Annexe

Règle de la Retro-Propagation du Gradient

Nous décrivons, ici, les différentes étapes requises par l'algorithme de la RPG :

Soit un ensemble Q de paires de vecteurs $\{(X_1,Y_1), (X_2,Y_2),\ldots\ldots, (X_q,Y_q)\}$ avec :

$X_p = (X_{p1},X_{p2},\ldots\ldots..,X_{p\,n})^{t} \in R^{n}$: un vecteur d'entrée.

$Y_p = (Y_{p1},Y_{p2},\ldots\ldots..,Y_{p\,m})^{t} \in R^{m}$: un vecteur de sortie désirée.

Nous voulons entraîner le réseau pour qu'il apprenne l'approximation d'une fonction de transfert qui associe à chaque vecteur d'entrée X_P un vecteur de sortie Y_P. Cela revient à trouver un ensemble approprié de poids pour le réseau en minimisant l'erreur quadratique (des poids) commise sur l'ensemble des exemples par la technique de la descente du gradient.

Pour simplifier les calculs, considérons un réseau à trois couches, et le même raisonnement est valable pour un nombre quelconque de couches.

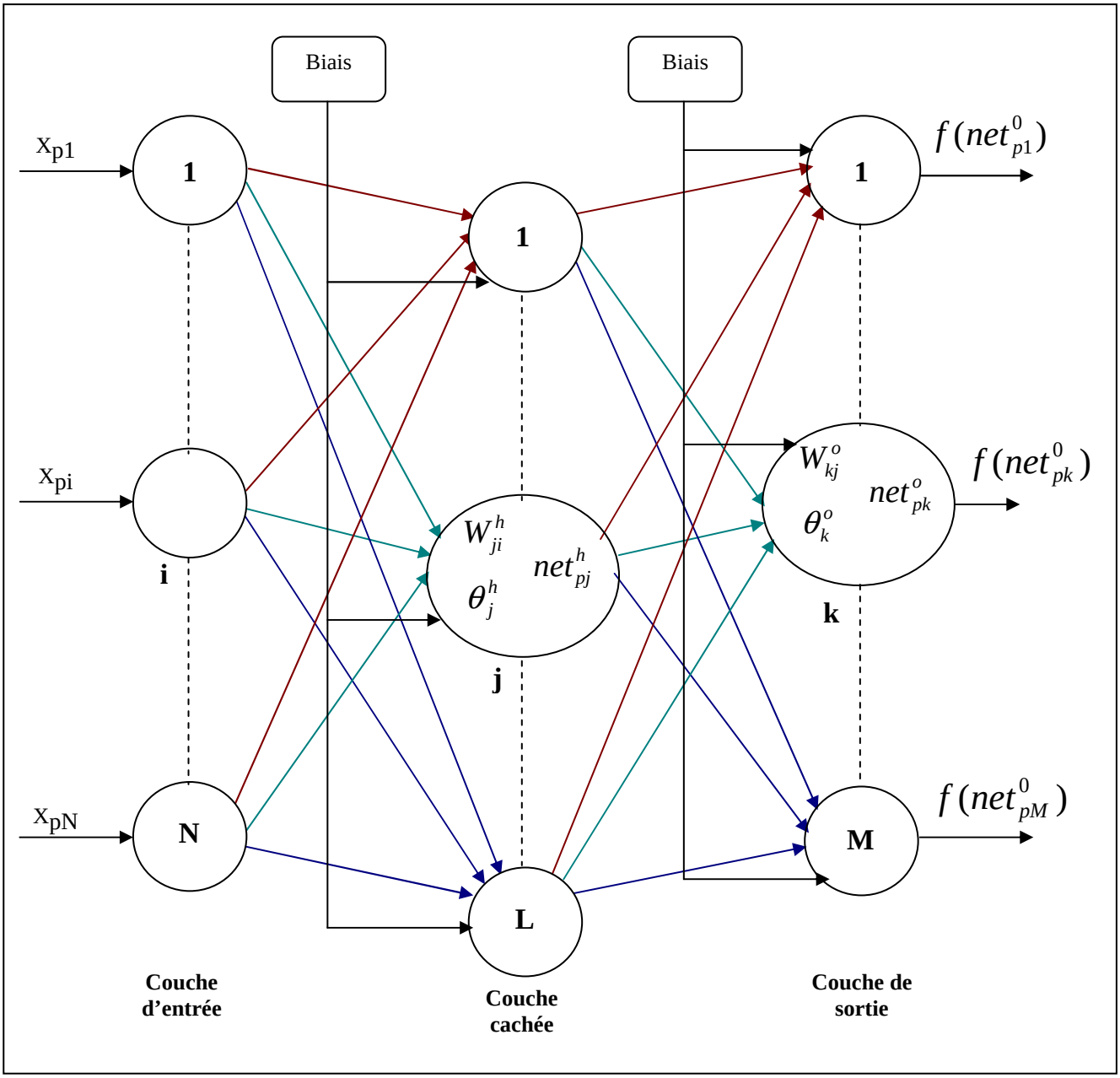


Figure 1: Réseau de neurones à une couche cachée

Dans ce qui suit, nous utiliserons la notation suivante :

Wji : le poids de la connexion entre le neurone i et le neurone j.

net^{h}_{pj} : l'entrée totale du neurone j pour l'exemple p sur la couche cachée (h : hidden layer)

net^{o}_{pk} : l'entrée totale du neurone k pour l'exemple p sur la couche de sortie (O : output layer)

θ^{h}_{j} : est le terme du biais du neurone j pour l'exemple p sur la couche cachée.

θ^{o}_{k} : est le terme du biais du neurone k pour l'exemple p sur la couche de sortie.

f : la fonction logistique.

μ : le pas d'apprentissage.

N : le nombre de neurones de la couche d'entrée.

L : le nombre de neurones de la couche cachée.

M : le nombre de neurones de la couche de sortie.

Soit un vecteur d'entrée X_p appliqué à la couche d'entrée du réseau.

Chaque neurone transmet sa valeur aux unités cachées, donc, l'entrée totale de la j^{ème} unité cachée sera :

$$net^{h}_{pj} = \sum_{i=1}^{N} W^{h}_{ji} \times X_{pi} + \theta^{h}_{j} \qquad (6)$$

Ainsi, la sortie de cette j^{ème} unité sera : $I_{Pj} = f\left(net^{h}_{Pj}\right)$ (7)

De la même manière, chaque neurone transmet sa valeur aux unités de la couche de sortie, donc la k^{ème} unité de sortie aura comme entrée totale :

$$net^{o}_{Pk} = \sum_{j=1}^{L} W^{o}_{kj} \times I_{Pj} + \theta^{o}_{k} \qquad (8)$$

et comme sortie : $O_{Pk} = f(net^{o}_{Pk})$ (9)

Mise à jour des poids de la couche de sortie :

L'erreur commise sur le k^{ème} neurone de sortie est :

$$\delta_{Pk} = \left(y_{Pk} - O_{Pk}\right) \qquad (10)$$

avec Y_{Pk} : la sortie désirée associée à la k^{ème} unité de sortie et O_{Pk} : la sortie effective.

L'erreur à minimiser est la somme quadratique des erreurs pour tous les neurones de sortie :

$$E_P = \frac{1}{2}\sum_{K=1}^{M} \delta^{2}_{pk} \qquad (11)$$

Remarque : Le facteur 1/2 est rajouté pour commodité dans les calculs des dérivés ci-dessous. Puisqu'une constante arbitraire apparaîtra dans le résultat final, la présentation de ce facteur n'infirme pas la dérivation.

Les poids de la couche de sortie sont ajustés selon la descente du gradient par la formule suivante [Fre 92] :

$$W_{Kj}^{o}(t+1) = W_{Kj}^{o}(t) + \Delta W_{Kj}^{o}(t) \qquad (12)$$

$$\Delta_P W_{Kj}^{o}(t) = \mu\left(\frac{\delta E_P}{\delta W_{Kj}^{o}}\right) = \mu\left(y_{Pk} - O_{Pk}\right) \times I_{Pj} \qquad (13)$$

Mise à jour des poids de la couche cachée :

Comme on ne connaît pas les sorties désirées pour la couche cachée, on ne peut pas connaître directement le terme d'erreur associé à chacune d'elles. Il faut donc exprimer l'erreur de la sortie de chaque neurone (d'une couche quelconque) en fonction de l'erreur de la dernière couche qui est la seule erreur directement mesurable.

Intuitivement, l'erreur totale Ep, devrait s'exprimer en fonction des valeurs de sortie de la couche cachée.

D'après les deux équations (10) et (11), on a :

$$E_P = 1/2 \sum_{K=1}^{M} \left(y_{Pk} - O_{Pk}\right)^2 \qquad (14)$$

$$E_P = 1/2 \sum_{K=1}^{M} \left(y_{Pk} - f\left(net_{Pk}^{o}\right)\right)^2 \qquad (15)$$

$$E_P = 1/2 \sum_{K=1}^{M} \left(y_{Pk} - f\left(\sum_{K=1}^{L} W_{Kj}^{o} i_{Pj} + \theta_K^{o}\right)\right)^2 \qquad (16)$$

Après un certain nombre de transformation et de minimisation de l'équation (16), nous déduirons l'équation de mise à jours des poids de la couche d'entrée.

$$W_{Kj}^{h}(t+1) = W_{Kj}^{h}(t) + \mu \delta_{Pk}^{h} X_{Pi} \qquad (17)$$

Notons que les termes d'erreur pour les unités de la couche cachée sont calculés avant que les connexions des poids aux unités de la couche de sortie ne sont mises à jour.

L'algorithme suivant résume les différentes étapes de la règle de rétro-propagation :

Etape 1 : Appliquer une valeur d'entrée $X_p = (X_{p1}, X_{p2}, \ldots\ldots, X_{p\,n})^t$ aux unités d'entrée.

Etape 2 : Calculer les entrées totales des neurones cachés : $net_{Pj}^{h} = \sum_{i=1}^{N} W_{Ji}^{h} \times X_{Pi} + \theta_J^{h}$ (18)

Etape 3 : Calculer les sorties des neurones cachés : $I_{Pj} = f\left(net_{Pj}^{h}\right)$ (19)

Etape 4 : Calculer les entrées totales de la couche de sortie : $net_{Pk}^{o} = \sum_{j=1}^{L} W_{Kj} \times I_{pj} + \theta_{k}^{o}$ (20)

Etape 5 : Calculer les sorties réelles des neurones de sortie : $O_{Pk} = f\left(net_{Pk}^{o}\right)$ (21)

Etape 6 : Calculer les termes d'erreurs sur chaque unité de sortie : $\delta_{Pk}^{o} = y_{Pk} - O_{Pk}$ (22)

Etape 7 : Calculer les termes d'erreurs sur chaque unité cachée : $\delta_{Pj}^{h} = f'\left(net_{pj}^{h}\right)\sum_{k=1}^{m} \delta_{Pk}^{o} W_{Kj}^{o}$ (23)

Etape 8 : Mettre à jour les poids de la couche de sortie : $W_{Kj}^{o} = W_{Kj}^{o} + \mu\delta_{Pk}^{o} I_{Pj}$ (24)

Etape 9 : Mettre à jour les poids de la couche cachée : $W_{Ji}^{h} = W_{Ji}^{h} + \mu\delta_{Pj}^{h} X_{Pi}$ (25)

Etape 10 : Répéter ce processus jusqu'à ce que l'erreur $E_P = 1/2 \sum_{K=1}^{M} \delta_{pk}^{2}$ (26)

devienne acceptable pour tous les exemples ; c'est à dire inférieure à l'erreur fixée avant l'apprentissage.

