) ont donné un taux de reconnaissance proche de 100%. Voir tableau 3.

Tableau 3 Taux de Reconnaissance par la méthode spectrale.

Méthode	**Taux de Reconnaissance %**		
	Dépendant du texte	Indépendant du texte	Ensemble ouvert (15 loc.)
LVQ1	99	100	100

VI.3. Conclusion

Nous pouvons voir sur les tableaux précédents les différents résultats obtenus par l'approche LVQ1 et ceci avec différents types de caractéristiques acoustiques.
La LVQ1 donne globalement des résultats satisfaisants proche de 100% sans nécessiter des unités connexionnistes complexes. Néanmoins nous avons remarqué que le bon fonctionnement de ce type de réseaux de neurones nécessite un très grand nombre d'exemples lors de l'apprentissage ; ce qui est un peu gênant pour le confort des locuteurs.

En résumé :

- la méthode prosodique semble donner des résultats précis avec un taux de l'ordre de 100% de bonne reconnaissance.
- la méthode cepstrale a manqué de précision en donnant un taux de reconnaissance proche de 80% ce qui est faible relativement aux autres résultats concurrents. Peut-être que le fait d'avoir choisi un seul vecteur propre seulement est insuffisant.
- la méthode spectrale est assez précise : taux de reconnaissance proche de 100%, mais le fait de ne choisir que le vecteur moyenne reste insuffisant en pratique.

Dans le cas des deux premières méthodes nous avons utilisé des caractéristiques acoustiques de nature hétérogène ; cela a nécessité de concevoir une nouvelle métrique que l'on a appelé ODHEF. La théorie correspondante a prouvé le bon choix de cette métrique dans de pareilles situations. De plus les expériences faites, pour chaque paramètre pris séparément, ont montré que cette métrique associative était très utile.

VII. Vérification automatique du Locuteur par MLP mono-locuteur

VII.1. Introduction

L'objectif de cette partie est la réalisation d'un système de vérification automatique du locuteur en utilisant des réseaux de neurones mono-locuteur à base de MLP.
Nous présentons dans cette section la procédure choisie pour réaliser notre objectif, ainsi que les différentes expériences que nous avons menées dans le but de trouver les paramètres adéquats pour le meilleur fonctionnement de notre système de vérification.
On a utilisé un perceptron multicouche (MLP) dont la structure est composée d'une couche d'entrée, une couche de sortie, et une couche cachée. L'apprentissage est supervisé et est basé sur la règle de rétropropagation du gradient. Le réseau calcule sur sa couche de sortie un résultat qui doit être mis en correspondance avec le vecteur d'entrée en corrigeant la structure du réseau d'une manière rétrograde.
La technique d'apprentissage et de test est résumée dans le schéma de la figure 9.
L'apprentissage est effectué selon les paramètres suivants : Pas d'apprentissage : Pa = 0.2. Erreur désirée: Err = 10^{-3}. Nombres d'itération: Iter = 100000. Seuil d'arrondissement: Sa = 0.5.

Le seuil d'arrondissement intervient lors de l'affichage de la sortie calculée par le réseau, en effet, les éléments de cette sortie sont réels, et ce seuil va arrondir ces éléments soit à 1 (si sortie $\geq$ Sa) ou bien à 0 (sinon).

Pour faire l'apprentissage d'un seul réseau on a utilisé 10 phrases, les 5 premières sont celles du locuteur concerné et les 5 autres sont différentes et choisies d'une manière aléatoire. Ainsi, en tout, nous aurons 370 phrases utilisées pour l'apprentissage
On fait rentrer, alors, ces phrases l'une après l'autre au réseau. L'apprentissage devra s'arrêter dès que le minimum de l'erreur quadratique Ep est atteint c.à.d dès que Ep est inférieur à l'erreur voulue Err.

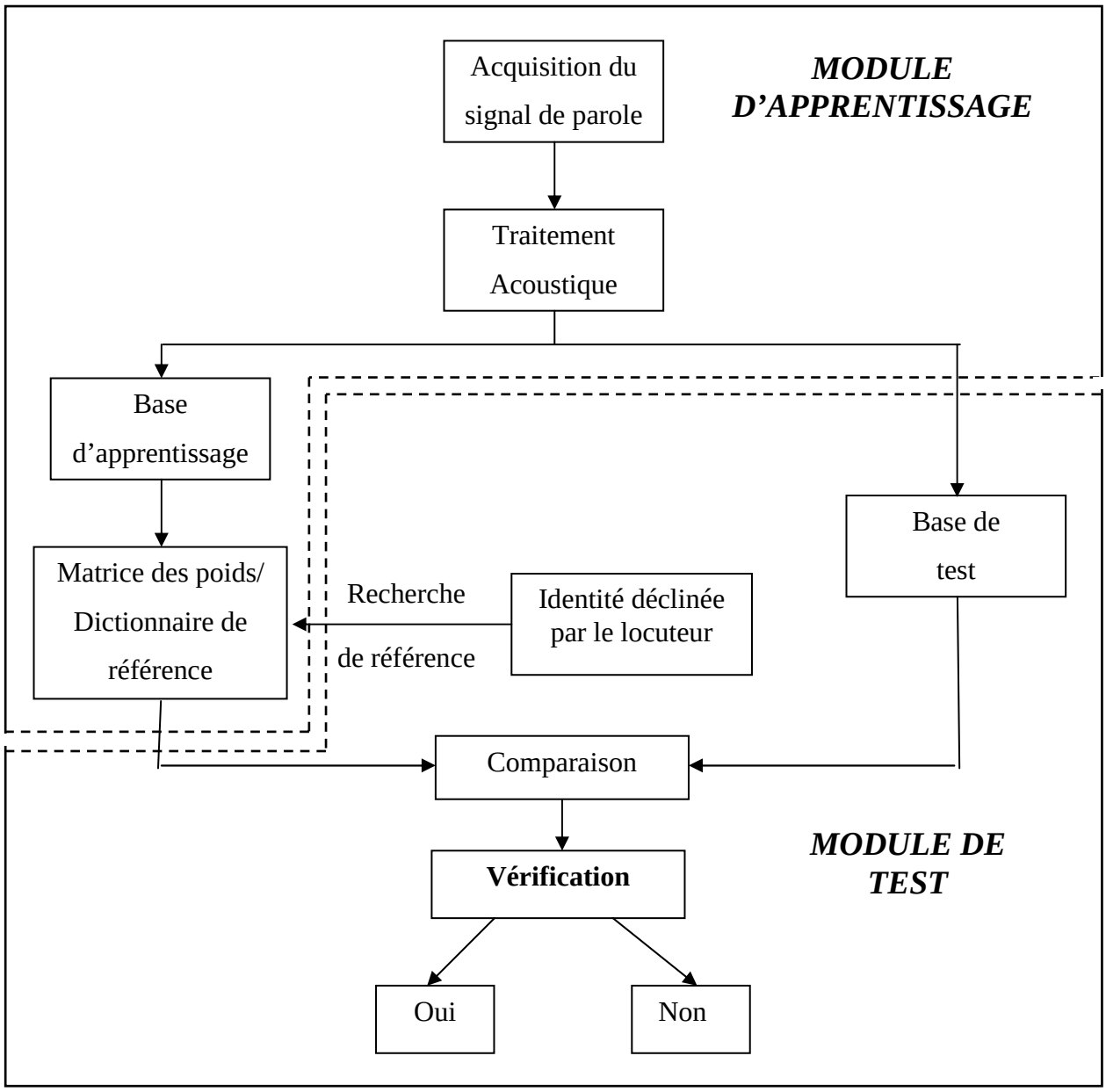

Figure 9 Synoptique général de la technique d'apprentissage et de test en VAL

VII.2 Expériences avec deux vecteurs propres

Le tableau ci dessous résume les taux de bonne acceptation, les taux de faux rejet et les taux de fausse acceptation obtenus en utilisant les deux premiers vecteurs propres de la covariance et en variant le nombre de neurones cachés de 20 à 35 neurones.

Tableau 4 Taux de vérification obtenus en utilisant les deux premiers vecteurs propres

	20 neurones	25 neurones	30 neurones	35 neurones
Taux de bonne acceptation	83.78 %	80.54 %	77.84 %	79.46 %
Taux de faux rejet	16.22 %	19.46 %	22.16 %	20.54 %
Taux moyen de fausse acceptation	28.33 %	26.36 %	25.44 %	24.79 %

En comparant les résultats obtenus dans le tableau 4, on voit que le taux de bonne acceptation varie entre 77,84% et 83,78% dont le meilleur taux est obtenu en utilisant 20 neurones cachés, alors que le taux moyen de fausse acceptation varie entre 24,79% et 28,33% dont la meilleure moyenne est obtenue en utilisant 35 neurones cachés.

Le choix du nombre de neurones cachés pour ce cas est difficile à prendre car le meilleur taux a été obtenu tantôt en utilisant 20 neurones cachés et tantôt en utilisant 35 neurones cachés. En comparant les résultats obtenus en utilisant 20 et 35 neurones cachés et en se basant sur le fait qu'en vérification automatique du locuteur il semble logique de pénaliser l'erreur de fausse acceptation plutôt que l'erreur du faux rejet, on peut dire que les meilleurs résultats ont été obtenus en utilisant 35 neurones cachés.

VII.3 Expériences avec la diagonale de la covariance

Le tableau ci dessous résume les taux de bonne acceptation, les taux de faux rejet et les taux de fausse acceptation obtenus en utilisant la diagonale de la covariance et en variant le nombre de neurones cachés de 20 à 35 neurones.

Tableau 5 Taux de vérification obtenus en utilisant la diagonale de la covariance

	20 neurones	25 neurones	30 neurones	35 neurones
Taux de bonne acceptation	78.92 %	78.92 %	80.54 %	81.08%
Taux de faux rejet	21.08 %	21.08%	19.46 %	18.92 %
Taux moyen de fausse acceptation	26.27 %	27.81 %	27.99 %	25.54 %

En comparant les résultats obtenus dans le tableau 5, on voit que le taux de bonne acceptation varie entre 78,92% et 81,08% dont le meilleur taux est obtenu en utilisant 35 neurones cachés, alors que le taux moyen de fausse acceptation varie entre 25,54% et 27,81% dont la meilleure moyenne est obtenue en utilisant 35 neurones cachés.
On peut conclure que pour ce cas - c'est à dire qu'en utilisant la diagonale de la covariance comme paramètre d'entrée du réseau- le meilleur choix du nombre de neurones cachés est de 35 neurones

VII.4 Expériences avec le vecteur moyenne

Le tableau ci dessous résume les taux de bonne acceptation, les taux de faux rejet et les taux de fausse acceptation obtenus en utilisant le vecteur moyenne et en variant le nombre de neurones cachés de 20 à 35 neurones.