- Chapitre II -

- La Reconnaissance Automatique du Locuteur : Position du Problème, Définition et Applications.

Cet important chapitre définit les notions de base en RAL en décrivant brièvement l'état de l'art dans ce domaine. Il 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 rencontrés dans ce domaine de recherche sont aussi exposés comme la variabilité intra-locuteur ou la variabilité due au matériel. Finalement, un point est consacré aux dernières tendances du domaine.

II.1. Position du problème et notion de variabilité

II.1.1. Position du problème

Le terme « reconnaissance » est défini comme étant l'identification de quelque chose, sachant qu'on doit connaître au préalable le modèle de référence de cette chose.

La Reconnaissance des formes consiste, alors, à identifier une forme donnée, après avoir déjà fabriqué le modèle de référence de ces formes. Cette deuxième phase s'appelle phase d'apprentissage.

La reconnaissance automatique de l'individu consiste à utiliser des caractéristiques physiques dans le but de faire une discrimination entre ces différents individus. Pour ce faire, plusieurs caractéristiques sont proposées dans la littérature : la photographie du visage, les empreintes du visage, les traits génétiques ou encore le signal de parole.

La caractéristique la plus pratique reste le signal de parole, vues la simplicité de son extraction et la possibilité d'offrir une authentification à distance.

L'authentification par la voix est appelée « Reconnaissance Automatique du Locuteur ». Cependant, ici nous rencontrons un problème majeur, qui est défini par la difficulté de trouver des caractéristiques pertinentes en discrimination : ce problème implique la nécessité de trouver des caractéristiques possédant une grande variabilité inter-locuteur et une faible variabilité intra-locuteur.

II.1.2. Notion de variabilité

Il existe deux types de variabilités (pour une caractéristique acoustique donnée) :

- **Variabilité Inter-Locuteur** : C'est la variance des moyennes (de cette caractéristique) pour les différents locuteurs.

- **Variabilité Intra-Locuteur** : C'est la moyenne des variances (de cette caractéristique) des différents locuteurs : ici, pour chaque locuteur une variance est calculée.

- **Coefficient de Wolf « CW »** : C'est le rapport de la Variabilité Inter-Locuteur sur la Variabilité Intra-Locuteur [Wol 72].

Une grande Variabilité Inter-Locuteur : indique qu'on peut, facilement, séparer les locuteurs par leurs caractéristiques.

Une petite Variabilité Intra-Locuteur : indique que chaque locuteur peut être représenté par une référence qui représente très bien ce locuteur.

Par conséquent, un grand coefficient de Wolf : indique, alors, que le paramètre choisi est pertinent en identification du locuteur et devrait donner un bon score de reconnaissance.

Dans ce qui suit, nous définirons les différentes étapes associées à la Reconnaissance du Locuteur, ainsi que les méthodes proposées dans ce domaine.

II.2. La Reconnaissance Automatique du Locuteur

II.2.1. 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.

Niveau de dépendance au texte :
Une première classification des systèmes de RAL repose sur le niveau de dépendance au texte. En premier lieu, on distingue généralement les systèmes dépendants du texte des systèmes indépendants du texte.

En mode dépendant du texte, la reconnaissance d'une personne est réalisée sur la base d'un message dont le contenu linguistique (mot de passe, phrase...) est connu du système.

En mode indépendant du texte, le système de reconnaissance n'a aucune connaissance sur le message linguistique prononcé par la personne.

Concernant le mode dépendant du texte, une terminologie plus fine peut être donnée à un système suivant l'application visée. Celle-ci est inspirée de la littérature ainsi que des travaux de Eagles [Eag 95] :

- systèmes à messages fixés : la personne est contrainte de prononcer un message, qu'elle aurait au préalable (mots de passe personnalises : [Jac 00], [Kha 00]) ou qui sera imposé par le système
- systèmes à messages promptes : un message, différent à chaque nouvelle session de reconnaissance, est imposé par le système sous forme visuelle [Mat 94b] ou auditive [Lin 97]. Ces systèmes ont pour première motivation de se protéger des attaques de personnes malveillantes (imposteurs) qui disposeraient d'un enregistrement de la voix d'une personne.
- systèmes à unités segmentales: la personne doit prononcer un message comportant soit une séquence de mots (séquence de chiffres), soit des traits phonétiques (séquence de phonèmes) connus du système.

La connaissance a priori partielle ou totale du message prononcé par la personne rend généralement les systèmes dépendants du texte plus performants que les systèmes indépendants du texte. En mode dépendant du texte, les systèmes s'affranchissent du problème de la variabilité linguistique.

II.2.2. Différentes Tâches en RAL

L'Identification Automatique du Locuteur et la Vérification Automatique du Locuteur sont les tâches pionnières du domaine de la RAL, où plusieurs chercheurs sont entrain de travailler : [Ata 76], [Dod 85], [O'Sh 86], [Ros 91], [Nai 94], [Fur 94], [Fur 97], [Dod 98].

Plus récemment, les besoins applicatifs ont fait naître de nouvelles tâches comme l'Indexation par Locuteur de flux audio (travaux de Johnson et de Delacourt) [Joh 99], [Del 00] ou le Suivi de Locuteurs - en Anglais : speaker tracking- (travaux de Rosenberg, Sonmez, Bonastre et Martin) [Ros 98a], [Son 99], [Bon 00a], [Bon 00b], [Mar 00] ou de nouvelles variantes telles que la détection d'un locuteur dans une conversation (travaux de Przybocki) [Prz 99], [Mar 00].

II.2.2.1. 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 2.1), 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 [Ros 98b].

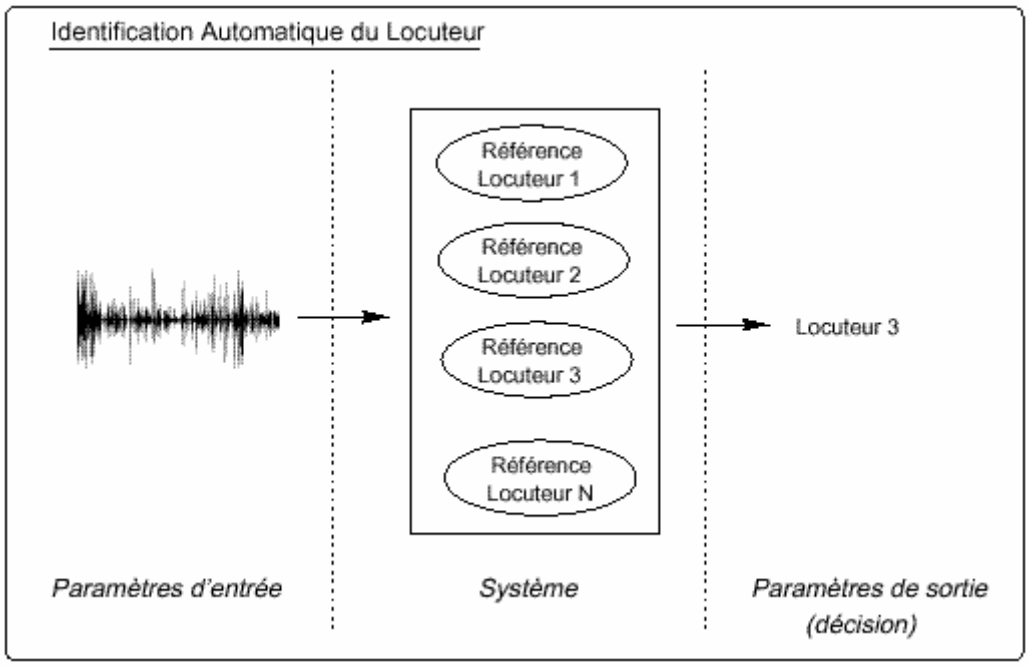


Figure 2.1: La tâche d'IAL. Principe de base de la tâche d'Identification Automatique du Locuteur.

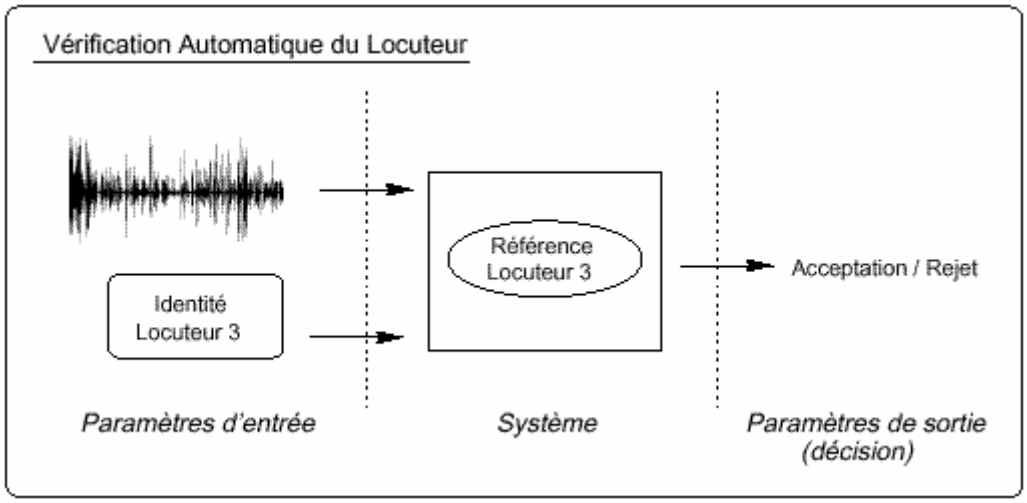


Figure 2.2: La tâche de VAL. Principe de base de la tâche de Vérification Automatique du Locuteur.

II.2.2.2. 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 2.2). 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. Dans le cas contraire, l'individu est considéré comme un imposteur et rejeté.

Applications

Les applications de VAL sont multiples et principalement commerciales [Bov 98] :

- 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.2.2.3. 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.2.2.4. Indexation par Locuteur et ses variantes

La tâche d'Indexation Automatique par Locuteur consiste à cibler les interventions des locuteurs dans un flux audio (figure 2.3). 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 (figure 2.4). 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.

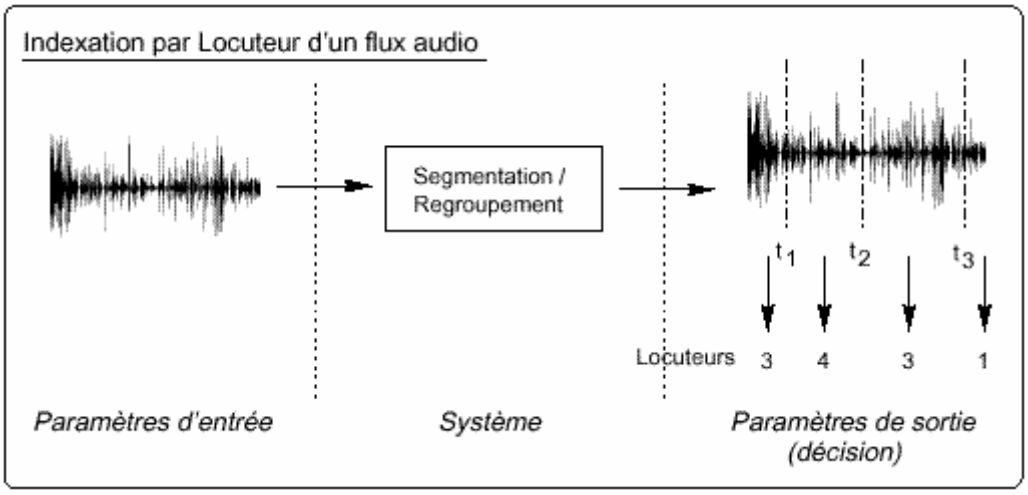


Figure 2.3: La tâche d'Indexation par Locuteur d'un flux audio. Principe de base.

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 ; comme c'est décrit par Bonastre [Bon 00b].

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 appliqués 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 ; comme c'est décrit par Rosenberg et Bonastre [Ros 98a], [Bon 00a].

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...) ; comme c'est décrit par Sonmez et Meignier [Son 99], [Mei 00].

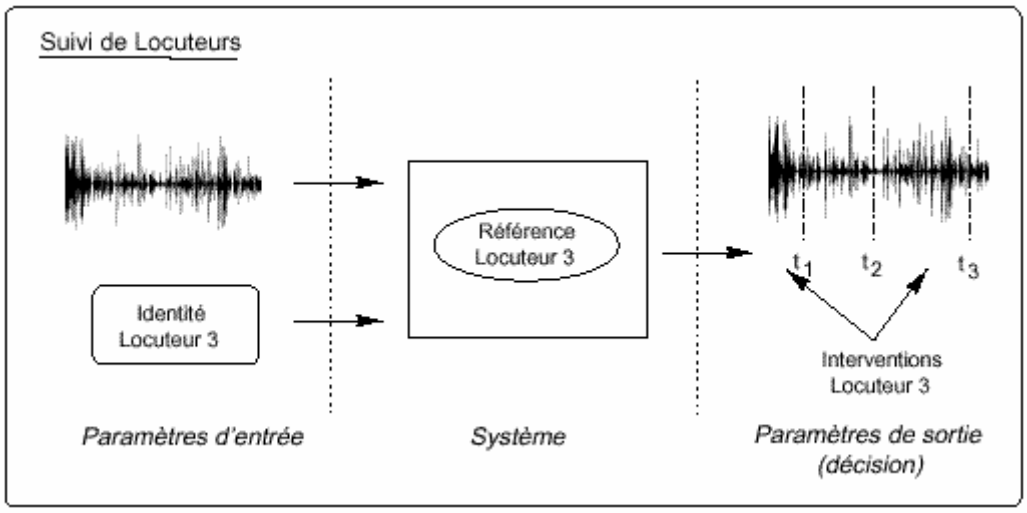


Figure 2.4: La tâche de suivi de locuteurs. Principe de base.

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.2.2.5. Applications criminalistiques

Un volet que nous n'avons pas encore évoqué est l'utilisation de la RAL dans les domaines judiciaires ou criminalistiques (voir travaux de Hollien, Kunzel, Boe et Champod) [Hol 90], [Kun 94], [Boe 98], [Cham 98]. 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 98], [Boe 99].

II.2.3. 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.2.4. Problèmes rencontres 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), comme c'est démontré par Homayounpour, Scherer, Karlsson et Banziger [Hom 95], [Sch 98], [Kar 98], [Ban 00], 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 (voir travaux de Furui, Rosenberg et Setlur) [Fur 77], [Ros 76], [Set 94]. 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 ; comme c'est démontré par Van Vuuren [Van 96].

Par exemple [Rey 96] [Auc 00], Reynolds et Auckenthaler ont démontré que des différences de types de combinés 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 ; comme expliqué par Fissore [Fis 99]. 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 porté sur la robustesse des systèmes de RAL à travers le réseau GSM (nous pouvons, toutefois, citer les travaux de Besacier et de Quatieri [Bes 00b] [Qua 00]). 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 [Hom 95].

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.

Contraintes imposées par le domaine applicatif

Le domaine applicatif et en particulier commercial impose des contraintes fortes quant à l'utilisation d'un système de RAL. Notamment, les sessions d'enrôlement doivent être peu contraignantes pour les clients d'une application. Dans ce sens, elles sont généralement très peu nombreuses et peu espacées dans le temps. Aussi, la quantité de signaux de parole collectés lors de ces sessions s'avère insuffisante pour une bonne estimation des modèles clients. D'autre part, au cours de ces sessions d'enrôlement, les conditions d'utilisation du système, notamment pour un service accessible par le réseau téléphonique, sont peu variées (appel du domicile du client, du lieu de travail, d'une cabine publique...). Aussi, peu de variabilité due au matériel utilisé est introduite dans les signaux d'apprentissage.

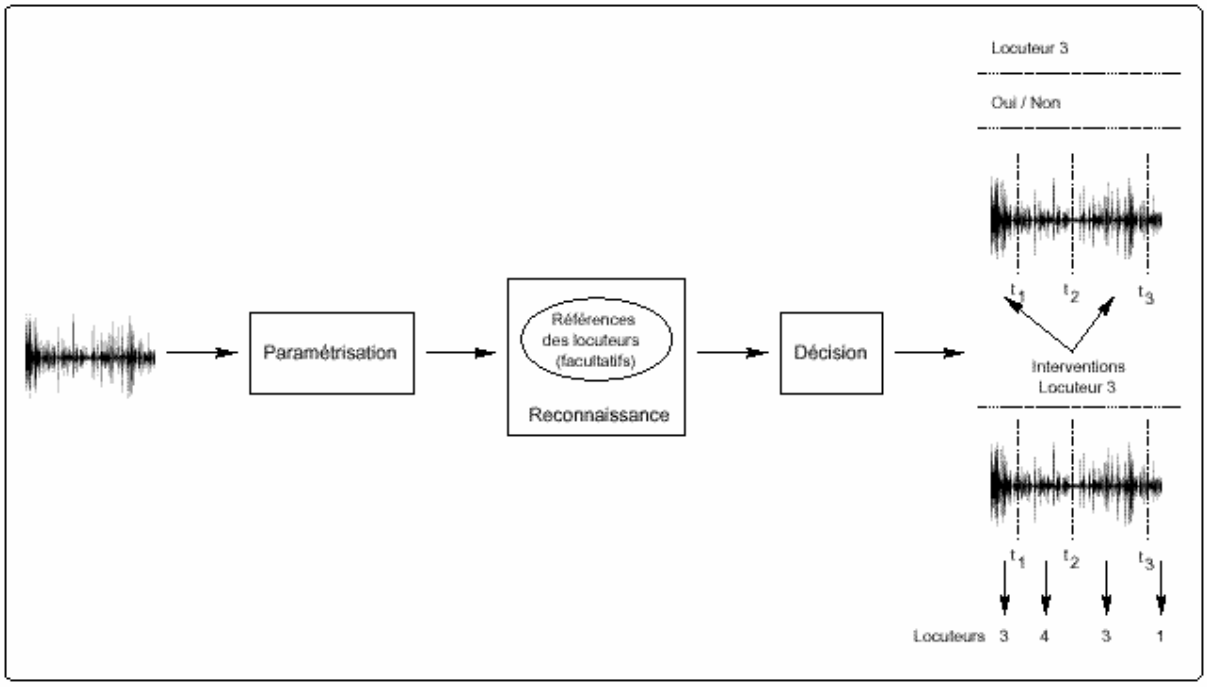


Figure 2.5: Structure d'un système de RAL. Illustration des trois grands processus œuvrant au sein d'un système de RAL.

II.3. 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 (illustration figure 2.5). 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 aux travaux de Delacourt [Del 00] 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.3.1. 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.3.1.1. 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 de Reynolds, Homayounpour et Charlet [Rey 94], [Hom 94], [Cha 97], 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 [Gre 77] : LPCC (Linear Predictive Cepstral Coefficients) ou LPC (Linear Predictive Coefficients) ;

- coefficients spectraux issus d'une analyse en banc de filtres : 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.3.1.2. 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 [Ata 76], 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.3.1.3. Paramètres dynamiques

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) ; comme c'est décrit par Furui, Soong et Bernasconi [Fur 81], [Soo 88], [Ber 90]. 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) ; comme c'est décrit par Magrin-Chagnolleau [Mag 99], la concatenation de trames successives de signal (voir les travaux de Hattori, Konig et Fredouille) [Hat 92], [Kon 98], [Fre 98], [Fre 00a]. Un panorama et un bref descriptif de ces différentes approches sont fournis dans [Fre 00].

II.3.2. 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.3.2.1. 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 (travaux de Furui, Booth et Yu) [Fur 81], [Boo 93], [Yu 95]. 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 (voir travaux de Soong, Mason et Matsui) [Soo 92], [Mas 89], [Mat 92]. 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.3.2.2. 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 ; comme c'est décrit par Pruzansky [Pru 63]. 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 [Bim 95]. 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 de covariance X et Y (voir travaux de Gish, Bimbot et Magrin) [Gis 86], [Bim 95], [Mag 96].

Par exemple, la mesure symétrisée μ_G, que nous utiliserons dans la suite de cette thèse, s'exprime sous la forme :

$$\mu_G(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{p}\left[tr(YX^{-1}) - \log\left(\frac{\det Y}{\det X}\right) + (\bar{y}-\bar{x})^T X^{-1}(\bar{y}-\bar{x})\right] - 1 \qquad (2.1\text{-a})$$

et

$$\mu_{G\beta}(\boldsymbol{x}, \boldsymbol{y}) = (M*\mu_G(\boldsymbol{x}, \boldsymbol{y}) + N*\mu_G(\boldsymbol{y}, \boldsymbol{x})) / (M+N) \qquad (2.1\text{-b})$$

L'avantage majeur des SOSM est leur simplicité de mise en œuvre. Performantes sur de courtes durées (3 secondes) [Bim 95], 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) ; comme c'est décrit par Reynolds [Rey 92], [Rey 95], [Rey 00]. 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 moyenne et une matrice de covariance. Lors de l'apprentissage, les paramètres des modèles clients (vecteur moyenne $\bar{x}_i$, matrice de covariance Σ_i, pondération p_i de chaque distribution gaussienne) sont généralement estimés à l'aide de l'algorithme EM (Expectation-Maximization) [Dem 77] 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. La vraisemblance pour qu'un vecteur de test, y_t, soit produit par le mélange de gaussienne χ s'exprime par :

$$\begin{aligned} L(y_t|\chi) &= \sum_{i=1}^{M} p_i L_i(y_t) \\ L_i(y_t) &= \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(y_t-\bar{x}_i)^T(\Sigma_i)^{-1}(y_t-\bar{x}_i)\right\} \end{aligned} \qquad (2.2)$$

où p_i, $\bar{x}$ et Σ_i représentent, respectivement, le poids, le vecteur moyenne (de dimension D) et la matrice de covariance (de dimension DxD) de la $i^{ème}$ distribution gaussienne. 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 (travaux de Rabiner) [Rab 89], 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.

NB : Les mélanges de gaussiennes peuvent être considérés comme un modèle de Markov caché à un seul état. De même, la quantification vectorielle décrite précédemment est souvent interprétée comme une dégénérescence des modèles de Markov cachés à un seul état pour lequel les probabilités émission sont remplacées par des mesures de distance [Fur 95], [Pie 98]. De par leur principe, les modèles de Markov cachés s'appliquent parfaitement au mode dépendant du texte, obtenant d'excellents résultats (d'après les résultats de Rosenberg et De Veth) [Ros 91], [De 94]. En revanche, l'utilisation des modèles HMM en mode indépendant du texte n'améliore pas les performances obtenues par des modèles plus simples à base de GMM [Rey 00].

II.3.2.3. 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 qui ont été utilisés par Oglesby et Frederickson [Ogl 91], [Fre 94]. 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 mises en oeuvre 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.3.2.4. 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 [Gre 80], [Bim 92], [Mon 92], [Gri 94], [Mag 96] et les réseaux prédictifs [Hat 92], [Art 93], [Ben 94], [Pao 96]. Ces deux techniques ont été très bien décrites par Fredouille dans [Fre 00].