Tableau 6 Taux de vérification obtenus en utilisant le vecteur moyenne

	20 neurones	25 neurones	30 neurones	35 neurones
Taux de bonne acceptation	88.65 %	90.27 %	90.27 %	98.38 %
Taux de faux rejet	11.35 %	9.73 %	9.73 %	1.62 %
Taux moyen de fausse acceptation	25.91 %	19.60 %	17.14%	11.70 %

En comparant les résultats obtenus dans le tableau 6, on voit que le taux de bonne acceptation varie entre 88.65% et 98.38% dont le meilleur taux est obtenu en utilisant 35 neurones cachés, alors que le taux moyen de fausse acceptation varie entre 11.70% et 25.91% dont la meilleure moyenne est obtenue en utilisant 35 neurones cachés.
On peut conclure que pour ce cas - c'est à dire en utilisant le vecteur moyenne comme paramètre d'entrée du réseau- les meilleurs résultats sont obtenus en utilisant 35 neurones.

VII.5 Discussion générale

Après avoir fait varier les paramètres d'entrées des réseaux (les deux premiers vecteurs propres, la diagonale de la covariance et le vecteur moyenne) ainsi que le nombre de neurones cachés on a pu aboutir aux conclusions suivantes :

- **Pour les deux premiers vecteurs propres :** On a obtenu un taux de bonne acceptation égal à 79,46% et un taux moyen de fausse acceptation égal à 24,79%, en utilisant 35 neurones cachés.
- **Pour la diagonale de la covariance :** On a obtenu un taux de bonne acceptation égal à 81,08% et un taux moyen de fausse acceptation égal à 25,54%, en utilisant 35 neurones cachés.
- **Pour la moyenne :** On a obtenu un taux de bonne acceptation égal à 98,38% et un taux moyen de fausse acceptation égal à 11,70%, en utilisant 35 neurones cachés.

D'après ces résultats on voit que les meilleurs résultats sont obtenus en utilisant le vecteur moyenne, ce qui implique que la moyenne est plus précise que les autres paramètres. Ceci est peut-être dû aux informations pertinentes contenues dans le vecteur moyenne et qui permettent de mieux caractériser le locuteur. Toutefois il est insuffisant de se baser uniquement sur ce dernier paramètre, mais l'idéal serait d'associer ces 3 paramètres en même temps à l'entrée du MLP.
Dans notre étude, il ressort que la configuration utilisant le vecteur moyenne permet d'obtenir un taux d'erreur inférieur à 2% et un taux moyen de fausse acceptation inférieur à 12%. En arrivant à ce résultat on peut dire que le système n'est pas bien sécurisé car la probabilité d'intrusion par des imposteurs est grande (de l'ordre de 12%). Pour cela on a pensé à améliorer encore ce résultat en jouant sur la couche de sortie. De nouvelles expériences de VAL ont été faites en faisant varier le seuil d'arrondissement, tout en observant l'évolution de la précision du réseau. Les résultats obtenus sont meilleurs ; nous donnons brièvement les nouveaux scores : TBA=98.38 et TFAmoy= 6.02.

VIII. Conclusion Générale

Dans le cadre de cette thèse, nous avons abordé un domaine de recherche non encore maîtrisé : La Reconnaissance Automatique du Locuteur. Nous avons ainsi tenté de concevoir plusieurs types de méthodes émanant de deux visions différentes : *une vision statistique et une vision connexionniste.*
Les méthodes élaborées sont les suivantes : une méthode basée sur la SOSM pour l'Identification du locuteur, trois méthode basée sur la LVQ pour l'Identification et une méthode basée sur le MLP pour la Vérification du locuteur

- La méthode Statistique, basée sur les mesures de similarité du 2nd ordre, nécessite peu d'apprentissage. Cette méthode testée, pour la reconnaissance automatique de 37 locuteurs différents, a donné d'excellents résultats (taux de 100%). De plus cette méthode est évolutive et reste très simple à mettre en œuvre.
- Les méthodes Connexionnistes basées sur la LVQ ont été testées pour la reconnaissance de 11 locuteurs différents. La complexité de ces méthodes est moyenne. Les résultats de reconnaissance sont bons sur la base de données testée. Mais les résultats pouvant être obtenus sur des bases de données plus grandes devraient être moins appréciables.
- La méthode Connexionniste basée sur le MLP a été testée pour la vérification de 37 locuteurs différents. La méthode est assez complexe et son temps d'apprentissage est relativement élevé. Les résultats de vérification sont bons sur la base de données testée. Ces résultats, néanmoins, restent insuffisants pour des applications de vérification industrielles qui exigent des taux d'erreur très proches de 0%.

Les performances des différentes techniques proposées sont données dans le tableau 7 : les 3 premières méthodes ont été testées durant le travail de recherche accompli au cours de cette thèse, tandis que les autres méthodes (GMM et HMM) sont des méthodes récemment testées en RAL par d'autres chercheurs du domaine.

Tableau 7 Performances de certaines méthodes en RAL

Méthode	**Précision**	**Complexité**	**Rapidité en Reconnaiss.**	**Rapidité en Apprentissage**	**Evolutivité**	**Dépendance du Texte**
SOSM*	Très Précise	Simple	Très Rapide	Très Rapide	Evolutive	Indépendante du Texte
LVQ*	Bonne Précision	Moyenne	Très Rapide	Moyenne	Non Evo.	Dépendante du Texte
MLP*	Bonne Précision	Complexe	Très Rapide	Lente	Non Evo.	Indépendante du Texte
HMM	Moyenne	Complexe	Rapide	Moyenne	_	Dépendante du Texte
GMM	Extrêmement Précise	Complexe	Rapide	Moyenne	Evolutive	Indépendante du Texte

**** N.B.** les 3 premières méthodes sont décrites selon les conditions d'expérimentation faites durant cette thèse.*

Ainsi, nous voyons que les méthodes les plus précises sont les méthodes statistiques SOSM et GMM. Notons que les GMM représentent actuellement une référence de pointe en RAL.
Par ailleurs, nous voyons que les méthodes connexionnistes sont moins précises : LVQ et MLP.
Du point de vue complexité et temps d'apprentissage, la SOSM demeure actuellement la meilleure méthode pouvant être utilisée en RAL.
Du point de vue indépendance du texte toutes les méthodes le sont sauf la LVQ et les HMM.
Quant à l'évolutivité, nous remarquons que les méthodes connexionnistes ne sont pas évolutives (i.e. dès qu'un nouveau locuteur de référence est ajouté à l'ensemble des adhérents, le système n'est plus fonctionnel alors). Les méthodes statistiques, par contre, sont évolutives, d'où un autre avantage de ces techniques.