II.3.4. Décision et mesure des performances

Comme souligné en introduction de cette section, le processus de décision est dépendant de la tâche visée :

II.3.4.1. Identification Automatique du Locuteur

En identification, un signal de test est comparé à toutes les références de locuteurs connus du système, résultant en un ensemble de mesures de similarité en entrée du processus de décision. Aussi, ce processus a pour tâche de rechercher la mesure de similarité maximale (ou minimale dans le cas de mesures de distance) et de désigner l'identité du locuteur correspondant. Dans ce contexte, la mesure des performances d'un système d'IAL est généralement donnée en termes de taux d'identification correcte. Ce taux s'obtient par la formulation suivante :

$$\text{Taux d'identification correcte} = \frac{\text{nombre de tests ayant amené à une identification correcte}}{\text{nombre de tests total}} \qquad (2.3)$$

II.3.4.2. Vérification Automatique du Locuteur

En vérification, le processus de décision consiste à comparer la mesure de similarité entre le signal de test et le modèle client à un seuil de décision. Si la mesure est supérieure au seuil, le locuteur est considère comme un client et est accepté. Dans le cas contraire, le locuteur est considéré comme un imposteur et est rejeté.

Plusieurs solutions sont proposées dans la littérature pour mesurer les performances d'un système de VAL.

Toutes s'appuient sur deux erreurs caractéristiques des systèmes de VAL que sont :

- l'erreur de fausse acceptation : le système reconnaît à tort une identité revendiquée (acceptation d'un imposteur).
- l'erreur de faux rejet : le système rejette à tort une identité revendiquée (rejet d'un client).

Ces erreurs conduisent aux taux d'erreurs de fausse acceptation p(FA) et de faux rejet p(FR) calcules pour un seuil de décision donné et définis par :

$$p(FA) = \frac{\text{nombre de tests ayant amené à une fausse acceptation}}{\text{nombre de tests imposteurs}} \qquad (2.4\text{-a})$$

$$p(FR) = \frac{\text{nombre de tests ayant amené à un faux rejet}}{\text{nombre de tests clients}} \qquad (2.4\text{-b})$$

Performances intrinsèques du système

Les performances intrinsèques d'un système de VAL sont estimées en comparant les mesures de similarité, issues de tests clients et imposteurs, à différentes valeurs de seuil. A chaque valeur de seuil est associé un couple (p(FA); p(FR)). L'ensemble des couples obtenus se représente sous la forme d'une courbe COR - Caractéristique Opérationnelle du Récepteur - (Receiver Operating Characteristic : ROC) [Ogl 95] ou d'une courbe DET (Detection Error Tradeoff) [Mar 97] sur lesquelles les taux p(FA) sont donnés en fonction des taux p(FR). Les performances intrinsèques d'un système sont très souvent résumées par un point particulier des courbes ROC et DET indiquant un taux identique d'erreurs de fausse acceptation et de faux rejet : taux d'égale erreur (Equal Error Rate : EER).

Performances d'exploitation du système

Dès lors que le seuil de décision est fixé a priori, les performances d'exploitation d'un système de VAL peuvent être évaluées au moyen du couple (p(FA); p(FR)) correspondant. Ce couple peut être exprimé sous la forme d'une moyenne des deux taux, appelée HTER [Bimbot et al., 1998] ou encore sous la forme d'une fonction de coût total du système faisant intervenir les coûts respectifs des erreurs de fausses acceptations et de faux rejets [Fre 00].

II.3.4.3. Suivi de locuteurs

Comme nous l'avons évoqué précédemment, le processus de décision repose, pour la tâche de suivi de locuteurs, soit sur une décision classique de VAL (à base de seuils de décision), soit sur un décodage basé sur un modèle HMM.

Décision classique de VAL

Le processus de décision compare la mesure de similarité obtenue sur chaque bloc ou segment à un seuil de décision. Si la mesure de similarité est supérieure au seuil, le bloc ou segment est attribué au locuteur cible. Dans le cas contraire, le bloc ou segment est rejeté.

Modèle HMM

Lors du processus de décision, un algorithme de type Viterbi attribue chaque bloc à un état particulier du modèle HMM (état parole, état non-parole ou état locuteur). La décision

consiste à accepter tous les blocs attribués à l'état du locuteur cible. D'une manière similaire aux systèmes de VAL, deux types d'erreurs peuvent être produites par un système de suivi de locuteurs : les erreurs de faux rejet ou de fausse acceptation.

Les taux de faux rejet et de fausse acceptation s'expriment ici différemment par :

$$p(FA) = \frac{\text{nombre de trames attribuées au locuteur cible}}{\text{nombre de trames ne correspondant pas au locuteur cible}} \qquad (2.5\text{-a})$$

$$p(FR) = \frac{\text{nombre de trames non attribuées au locuteur cible}}{\text{nombre de trames correspondant au locuteur cible}} \qquad (2.5\text{-b})$$

Dans ces conditions, les systèmes de suivi de locuteurs bénéficient des mêmes mesures de performances que les systèmes de VAL (courbes ROC, DET...).

II.3.5. Evaluation des systèmes de RAL

A l'heure actuelle une seule grande campagne d'évaluation est dédiée aux systèmes de RAL. Cette campagne, organisée chaque année par l'institut américain NIST (National Institute of Standards and Technologies) depuis 1996, est accessible aux laboratoires de recherche publics ou privés. Basées initialement sur l'évaluation des systèmes de VAL dans un contexte conversationnel et téléphonique, ces campagnes se sont rapidement étendues aux tâches de détection d'un locuteur dans une conversation, de suivi de locuteurs, et plus récemment d'Indexation Automatique par Locuteur d'un flux audio. Avant chaque nouvelle campagne d'évaluation, un corpus de développement est défini et distribué aux participants, permettant la mise au point des systèmes de RAL.

Lors de l'évaluation à proprement parler, chaque laboratoire participant reçoit un nouveau corpus de données comprenant les signaux d'apprentissage et de test pour chacune des tâches à réaliser. Les systèmes de RAL sont alors soumis, en aveugle, à diverses séries de tests dont les résultats doivent être retournés en un temps donné (généralement trois semaines). Ces campagnes sont intéressantes à plusieurs niveaux. Elles permettent en premier lieu d'avoir accès à des sous-corpus de la base de données Switchboard, organises spécialement pour les tâches de RAL. Ces sous-corpus constituent une base de plus de 1000 locuteurs (hommes et femmes) dans un contexte conversationnel et téléphonique. D'autre part, il devient facile, sur une base de données commune, de comparer les performances des différents systèmes et d'apprécier les nouvelles tendances et techniques proposées.

Par exemple, les techniques GMM ont fait l'unanimité pour la tâche de détection de locuteurs au cours des évaluations de cette année (campagne évaluation NIST 2000).

II.4. 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 sont rapportés par Reynolds sur l'utilisation d'un modèle générique de locuteurs pour pallier le problème des données manquantes [Rey 97], [Rey 00]. 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) ; comme c'est décrit par Gauvin [Gau 94] ou MLLR (Maximum Likelihood Linear Regression) ; comme c'est décrit par Leggetter [Leg 95].

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 supervisé [Fre 00b].

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.

II.5. Travaux antérieurs

La RAL a débuté, il y'a plus de 25 ans, avec des méthodes d'identification élémentaires dont la dimension des bases de données ne dépassait pas les 10 locuteurs. Les scores cités à l'époque approximaient les 90% de bonne reconnaissance (travaux de Atal) [Atal 76].

Depuis, les chercheurs n'ont guère cessé de proposer de nouvelles techniques et de nouvelles caractéristiques pour l'IAL. Ainsi les méthodes statistiques ont révélé une grande performance de reconnaissance en donnant des scores proche de 100% sur des populations de plus de 600 locuteurs (travaux de Reynolds et Bimbot) [Rey 95] [Bim 95].

Des chercheurs impressionnés par les réseaux de neurones, comme Oglesby, ont proposé des méthodes connexionnistes à base de réseaux RBF, qui ont donné un score de 91% sur 40 locuteurs [Ogl 91]. Mais les résultats obtenus par ce type de techniques n'étaient pas remarquables.

De nouvelles tendances, telles que les modèles hybrides (statistico-connexionnistes), sont apparues en dévoilant de grandes performances sur les populations de grande dimension : les scores cités avoisinent les 100% [Ben 94]. Cependant la complexité des réseaux de neurones rend leur utilisation hésitante.

A l'heure actuelle, les chercheurs préfèrent toujours utiliser les méthodes statistiques (surtout à base de GMM) qui sont considérées comme étant des méthodes de référence et de pointe.

- Chapitre III -

- Approche Statistique -

III.1. Introduction

L'étude statistique consiste à trouver des mesures de vraisemblance, basées exclusivement sur des caractéristiques statistiques, pour l'identification du locuteur en milieu sain et en milieu corrompu. Dans cet objectif, nous avons élaboré un système automatique pour la reconnaissance automatique du locuteur, basé exclusivement sur des mesures statistiques de vraisemblance appelées SOSM ou : *Second Order Statistical Measures* [Say 03c].

L'approche statistique a montré une très bonne performance 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.

Dans la suite de ce chapitre, nous définirons les mesures de vraisemblance statistiques, la technique SOSM, ainsi que le protocole d'apprentissage et de test utilisé dans nos expériences.

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 :

$$\bar{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \qquad \text{(3.1.a)} \quad \text{et} \qquad X = \frac{1}{M}\sum_{t=1}^{M}(x_t - \bar{x})(x_t - \bar{x})^T \qquad \text{(3.1.b)}$$

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

$$\bar{y} = \frac{1}{N}\sum_{t=1}^{N} y_t \qquad (3.2.a) \quad \text{et} \quad Y = \frac{1}{N}\sum_{t=1}^{N} (y_t - \bar{y})(y_t - \bar{y})^T \qquad (3.2.b)$$

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 $\boldsymbol{x}$ (respectivement $\boldsymbol{y}$) 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 $\boldsymbol{x}$ et $\boldsymbol{y}$ peut être exprimée comme la fonction Φ suivante.

$$\mu(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\bar{x}, X, M, \bar{y}; Y, N) \qquad (3.3)$$

Elle est non-négative, c'est à dire :

$$\forall\, \boldsymbol{x},\ \forall\, \boldsymbol{y},\ 0 \le \mu(\boldsymbol{x}, \boldsymbol{y}), \qquad (3.4)$$

et elle satisfait la propriété (3.5):

$$\forall\, \boldsymbol{x},\ \mu(\boldsymbol{x}, \boldsymbol{x}) = 0. \qquad (3.5)$$

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

$$\forall\, \boldsymbol{x},\ \forall\, \boldsymbol{y},\ \mu(\boldsymbol{x}, \boldsymbol{y}) = \mu(\boldsymbol{y}, \boldsymbol{x}). \qquad (3.6)$$

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

III.2.3.1.1. Définition

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

$$G(y_t/\boldsymbol{x}) = \frac{1}{(2\Pi)^{p/2}(\det X)^{1/2}} \times \exp(-\frac{1}{2}(y_t - \bar{x})^T X^{-1}(y_t - \bar{x})) \qquad (3.7)$$

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 :

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

$$= -\frac{1}{2}\left[p\log 2\Pi + \log(\det X) + \frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{x})^T X^{-1}(y_t - \bar{x}) \right] \qquad (3.8)$$

Par ailleurs, en remplaçant $y_t - \bar{x}$ par $y_t - \bar{y} + \bar{y} - \bar{x}$ et en utilisant la propriété mathématique (3.9)

$$\frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{y})^T X^{-1}(y_t - \bar{y}) = tr(YX^{-1}) \qquad (3.9)$$

Nous aurons alors

$$\bar{G}_{\boldsymbol{x}}(y_1^N) + \frac{p}{2}\log 2\Pi = -\frac{1}{2}\left[\log(\det X) + tr(YX^{-1}) + (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x})\right] \qquad (3.10)$$

de même aussi l'expression suivante :

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

$$= \frac{1}{p}\left[\log\left(\frac{\det Y}{\det X}\right) - tr(YX^{-1}) - (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x})\right] + 1 \qquad (3.11)$$

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

$$\mu_G(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{p}\left[tr(YX^{-1}) - \log\left(\frac{\det Y}{\det X}\right) + (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x})\right] - 1 \qquad (3.12)$$

$$= \frac{1}{p}\left[tr(\Gamma) - \log(\det\Gamma) + \delta^T X^{-1}\delta\right] - 1 \qquad (3.13)$$

$$= a - \log g + \frac{1}{p}\delta^T X^{-1}\delta - 1 \qquad (3.14)$$

alors nous aurons :

$$\textbf{Argmax}_x \{\overline{G}_x(y_1^N)\} = \textbf{Argmin}_x \{\mu_G(\boldsymbol{x}, \boldsymbol{y})\} \qquad (3.15)$$

Par conséquent cette fonction μ_G peut être assimilée à une mesure de distance inversement proportionnelle à la probabilité de vraisemblance.

Sachant que

$a(\lambda_1, \lambda_2, ..., \lambda_p) = \frac{1}{p}\sum_{i=1}^{p}\lambda_i$ *représente la moyenne arithmétique,* (3.16)

$$\Gamma = X^{-\frac{1}{2}} Y X^{-\frac{1}{2}}, \qquad (3.17)$$

les λ_i *représentent les valeurs propres de* Γ,

X *représente la matrice de covariance représentant* $\boldsymbol{x}$,

Y *représente la matrice de covariance représentant* $\boldsymbol{y}$,

$g(\lambda_1, \lambda_2, ..., \lambda_p) = \left(\prod_{i=1}^{p}\lambda_i\right)^{1/p}$ *représente la moyenne géométrique,* (3.18)

et $\delta = \bar{y} - \bar{x}$. (3.19)

III.2.3.1.2. Propriété de la μ_G

La mesure gaussienne de vraisemblance μ_G est non symétrique. C'est-à-dire que, si on considère son terme dual μ_G ($\boldsymbol{y}$, $\boldsymbol{x}$),

$$\mu_G(\boldsymbol{y}, \boldsymbol{x}) = \frac{1}{h} + \log g + \frac{1}{p}\delta^T Y^{-1}\delta - 1 \qquad (3.20)$$

alors on remarque que $\mu_G(\boldsymbol{x}, \boldsymbol{y}) \neq \mu_G(\boldsymbol{y}, \boldsymbol{x})$

sachant que

$$h(\lambda_1, \lambda_2, ..., \lambda_p) = \left(\frac{1}{p} \sum_{i=1}^{p} \left(\frac{1}{\lambda_1} \right)_i \right)^{-1} \text{ représente la moyenne harmonique} \qquad (3.21)$$

III.2.3.1.3. Inconvénient de la μ_G

Quand on traite un signal de parole distordu ou bruité, les vecteurs moyenne $\bar{x}$ et $\bar{y}$ peuvent être fortement influencés par les caractéristiques du canal, tandis que les matrices de covariance X et Y sont habituellement plus robustes aux variations entre les conditions d'enregistrement et les lignes de transmission du canal. Ainsi la différence $\delta = \bar{y} - \bar{x}$ peut devenir un terme d'erreur pour la μ_G.

Par conséquent, la mesure de vraisemblance gaussienne par covariance μ_{GC} peut donc être dérivée de la mesure précédente par l'équation (3.22), en supposant que la variabilité inter-locuteur de la moyenne est nulle.

Soit :

$$\mu_{GC}(\boldsymbol{x}, \boldsymbol{y}) = a - \log g - 1 \qquad (3.22)$$

Cette mesure (mesure de vraisemblance gaussienne par covariance) est aussi non symétrique :

$$\mu_{GC}(\boldsymbol{y}, \boldsymbol{x}) = \frac{1}{h} + \log g - 1 \neq \mu_{GC}(\boldsymbol{x}, \boldsymbol{y}) \qquad (3.23)$$

III.2.3.2. La mesure arithmétique- géométrique sphérique

Cette mesure est donnée par [Bim 95] :

$$\mu_{SC}(\boldsymbol{x}, \boldsymbol{y}) = \log\left(a \big/ g \right), \qquad (3.24)$$

De même, cette mesure n'est pas symétrique :

$$\mu_{SC}(\boldsymbol{y}, \boldsymbol{x}) = \log\left(g \big/ h \right) \neq \mu_{SC}(\boldsymbol{x}, \boldsymbol{y}). \qquad (3.25)$$

III.2.3.3. La mesure de déviation absolue

Elle est donnée par [Bim 95] :

$$\mu_{DC}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{p}\sum_{i=1}^{p}\left|\lambda_i - 1\right|, \tag{3.26}$$

Cette mesure est aussi non symétrique car

$$\mu_{DC}(\boldsymbol{y}, \boldsymbol{x}) = \frac{1}{p}\sum_{i=1}^{p}\left|\frac{1}{\lambda_i} - 1\right| \neq \mu_{DC}(\boldsymbol{x}, \boldsymbol{y}) \tag{3.27}$$

III.2.4. La symétrisation

III.2.4.1. Motivation

Toutes les mesures vues précédemment ont une propriété commune d'être non symétriques, autrement dit, les rôles joués par les données de l'apprentissage et les données du test ne sont pas interchangeables. Cependant, notre intuition logique serait que la mesure de similarité devrait être symétrique.

L'asymétrie des mesures μ_G, μ_{GC}, μ_{SC} et μ_{DC} peut être expliquée par le fait que ces mesures sont basées sur des essais statistiques qui supposent que la référence (le modèle du locuteur $\boldsymbol{x}$) est exacte, tandis que le modèle test (le locuteur $\boldsymbol{y}$) est une estimation. Mais en pratique, les deux modèles de référence et de test sont des estimations.

De plus, il est clair que la fiabilité d'un modèle de référence est dépendante du nombre de données qui a été employé pour estimer ces paramètres. Ceci est confirmé expérimentalement par les différences en performance qui peuvent être observées dans les tests d'identification du locuteur entre $\mu(\boldsymbol{x}, \boldsymbol{y})$ et $\mu(\boldsymbol{y}, \boldsymbol{x})$, surtout si M et N (le nombre de vecteurs de référence et celui de test) sont disproportionnés (c'est à dire si $\rho = \frac{N}{M}$ est très différent de 1).

III.2.4.2. Procédures de symétrisation

La première possibilité pour la symétrisation de la mesure $\mu(\boldsymbol{x}, \boldsymbol{y})$, est de construire la moyenne $\mu_{[0.5]}(\boldsymbol{x}, \boldsymbol{y})$ entre la mesure et son terme dual :

$$\mu_{[0.5]}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\mu(\boldsymbol{x}, \boldsymbol{y}) + \frac{1}{2}\mu(\boldsymbol{y}, \boldsymbol{x}) = \mu_{[0.5]}(\boldsymbol{y}, \boldsymbol{x}) \tag{3.28}$$

La mesure de vraisemblance gaussienne symétrisée de cette façon, devient :

$$\mu_{G[0.5]}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\left[a + \frac{1}{h} + \frac{1}{p}\delta^T\left[X^{-1} + Y^{-1}\right]\delta\right] - 1, \tag{3.29}$$

Tandis que la mesure de vraisemblance à covariance devient :

$$\mu_{GC[0.5]}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\left[a + \frac{1}{h} \right] - 1, \qquad (3.30)$$

Cette procédure de symétrisation peut améliorer nettement les performances de classification, en comparaison avec les deux termes non symétriques pris individuellement.

Cependant, nous observons, en pratique, une dégradation des performances quand les longueurs diffèrent considérablement (i.e. $\rho \neq 1$).

Ainsi, si $\rho \leq 1 \Rightarrow M \geq N$ alors $\mu(\boldsymbol{x}, \boldsymbol{y})$ sera meilleure que $\mu(\boldsymbol{y}, \boldsymbol{x})$ et inversement si $\rho \geq 1$.

Sous le manque d'un cadre théorique rigoureux, les vérifications ont été limitées aux expériences empiriques. Une forme arbitraire a été, alors, postulée pour la généralité des mesures symétriques, lestées par des coefficients qui sont fonction du nombre de vecteurs d'apprentissage et de test (respectivement *M* et *N*).

Cette nouvelle symétrisation notée $\mu_\beta(\boldsymbol{x}, \boldsymbol{y})$ est donnée par l'expression 3.31.

$$\mu_\beta(\boldsymbol{x}, \boldsymbol{y}) = (M * \mu(\boldsymbol{x}, \boldsymbol{y}) + N * \mu(\boldsymbol{y}, \boldsymbol{x})) / (M+N) \qquad (3.31)$$

Cette nouvelle mesure est bien adaptée au cas où la durée de test et celle de l'apprentissage sont très différents.

III.3. Pré-Traitement

Le pré-traitement est constitué des étapes d'analyse du signal de parole précédant l'apprentissage et le test d'identification. Ainsi, chaque phrase est analysée selon les étapes suivantes :

- Fenêtrage du signal de parole sur des trames de 32 ms (du type Hanning) avec recouvrement de 50%.
- Détection et suppression des zones de silence par un détecteur d'activité basé sur l'énergie moyenne (figure 3.1).
- Evaluation du spectre d'énergie par une transformée de Fourier rapide sur des trames d'analyse de 512 échantillons.
- Le spectre d'énergie passe ensuite à travers un banc de filtres constitué de plusieurs filtres du type Hamming (voir fig. 3.2).