Nous faisons rappeler, par ailleurs, que durant cette étude, nous avons pu trouver la résolution spectrale optimale à adopter en IAL ainsi que les différents bruits à éviter avant toute tentative de reconnaissance. Nous avons aussi élaboré une nouvelle métrique bien adaptée aux caractéristiques acoustiques hétérogènes et qui a permis d'utiliser simultanément plusieurs paramètres différents d'une manière cohérente.

Finalement quelle méthode choisir ?
En conclusion, cette étude montre que les réseaux de neurones ne sont pas très intéressants en RAL sauf si une association avec des méthodes statistiques est envisagée. Certaines expériences, dans ce sens, ont montré que cela est possible et qu'ainsi le taux de reconnaissance est bien amélioré.

Toutefois, l'amélioration apportée en précision est tellement faible (pour une complexité plus grande) que l'on se demande s'il est vraiment utile d'ajouter toute cette complexité aux systèmes statistiques !
Les travaux actuels tendent à se pencher vers les mélanges de gaussiennes ou GMM qui ont bien fait leur preuve en IAL et en VAL.

Références bibliographiques

[Art, 93] Artieres T., Gallinari P. Neural models for extracting speaker characteristics in speech modelization systems. European Conference on Speech Communication and Technology (Eurospeech), pages 2263{2266, 1993, Berlin (Allemagne).

[Ata, 76] Atal B. S. Automatic recognition of speakers from their voices. IEEE transactions, volume 64(4), pages 460{475, 1976.

[Auc, 00] Auckenthaler R., Carey M., Lloyd-Thomas H. Score normalization for text-independent speaker verification system. Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop, 10(1-3), 2000.

[Ban, 00] Banziger T., Klasmeyer G., Johnstone T., Kamceva T., Scherer K. R. Améliorer les systèmes de vérification automatique du locuteur en intégrant la variabilité émotionnelle : Méthodes et premiéres données. XXIIIémes Journées d'Etudes sur la Parole (JEP), pages 341{344, 2000, Aussois (France).

[Ben, 92] Younès Bennani. « Approches Connexionnistes Pour La Reconnaissance Automatique Du Locuteur : Modelisation Et Identification ». Thèse de Doctorat de l'université de Paris Sud.1992.

[Ben, 94] Bennani Y., Gallinari P. Connexionist approaches for automatic speaker recognition. Workshop on Automatic Speaker Recognition, Identification, Verication, pages 95{102, Avril 1994, Martigny (Suisse).

[Ben, 90] Bennani Y., Soulie F. F., Gallinari P. A connectionist approach for automatic speaker identification. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 265{268, 1990.

[Ber, 90] Bernasconi C. On instantaneous and transitional spectral information for text-dependent speaker verification. Speech Communication, volume 9(2), pages 129{139, 1990.

[Bes, 00b] Besacier L., Grassi S., Dufaux A., Ansorge M., Pellandini F. GSM speech coding and speaker recognition. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000b, Istanbul (Turquie).

[Bim, 98] Bimbot F., Hutter H.-P., Jaboulet C., Koolwaaij J., Lindberg J., Pierrot J.-B. An overview of the CAVE project research activities in speaker verification. Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pages 215{220, Avril 1998, Avignon (France).

[Bim, 95] Bimbot F., Magrin Chagnolleau I., Mathan L. Second-order statistical measures for text-independent speaker identification. Speech Communication, volume 17(1-2), pages 177{192, Août 1995.

[Bim, 92] Bimbot F., Mathan L., De Lima A., Chollet G. Standard ant target driven AR-Vector Models for speech analysis and speaker recognition. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 5{8, 1992, San Francisco (USA)..Bibliographie 177

[Boe, 98] Boe L. J. L'identification juridique de la voix : le cas fran¢ cais - historique, problématiques et propositions. Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pages 222{239, Avril 1998, Avignon (France).

[Boe, 99] Boe L. J., Bimbot F., Bonastre J.-F., Dupont P. De l'évaluation des systémes de vérification du locuteur _ a la mise en cause des expertises vocales en identification juridique. Revue Langues, volume 2(4), Décembre 1999.

[Bon, 97] Bonastre, J.F., Besacier, L., 1997. Traitement Indépendant de Sous-bandes Fréquentielles par des méthodes Statistiques du Second Ordre pour la Reconnaissance du Locuteur. Actes du 4ème Congrès Français d'Acoustique, pp 357-360, Marseille 14-18 April 1997.

[Bon, 00a] Bonastre J.-F., Delacourt P., Fredouille C., Meignier S., Merlin T., Wellekens C. J. Difiérentes stratégies pour le suivi du locuteur. Reconnaissance des Formes et Intelligence Artificielle (RFIA), pages 123{129, 2000a, Paris (France).

[Bon, 00b] Bonastre J.-F., Delacourt P., Fredouille C., Merlin T., Wellekens C. J. A speaker tracking system based on speaker turn detection for NIST evaluations. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000b, Istanbul (Turquie).