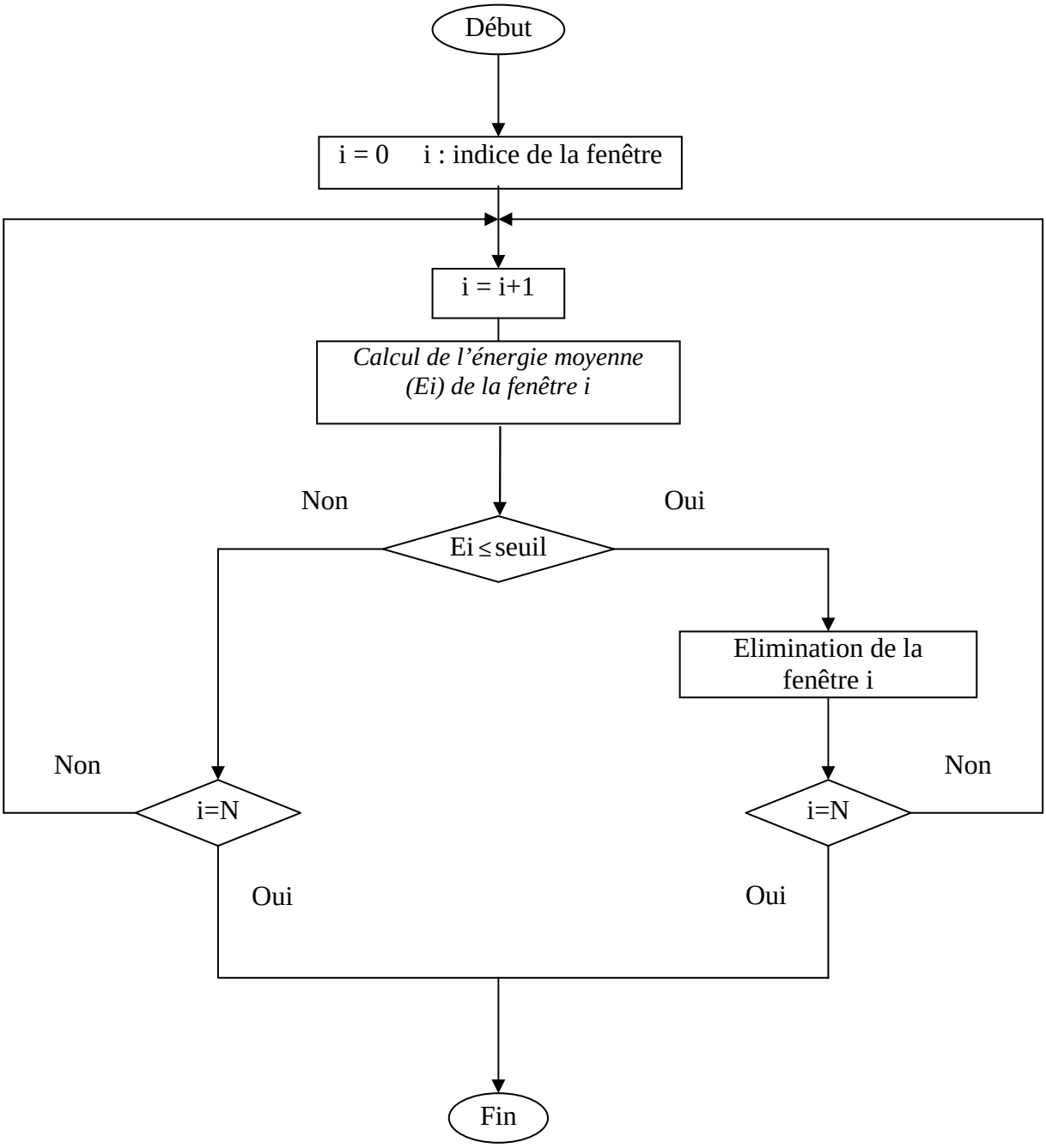


Figure 3.1 : Organigramme de détection et élimination des silences.
Remarque : dans notre cas, seuil = 72% × max(E) (résultat expérimental).

- Evaluation du spectre d'énergie par une transformée de Fourier rapide sur des trames d'analyse de 512 échantillons.

- Le spectre d'énergie passe ensuite à travers un banc de filtres constitué de plusieurs filtres du type Hamming (voir fig. 3.2).

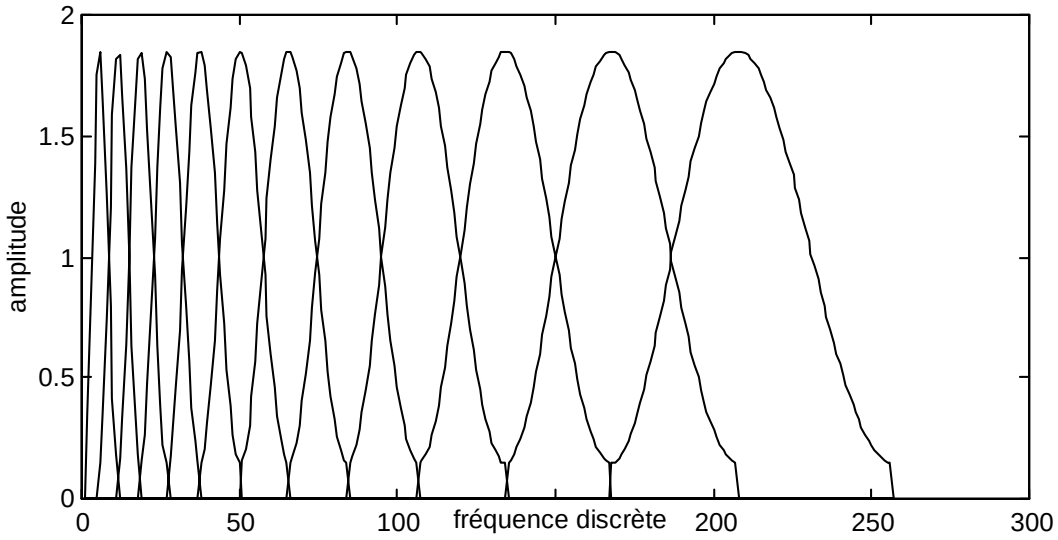


Fig. 3.2 : Banc de Filtres du type Hamming (dans ce cas : 12 filtres).

- La sortie de chaque filtre est calculée en sommant toutes les composantes fréquentielles à l'intérieur de la bande passante du filtre, pondérées par les coefficients du filtre, comme ce qui suit :

$$Y_k(m) = \sum_i c_{ik} Y^i(m), \tag{3.32}$$

- où $Y_k(m)$ est la sortie du k^e canal du banc de filtres pour la m^e fenêtre. $Y^i(m)$ est l'énergie du signal (carré du module du spectre) dans le i^e coefficient de Fourier pour la m^e fenêtre. c_{ik} est le gain du k^e filtre dans le i^e échantillon fréquentiel et i est sommé sur la bande passante du filtre.

Un écrêtage à seuil d'énergie [Dau 83] est utilisé pour limiter le rang dynamique du signal de parole. A la sortie de l'écrêtage on a

$$\tilde{S}_k(m) = \max[S_k(m), T^*_k], \tag{3.33}$$

$$\text{où } T^*_k = \max_m[S_{k(m)}] - 30\text{dB} \tag{3.34}$$

et où S_k représente la sortie Y_k dans l'échelle logarithmique (en dB).

Les canaux à faible énergie en sortie sont ainsi éliminés car ils correspondent souvent à l'effet du bruit.

Une normalisation de ces coefficients par leur énergie moyenne est effectuée, parce que le niveau d'énergie de parole varie d'une prononciation à l'autre et ceci est indésirable dans le cas de la reconnaissance [Lee 95].
Ceci se fait, d'abord, en calculant l'énergie moyenne $\overline{S}(m)$ de chaque trame sur les K canaux par :

$$\overline{S}(m) = \frac{1}{K}\sum_{k=1}^{K} \widetilde{S}_k(m), \tag{3.35}$$

La normalisation d'énergie est donnée, alors, par

$$S^n{}_k(m) = \widetilde{S}_k(m) - \overline{S}(m); \tag{3.36}$$

- Finalement les coefficients $S^n(m)$ pour chaque trame “m” sont stockés dans des vecteurs de dimension K, appelés les MFSC ou *Mel Frequency Spectral Coefficients*. Ce sont les caractéristiques du locuteur utilisées dans cette étude (figure 3.3).

Nous faisons remarquer que le signal de parole ne subit pas de préaccentuation.

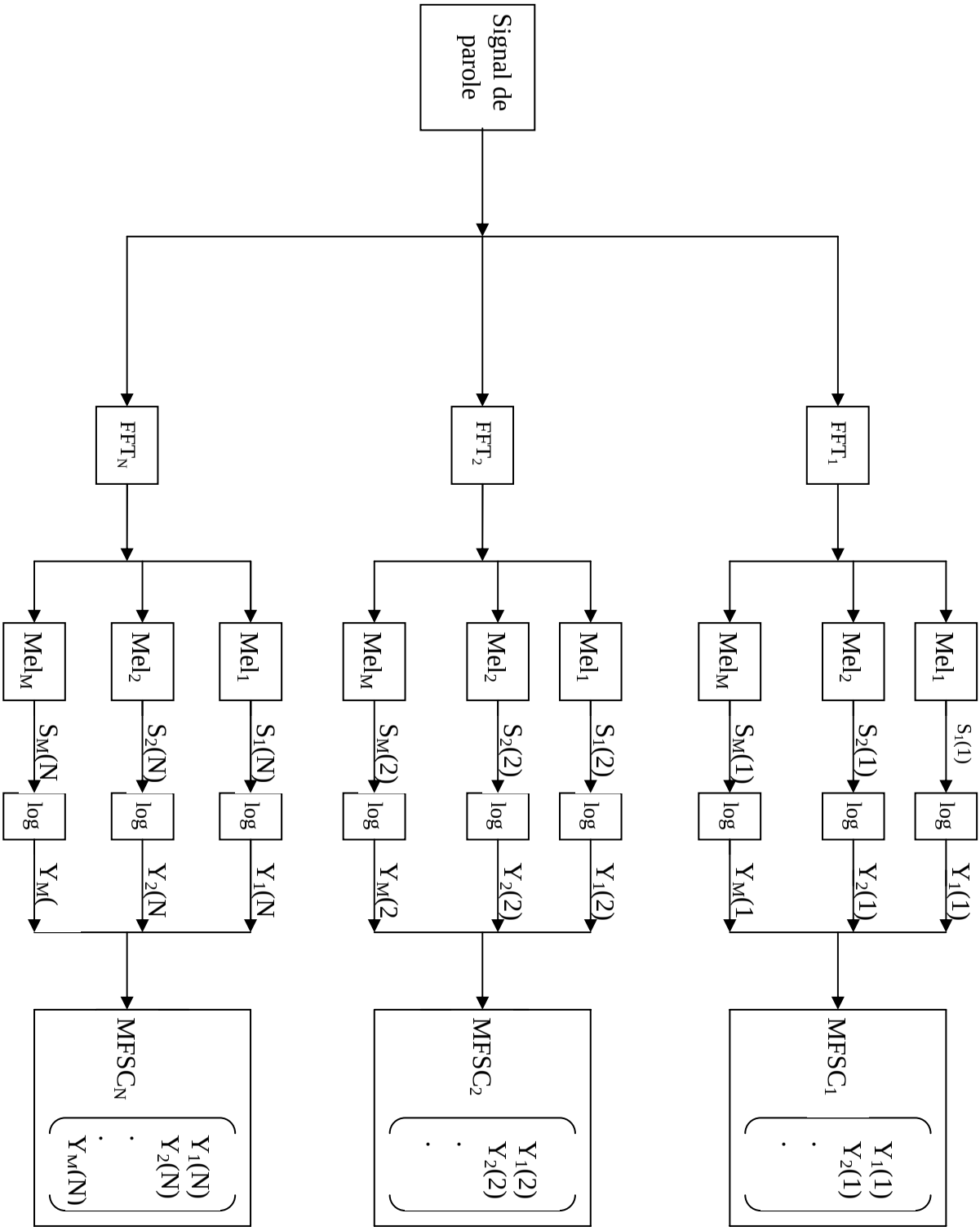


Figure 3.3 : Principe d'extraction des MFSC.

III.4. Apprentissage et Test

III.4.1. 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 l'ensemble des références (voir figure 3.4).

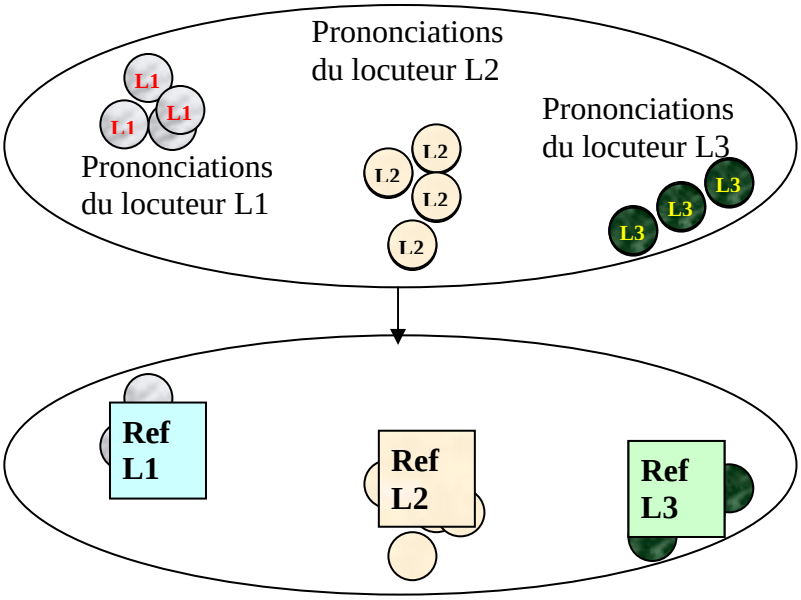


Figure 3.4 Construction des références : chaque ensemble de phrases (prononcées par le même locuteur) est représenté par une référence correspondant à ce locuteur.

III.4.2. 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. Ceux-ci seront comparés à ceux de l'ensemble des références afin d'identifier le locuteur en question.

La règle du plus proche voisin est ici appliquée : Ainsi on recherche la plus petite distance entre le locuteur inconnu et les différentes références (figure 3.5). La distance utilisée est la µG dans ses différentes variantes.

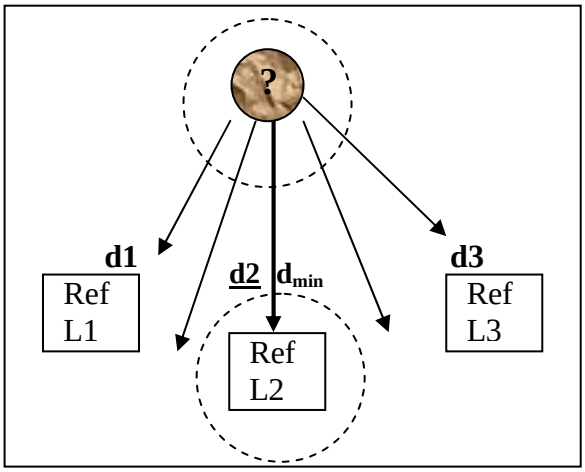


Figure 3.5 : Recherche de la distance minimale.

Les phases d'apprentissage et de test en identification du locuteur (par la méthode statistique) sont résumées dans les figures 3.6 et 3.7 (respectivement).

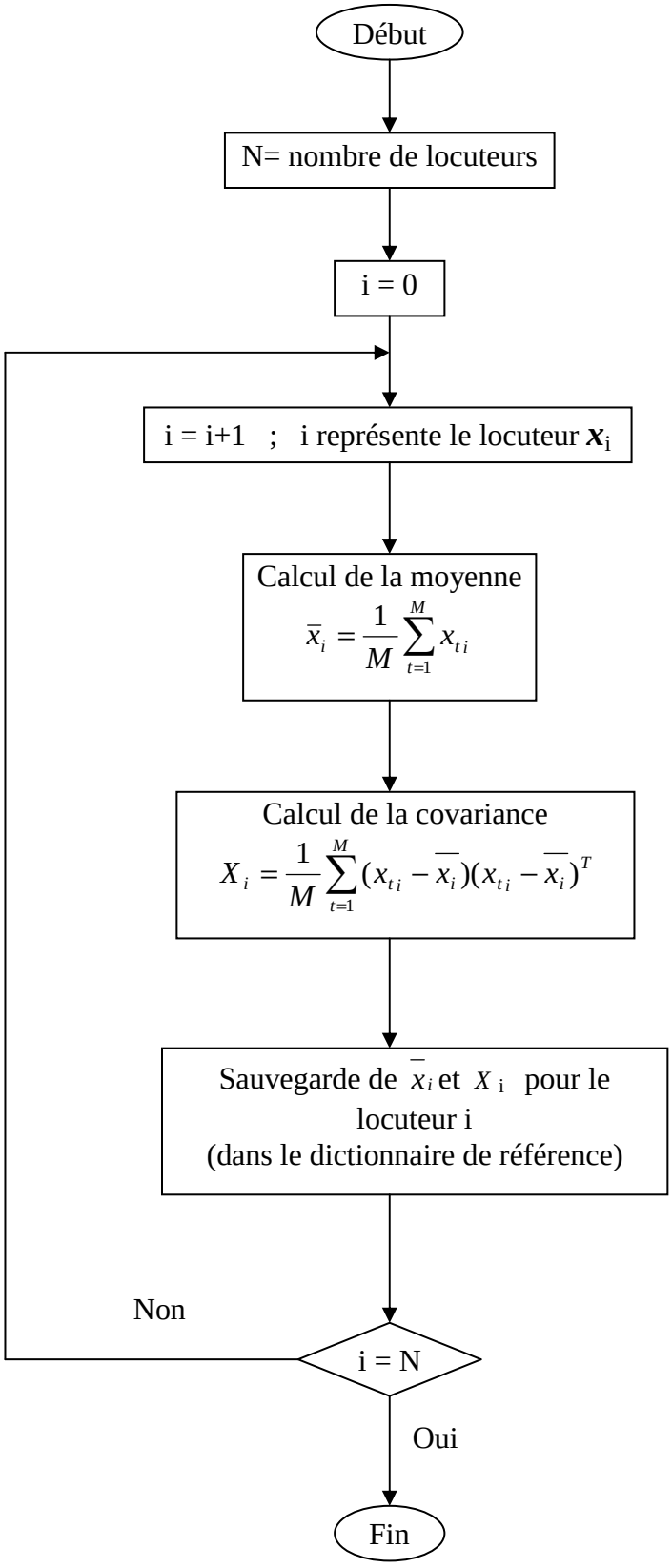


Figure 3.6 : Organigramme d'apprentissage.

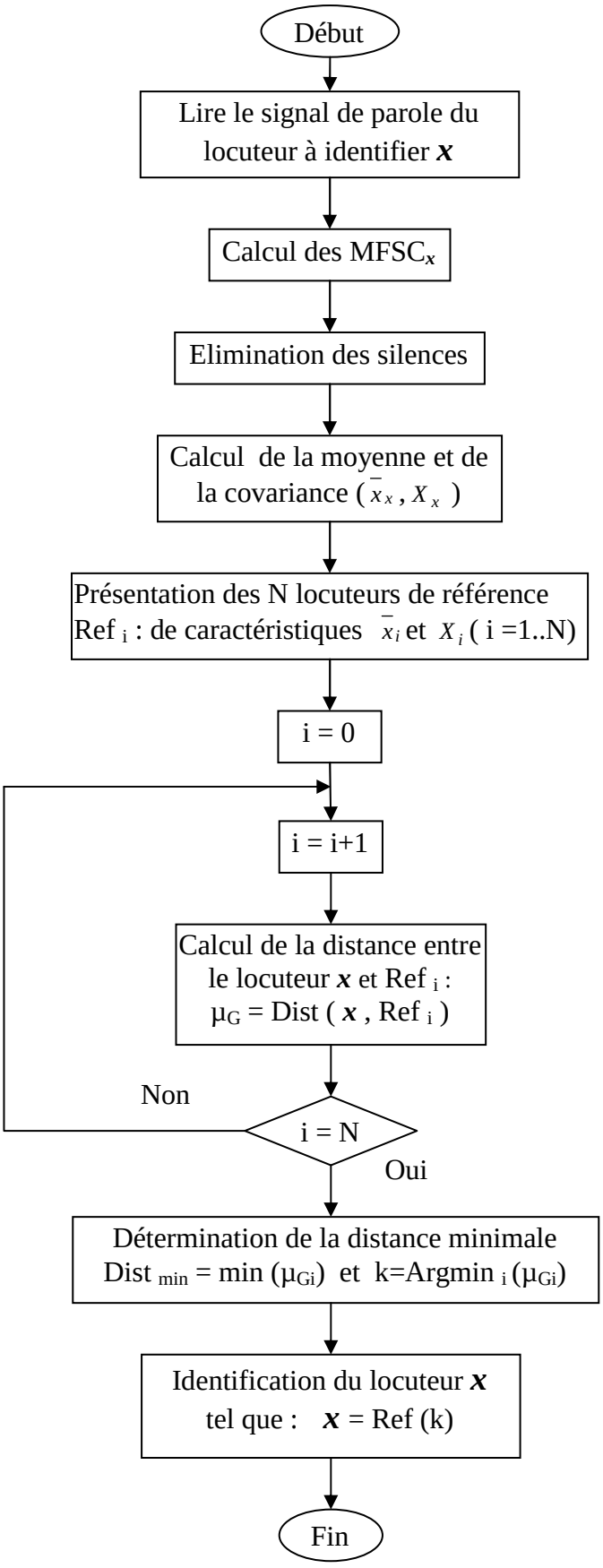


Figure 3.7 : Organigramme d'identification du locuteur dans un ensemble fermé

- Chapitre IV -

- Approche Connexionniste -

N.B. *Dans cette partie, on a supposé pouvoir simuler le fonctionnement du cerveau humain, mais en réalité on est bien loin, encore, pour comprendre le magnifique fonctionnement de ce dernier…*

IV.1. Introduction

L'étude connexionniste consiste à trouver et tester des modèles de réseaux de neurones fiables et bien adaptés à la tâche d'identification et de vérification du locuteur. Ici, on s'est inspiré du fait que le cerveau humain est capable de reconnaître aisément un personnage familier dès qu'il l'entend parler.
Une importante question se pose alors : est-il possible de comprendre et simuler son fonctionnement neuronal ?

Depuis plusieurs années, les neurophysiologistes ont tenté de mettre au point un modèle mathématique de neurones qui simule assez bien le cerveau, pour mieux comprendre certains fonctionnements du système nerveux.
Du modèle à la réalisation de machines pouvant être soumises à l'expérience (pour en tester la validité et le comportement), le cheminement de la pensée actuelle des chercheurs est simple à comprendre. Les machines ainsi conçues sont des architectures parallèles capables d'apprendre à prononcer assez correctement un texte, en dépit des pièges de la phonétique, de reconnaître un locuteur malgré son état (psychique, émotionnel …) différent d'un moment à l'autre.

Les réseaux de neurones ne sont pas des dispositifs biologiques mais bien 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.

Dans ce qui suit, nous donnerons quelques notions de base, nécessaires à l'utilisation de ces derniers en RAL.

IV.2. Notions de base sur les réseaux de neurones

IV.2.1. Aspect biologique

IV.2.1.1. Le neurone

Les cellules nerveuses, appelées neurones, sont les éléments de base du système nerveux central. Celui-ci en posséderait environ cent milliards [Dav 90].
Les neurones possèdent de nombreux points communs dans leur organisation générale et leur système biochimique avec les autres cellules. Ils présentent cependant des caractéristiques qui leur sont propres et qui se retrouvent au niveau des cinq fonctions spécialisées qu'ils assurent :

– Recevoir des signaux en provenance de neurones voisins.

– Intégrer ces signaux.

– Engendrer un influx nerveux.

– Le conduire.

– Le transmettre à un autre neurone capable de le recevoir.

IV.2.1.2. Structure des neurones

Un neurone est constitué de trois parties : le corps cellulaire, les dendrites et l'axone (fig. 4.1).

Le corps cellulaire :

Il contient le noyau du neurone et effectue les transformations biochimiques nécessaires à la synthèse des enzymes et des autres molécules qui assurent la vie du neurone. Sa forme est pyramidale ou sphérique dans la plupart des cas.

Les dendrites :

Chaque neurone possède une '' chevelure ''de dendrites. Celles-ci sont de fines extensions tubulaires. Elles se ramifient, ce qui les amène à former une espèce d'arborescence autour du corps cellulaire. Elles sont les récepteurs principaux du neurone pour capter les signaux qui lui parviennent.