[Boo, 93] Booth I., Barlow M., Watson B. Enhancements to DTW and VQ decision algorithms for speaker recognition. Speech Communication, volume 13(3-4), pages 427{433, Décembre 1993.

[Bov, 98] Boves L. Commercial applications of speaker verification : overview and critical success factors. Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pages 150{159, Avril 1998, Avignon (France).

[Cha, 98] Champod C., Meuwly D., Weintraub M., Sonmez K. The inference of identity in forensic speaker recognition. Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pages 125{134, Avril 1998, Avignon (France).

[Cha, 97] Charlet D. Authentification vocale par téléphone en mode dépendant du texte. Thése de doctorat, Ecole Nationale Supérieure des Télécommunications (ENST), 1997, Paris (France).

[Cou, 87] F. de COULON 1987, "Théorie et Téorie du Signal, Ed. Presses Polytechniques Romandes, Lausanne 1987.

[Cre, 94] M.j Creaney and R.N Gorgui-Naguib. « A Scaly Artificiel Neural Network For Speaker Independent ». PROCEEDINGS OF THE 1994 IEEE WORKSHOP. Department of electrical and Electronic Engineering. University of Newcastle Upon Tyne. Newcastle-Upon-Tyne NEI 7 RU U.K. page 44331.

[Dau, 83] B.A. DAUTRICH, L.R. RABINER and T.B. MARTIN 1983, "The effects of selected signal processing techniques on the performance of a filter bank based isolated word recognizer", Bell System Technical Journal, 1983.

[Dav, 90] Eric Davalo, Patrick Naim. « Des Reseaux De Neurones ». Eyrolles 1990. [Dev, 94] De Veth J., Bourlard H. Comparison of hidden Markov model techniques for automatic speaker verification. Workshop on Automatic Speaker Recognition, Identification, Verification, pages 11{14, Avril 1994, Martigny (Suisse).

[Del, 00] Delacourt P. La segmentation et le regroupement par locuteurs pour l'indexation de documents audio. Thése de doctorat, Institut Eurecom, 2000, Nice (France)..178 Bibliographie

[Dem, 77] Dempster A. P., Laird N. M., Rubin D. B. Maximum-likelihood from incomplete data via the EM algorithm. Journal of Acoustical Society of America (JASA), volume 39, pages 1{38, 1977.

[Dod, 85] Doddington G. R. Speaker recognition. Identifying people by their voices. IEEE transactions, volume 73(11), pages 1651{1664, 1985.

[Dod, 98] Doddington G. R. Speaker recognition evaluation methodology { An overview and perspective {. Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pages 60{66, Avril 1998, Avignon (France).

[Eag, 95] Eagles. Assessment of speaker verification systems. EAGLES Spoken Language Systems, 1995.

[Fis, 86] W. FISHER, V. ZUE , J. BERNSTEIN and D. PALLET 1986, “An acoustic-phontic database”, JASA, suppl. A, Vol. 81(S92) 1986.

[Fis, 99] Fissore L., Ravera F., Vair C. Speech recognition over GSM : specific features and performance evaluation. Workshop on robust methods for speech recognition in adverse conditions, pages 127{130, Mai 1999, Tampere (Finlande).

[Fre, 92] James A. Freeman & David M. Skapura. « Neural Networks ». Algorithms, Applications and programming techniques. 1992. [Fre, 94] Frederickson S. E., Tarassenko L. Radial basis functions for speaker identification. Workshop on Automatic Speaker Recognition, Identification, Verification, pages 107{110, Avril 1994, Martigny (Suisse).

[Fre, 98] Fredouille C., Bonastre J.-F. Use of dynamic information with second order statistical methods in speaker identification. Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pages 50{54, Avril 1998, Avignon (France).

[Fre, 00a] Fredouille C., Bonastre J.-F., Merlin T. AMIRAL : a block segmental multirecognizer architecture for automatic speaker recognition. Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop, 10(1-3), 2000a..Bibliographie 179

[Fre, 00b] Fredouille C., Mariéthoz J., Jaboulet C., Hennebert J., Bonastre J.-F., Mokbel C., Bimbot F. Behavior of a Bayesian adaptation method for incremental enrollment in speaker verification. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000b, Istanbul (Turquie).

[Fur, 77] Furui S. An analysis of long-term variation of feature parameters of speech and its application to talker recognition. Electron. Communication, volume 57-A, pages 34{42, 1977.

[Fur, 81] Furui S. Cepstral analysis technique for automatic speaker verification. IEEE Transactions Acoustics, Speech, and Signal Processing (ASSP), volume 29(2), pages 254{272, Avril 1981.

[Fur, 94] Furui S. An overview of speaker recognition technology. Workshop on Automatic Speaker Recognition, Identification, Verification, pages 1{9, Avril 1994, Martigny (Suisse).

[Fur, 95] Furui S. An overview of speaker recognition technology. Automatic speech and speaker recognition - Advanced topics, 1995.

[Gau, 94] Gauvain J. L., Lee C. H. Maximum a Posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Transactions on Speech and Audio Processing, volume 2(2), pages 291{298, Avril 1994.

[Gis, 86] Gish H., Krasner M., Russel W., Wolf J. Methods and experiments for text-independent speaker recognition over telephone channels. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 865{868, 1986, Tokyo (Japan).

[Gre, 77] Grenier Y. Identification du locuteur et adaptation au locuteur d'un systéme de reconnaissance phonétique. Thése de doctorat, Ecole Nationale Supérieure des Télécommunications (ENST), 1977, Paris (France)..180 Bibliographie

[Gre, 80] Grenier Y. Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur. IXémes Journées d'Etudes sur la Parole (JEP), pages 163{171, 1980, Strasbourg (France).

[Gri, 94] Grifin C., Matsui T., Furui S. Distance measures for text-independent speaker recognition based on MAR model. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 309{312, 1994, Adélaide (Australie).

[Hat, 92] Hattori H. Text-independent speaker recognition using neural networks. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 153{156, 1992, San Francisco (USA).

[Her, 94] A.Hervé. « Les Reseaux De Neurones ». presses universitaire de Grenoble 1994.

[Hér, 94] Hérault J, Jutten C. « Reseaux Neuronaux Et Traitement Du Signal ». Editions Hérmés, Paris, 1994.

[Hol, 90] Hollien H. The acoustics of crime. Applied psycholinguistics and communication disorders, 1990.

[Hom, 95] Homayounpour M. M. Vérification vocale d'identité : dépendante et indépendante du texte. Thése de doctorat, Université de Paris-Sud centre d'Orsay, 1995, Paris (France).

[Hom, 94] Homayounpour M. M., Chollet G. Performance comparison of some relevant spectral representations for speaker verification. Workshop on Automatic Speaker Recognition, Identification, Verification, pages 27{30, Avril 1994, Martigny (Suisse).

[Jac, 00] Jacob B., Mariéthoz J., Gravier G., Bimbot F. Robustesse de la vérification du locuteur par un mot de passe personnalisé. XXIIIémes Journées d'Etudes sur la Parole (JEP), pages 357{360, 2000, Aussois (France).

[Joh, 99] Johnson S. E. Who spoke when ? - automatic segmentation and clustering for determining speaker turns. European Conference on Speech Communication and Technology (Eurospeech), Septembre 1999, Budapest (Hongrie)..Bibliographie 181

[Kar, 98] Karlsson I., Banziger T., Dankovicov_a J., Johnstone T., Lindberg J., Melin H., Nolan F., Scherer K. Speaker verification with elicited speaking-styles in the Verivox project. Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pages 207{210, Avril 1998, Avignon (France).

[Kha, 00] Kharroubi J., Chollet G. Utilisation de mots de passe personnalisés pour la vérification du locuteur. XXIIIémes Journées d'Etudes sur la Parole (JEP), pages 331{334, 2000, Aussois (France).

[Koh, 84] Kohonen, T.: Self-Organization and Associative Memory. Springer Series in Information Sciences 8, Springer Verlag, New York, 1984.

[Koh, 88] Kohonen, T.: The "Neural" Phonetic Typewriter, In: Computer, 1988, S. 11-22.

[Kon, 98] Konig Y., Heck L. P., Weintraub M., Sonmez K. Nonlinear discriminant feature extraction for robust text-independent speaker recognition. Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pages 72{75, Avril 1998, Avignon (France).

[Kun, 94] Kunzel H. J. Current approaches to forensic speaker recognition. Workshop on Automatic Speaker Recognition, Identification, Verification, pages 135{141, Avril 1994, Martigny (Suisse).

[Lee, 95] H.S. LEE and A.C. TSOI 1995, “ Application of mujti-layer perceptron in estimating speech / noise characteristics for speech recognition in noisy environment ” . Speech Communication, Volume. 17, Number, 1-2, August 1995, pp. 59-76.

[Leg, 95] Leggetter C. J., Woodland P. C. Maximum Likelihhod Linear Regression for speaker adaptation of continuous Hidden Markov Models. Computer Speech and Language, volume 9, pages 171{185, 1995.

[Lin, 97] Lindberg J., Melin H. Text-prompted versus sound prompted passwords in speaker verification system. European Conference on Speech Communication and Technology (Eurospeech), pages 22{25, Septembre 1997, Rh^odes (Gr^ece).

[Mag, 99] Magrin Chagnolleau I., Durou G. Time-frequency principal components of speech : application to speaker identification. European Conference on Speech Communication and Technology (Eurospeech), pages 759{762, Septembre 1999, Budapest (Hongrie).

[Mag, 96] Magrin Chagnolleau I., Wilke J., Bimbot F. Further investigation on AR-vector models for text-independent speaker identification. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 401{404, 1996, Atlanta (USA).

[Mar, 00] Martin A., Przybocki M. The NIST 1999 speaker recognition evaluation - an overview. Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop, 10(1-3), 2000.

[Mar, 97] Martin A. F., Przybocki M. A. The DET curve in assessment of detection task performance. European Conference on Speech Communication and Technology (Eurospeech), pages 1895{1898, Septembre 1997, Rh^odes (Gréce).

[Mas, 89] Mason J. S., Oglesby J., Xu L. Codebooks to optimise speaker recognition. European Conference on Speech Communication and Technology (Eurospeech), pages 267{270, 1989, Paris (France).

[Mas, 92] J.Master. « Pratical Neural Network Recipes ». Acadimic press 1992.

[Mat, 92] Matsui T., Furui S. Comparison of text-independent speaker recognition methods using VQ-distorsion and discrete-continuous HMMs. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 157{160, 1992, San Francisco (USA).

[Mat, 94b] Matsui T., Furui S. Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 125{128, 1994b, Adélaide (Australie).

[Mei, 00] Meignier S., Bonastre J.-F., Fredouille C., Merlin T. Evolutive HMM for speaker tracking system. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000, Istanbul (Turquie).

[Mon, 92] Montacié C., Deléglise P., Bimbot F., Caraty M.-J. Cinematic techniques for speech processing : temporal decomposition and multivariate linear prediction. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 153{156, 1992, San Francisco (USA)..Bibliographie 183

[Nai, 94] Naik J. Speaker verification over the telephone : databases, algorithms and performance assessment. Workshop on Automatic Speaker Recognition, Identification, Verification, pages 31{38, Avril 1994, Martigny (Suisse).

[Ogl, 95] Oglesby J. What's in a number ? : moving beyond the Equal Error Rate. Speech Communication, volume 17(1-2), pages 193{209, Ao^ ut 1995.

[Ogl, 90] Oglesby J., Mason J. S. Optimisation of neural models for speaker identification. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 261{264, 1990.

[Ogl, 91] Oglesby J., Mason J. S. Radial basis function networks for speaker recognition. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 393{396, 1991, Toronto (Canada).

[O'Sh, 86] O'Shaughnessy D. Speaker recognition. IEEE Transactions Acoustics, Speech, and Signal Processing (ASSP), pages 4{17, Octobre 1986.

[Oua, 01] Ouamour, S., Sayoud, H., Boudraa, M., 2001. Suivi de Locuteur par la statistique d'ordre 2. RJC'2001, pp 130-133, Mons 11-14 sept. 2001.

[Pao, 96] Paoloni A., Ragazzini S., Ravaioli G. Predictive neural networks in text independent speaker verification : an evaluation on the SIVA database. International Conference on Spoken Language Processing (ICSLP), pages 2423{2426, 1996, Philadelphia (USA).

[Pie, 98] Pierrot J.-B. Elaboration et validation d'approches en vérification du locuteur. Thése de doctorat, Ecole Nationale Supérieure des Télécommunications (ENST), 1998, Paris (France).

[Pru, 63] Pruzansky S. Pattern matching procedure of automatic talker recognition. Journal of Acoustical Society of America (JASA), volume 35, pages 354{358, 1963.

[Prz, 99] Przybocki M. A., Martin A. F. Two-channel telephone data for speaker detection and speaker tracking. European Conference on Speech Communication and Technology (Eurospeech), Septembre 1999, Budapest (Hongrie).

[Qua, 00] Quatieri T. F., Singer E., Dunn R. B., Reynolds D. A., Campbell J. P. Speaker and language recognition using speech codec parameters. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000, Istanbul (Turquie)..184 Bibliographie

[Rab, 89] Rabiner L. R. A tutorial on Hidden Markov Models and selected applications in speech recognition. IEEE transactions Speech Audio Processing, volume 77(2), pages 257{285, 1989.

[Rey, 92] Reynolds D. A. A Gaussian mixture modeling approach to text-independent speaker identification. Thése de doctorat, Georgia Institute of Technology, 1992, (USA).

[Rey, 94] D.A. REYNOLDS 1994, "Speaker identification and verification using Gaussian Mixture speaker models", Workshop on Automatic Speaker Recognition, identification and verification", April 1994, Martigny, Switzerland, pp. 27-30.

[Rey, 94'’] Reynolds D. A. Experimental evaluation of features for robust speaker identification. IEEE transactions Speech Audio Processing, volume 2, pages 639{643, 1994.

[Rey, 95] Reynolds D. A. Speaker identification and verification using gaussian mixture speaker models. Speech Communication, volume 17(1-2), pages 91{108, 1995.

[Rey, 96] Reynolds D. A. The eécts of handset variability on speaker recognition performance : experiments on the Switchboard corpus. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1996, Atlanta (USA).

[Rey, 97] Reynolds D. A. Comparison of background normalization methods for text-independent speaker verification. European Conference on Speech Communication and Technology (Eurospeech), Septembre 1997, Rh^odes (Gréce).

[Rey, 00] Reynolds D. A., Quatieri T. F., Dunn R. B. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop, 10(1-3), 2000.

[Ros, 76] Rosenberg A. E. Automatic speaker verification, a review. Proceedings IEEE, volume 64(4), pages 475{487, 1976.

[Ros, 98a] Rosenberg A. E., Magrin-Chagnolleau I., Parthasarathy S., Huang Q. Speaker detection in broadcast speech databases. International Conference on Spoken Language Processing (ICSLP), pages 1339{1342, 1998a, Sydney (Australia).

[Ros, 91] Rosenberg A. E., Soong F. K. Recent research in automatic speaker recognition. Advances in speech signal processing, 1991.

[Say, 94] H. Sayoud, et al. « Etude comparative de plusieurs détecteurs de F0 », ppIV-32à 36 , ICSS'1994 Alger 24-26 sept 1994.

[Say, 94'] H. Sayoud, et al. « PDA à Ambiguïté Modifiée », ppIV-37 à 40 ICSS'1994, Alger 24-26 sept 1994.

[Say, 95] H. Sayoud, et al. « Détecteurs de Pitch à convergence de fréquence » MCEA'1995, Grenoble septembre 1995.

[Say, 96] H. Sayoud, et al. « Interprétation des erreurs de PDA »

[Say, 97] H. Sayoud, et al. CFA'1997, Marseille 14-18 avr. 1997. « L'erreur spectrale 3D », pp 377-380, JTEA'1996, Tunis 8-9 nov. 1996.

[Say, 98] H. Sayoud, et al. “Error Correction Algorithmses for PDAs”, CESA'98 Tunisia April 1-4 1998, pp 337-341, Volume 4, 1998.

[Say, 98’] H. Sayoud, et al. “ Méthodes comparatives en reconnaisance du locuteur ”, JTEA’98 Nabeul, Tunis, Tunisie 6 et 7 Nov. 1998.

[Say, 98’’] H. Sayoud, et al. « Post-Processing for PDA’s », SSST’1998, Virginia 8-10 Mars 1998.

[Say, 98”’] H. Sayoud, et al. On the use of statistical ratio for classification in automatic speaker recognition, MCEA, Marrakech 17-19 sept 1998.

[Say, 98””] H. Sayoud, et al. Inter and Intra-speaker variability of some phonetic parameters in standard Arabic. Classification in speaker recognition. CESA’98, 1-4 April 1998, Nabeul, Tunisia pp 216-219.

[Say, 98””’] H. Sayoud, et al. Tests comparatifs de differentes methodes en identification du locuteur, NWSIP’98, 2 dec 1998, Sidi-bel- abbas, Algerie.

[Say, 99] H. Sayoud, et al. « Reconnaissance automatique du locuteur sur la bande téléphonique et microphonique », CSCA’1999, Alger 18-19 oct. 1999.

[Say, 99’] H. Sayoud, et al. Tentatives en reconnaissance du locuteur, SSA2’99, Blida 10-12 Mai 99, Algerie.

[Say, 00] H. Sayoud, et al. Reconnaissance Automatique du Locuteur en Milieu Bruité. Juin 2000, pp. 345-348. www.icp.inpg.fr/jep2000/. JEP’2000, Aussois France.

[Say, 01] H. Sayoud, et al. Discrimination Parole / Non Parole en suivi de locuteur. 8-11 Septembre 2001, p-to-p : 130-132 (in RJC’2001). RJC’2001 Mons Belgique.

http://tcts.fpms.ac.be/rjc/communication.html.

[Say, 01’] H. Sayoud, et al. Suivi de locuteur par la statistique d’ordre 2. 8-11 Septembre 2001, p-to-p: 134-135 (in RJC’2001). RJC’2001 Mons Belgique.

http://tcts.fpms.ac.be/rjc/communication.html.

[Say, 02] H. Sayoud, et al. Classification des Sons par rapport à la Parole Pure basée sur la statistique d’ordre 2. CFA’2002, 8-13 avril 2002, pp 771 à 774. Lille France.

www.isen.fr/cfa2002/cadres_cfa.htm.

[Say, 02’] H. Sayoud, et al. Indexation des Documents Audio en vue d’un Filtrage par Locuteur. 24-27 juin 2002, pp 141-148. TALN’2002, Nancy France. www.loria.fr/projets/TALN/.

[Say, 02”] H. Sayoud, et al. Automatic Speaker Indexing in Corrupted Speech. 8-12 September 2002, Brésil, pp 877-882. ITS’2002, Natal Brésil. http://www.its2002.ufrn.br/.

[Say, 02”’] H. Sayoud, et al. A Robust Method for Speaker Tracking. 27-29 Août 2002, Egypte, pp 187-192. ICCTA’2002, Alexandria Egypt. http://iccta.aast.edu/home.html.

[Say, 02””] H. Sayoud, et al. Speaker Indexing in a noisy environm. Investigation of 3 types of noise ICAMSL’02, Dec. 19-21, 2002, Tenerife Espagne. www.wseas.org/conferences/2002/tenerife/icamsl/

[Say, 03] H. Sayoud, et al. Application des Mesures Statistiques pour la détection des points de changement des locuteurs. AMAM’2003, February 10-13, 2003 in Nice, France.

http://acm.emath.fr/amam/index.php.

[Say, 03’] H. Sayoud, et al. Application of Statistical Measures in Speaker Identification. AMAM’2003, February 10-13, 2003 in Nice (AMAM’2003), France. http://acm.emath.fr/amam/index.php.

[Say, 03”] H. Sayoud, et al. Application of the MLVQ1 in Speaker Identification. NOLISP’03, 20-23 Mai 2003. Le Croisic, France.

[Say, 03”’] H. Sayoud, et al. ‘ASTRA’ An Automatic Speaker Tracking System based on SOSM measures and an Interlaced Indexation. Publication dans la revue Acta Acustica, No4, Vol 89 2003. pp 702-710.

[Sch, 98] Scherer K. R., Johnstone T., Sangsue J. L'état émotionnel du locuteur : facteur négligé mais non négligeable pour la technologie de la parole. XXIIémes Journées d'Etudes sur la Parole (JEP), pages 249{257, Juin 1998, Martigny (Suisse). 185

[Set, 94] Setlur A., Jacobs T. Results of a speaker verification service trials using HMM models. Workshop on Automatic Speaker Recognition, Identification, Verification, pages 639{642, Avril 1994, Martigny (Suisse).

[Son, 99] Sonmez K., Heck L. P., Weintraub M. Speaker tracking and detection with multiple speakers. European Conference on Speech Communication and Technology (Eurospeech), Septembre 1999, Budapest (Hongrie).

[Soo, 88] Soong F. K., Rosenberg A. E. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Acoustics Transactions, Speech, and Signal Processing (ASSP), volume 36(6), pages 871{879, Juin 1988.

[Soo, 92] Soong F. K., Rosenberg A. E., Rabiner L. R., Juang B. H. A vector quantization approach to speaker recognition. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 387{390, 1992, Tampa (USA).

[Van, 96] Van Vuuren S. Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch. International Conference on Spoken Language Processing (ICSLP), pages 1788{1791, 1996, Philadelphia (USA).

[Yu , 95] Yu K., Mason J. S., Oglesby J. Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation. IEE vision, image and signal processing, 1995, Berlin (Allemagne).

[Zaa, 00] Ibrahim ZAAF & Cherif SMAILI. « Systeme De Reconnaissance De Mots Isoles Base Sur Deux Approches : Reseaux De Neurones (Ann) Et Programmation Dynamique (Dtw) ». Université des sciences et de la technologie Houari Boumedienne. Institut d’informatique, Promotion 2000.