L'axone :

C'est la fibre nerveuse ; il sert de moyen de transport pour les signaux émis par le neurone. Il se distingue des dendrites par sa forme et sa longueur en se ramifiant à son extrémité, là où il communique avec les autres neurones [Dav 90].

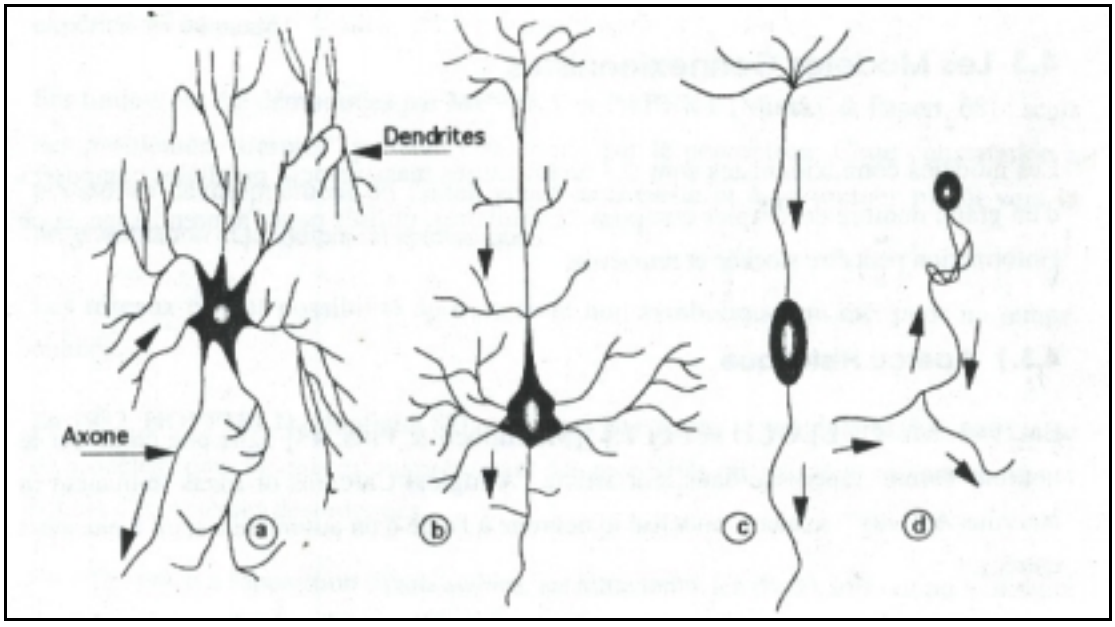


Figure 4.1 : Quelques silhouettes de neurones
Neurones multipolaires : a) motoneurone de la corne antérieure de la moelle
b) cellule pyramidale du cerveau.
Neurones bipolaires : c) cellule olfactive. d) neurone en T du ganglion spinal

IV.2.1.3. Fonctionnement

Quand un neurone a reçu suffisamment de signaux lancés par des milliers d'autres neurones à travers les synapses (figure 4.2), il envoie à son tour une décharge d'énergie vers les autres neurones qui lui son connectés par l'intermédiaire de son axone tel que le montre la figure 4.3.

En ce sens, le neurone ressemble à une batterie à seuil. Quand la batterie atteint une charge donnée, elle délivre une impulsion de courant et revient à une charge plus faible. Dans cette description du fonctionnement du neurone, la fonction d'activation est définie comme étant une somme pondérée des signaux reçus par chacune des dendrites.

En fait, seules quelques synapses sont prépondérantes dans l'activation des neurones. La synapse est la zone de contact entre deux neurones (ou entre un neurone et une cellule). C'est elle qui permet aux signaux électriques de circuler d'un neurone à l'autre, d'un neurone à un muscle…Chaque axone est pourvu de branches terminées par des "boutons synaptiques" par l'intermédiaire desquels le neurone se lie à un autre neurone ou à un muscle.

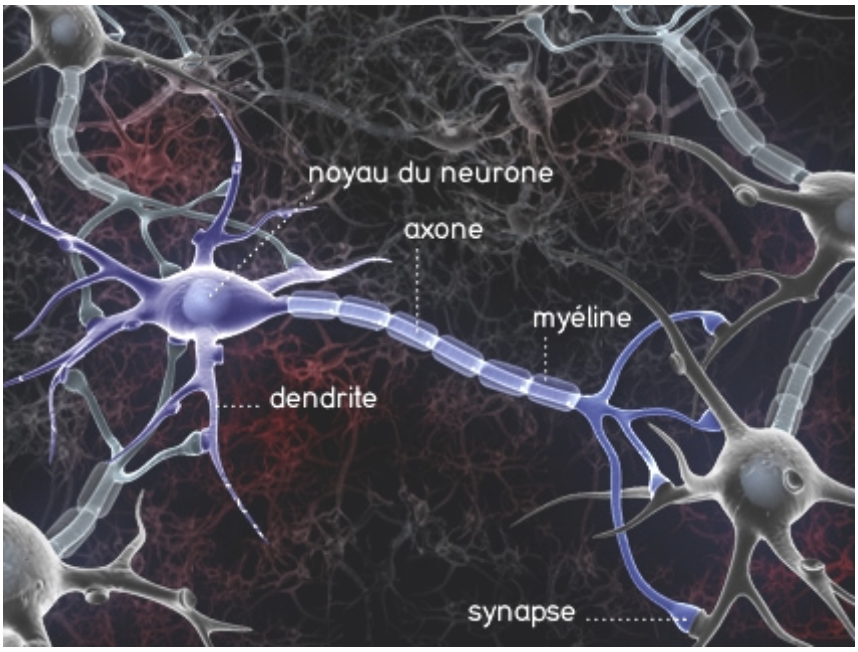


Figure 4.2 : Communication neuronale à l'aide des synapses

Le neurone est capable de s'en apercevoir et de modifier la fonction d'activation pour favoriser d'avantage les synapses qui l'ont stimulé. En ce sens, le neurone est capable d'apprendre : il reconnaîtra de plus en plus vite les signaux d'entrées qui génèrent l'impulsion de sortie. C'est de cette modélisation que s'inspirent les réalisations des neurones artificiels.

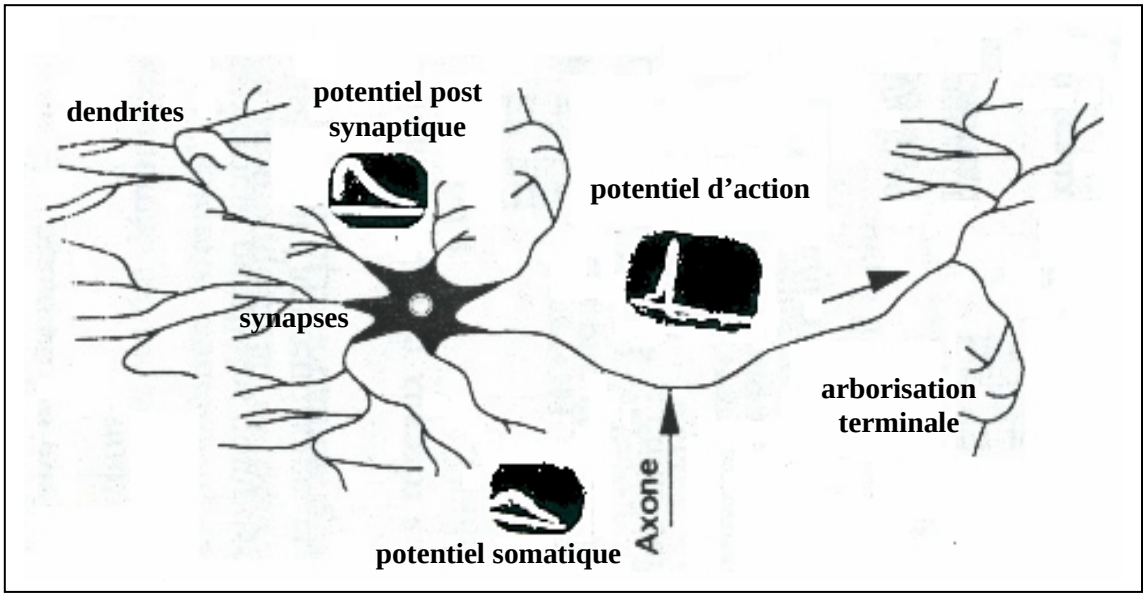


Figure 4.3 : Connexion entre neurones

IV.2.2. Modélisation

IV.2.2.1. Aperçu historique

L'observation des neurones biologiques a donné naissance aux réseaux de neurones artificiels. Ceux-ci visent à modéliser certaines structures de base du cerveau pour en reproduire les mécanismes.

Ainsi, en 1943, Mc Culloch et Pitts dans une étude sur le neurone formel avaient modélisé le neurone à l'aide d'un automate à seuil à plusieurs entrées. Ce neurone formel est l'unité de base des modèles connexionnistes.

En 1957, Rosenblatt propose un prototype de machine adaptative : le perceptron. Il s'agit d'une machine possédant trois couches d'unités, qui effectuent simultanément des traitements ; c'est donc une machine parallèle dont certaines connexions sont variables, ce qui la dote de plus d'une capacité d'apprentissage. Le perceptron a fourni la preuve qu'une machine, en modifiant sa structure, peut constituer une connaissance implicite des expériences du passé. Mais en 1969 Minsky et Papert démontrèrent les limites du perceptron et les recherches sur les réseaux de neurones artificiels furent abandonnées.

Cependant, en 1982 Hopfield relance l'intérêt pour cette approche. En effet, grâce à l'apparition d'états stables, les physiciens ont pu simuler une loi mathématique pour le comportement des systèmes dynamiques réels : le *modèle réseau*.

C'est à partir de ces travaux que le modèle réseau a pris toute son importance et qu'ont pu être réalisés les progrès les plus spectaculaires enregistrés ces dernières années dans le domaine [Ben 92], [Dav 90].

IV.2.2.2. La structure du modèle général

Les réseaux connexionnistes peuvent être définis comme un ensemble d'unités (ou automates) réalisant des calculs élémentaires, structurées en couches successives capables d'échanger des informations au moyen des connexions qui les relient.

Chaque unité effectue un traitement local d'information. Il existe plusieurs types d'unité :

Des unités d'entrée : auxquelles sont transmises les données à traiter, en provenance de sources externes au réseau.

Des unités de sortie : qui contiennent l'information traitée utilisable par d'autres systèmes connectés au réseau.

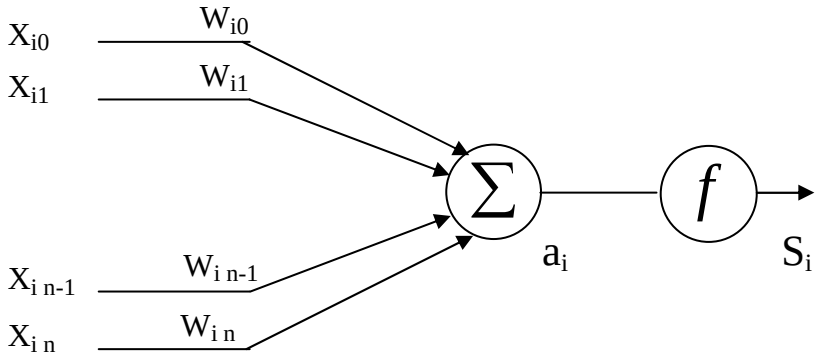


Figure 4.4 : Exemple d'unité neuronale

X_j : entrée provenant du neurone j.

W_{ij} : poids de la connexion de j vers i.

Des unités cachées : dont les entrées et les sorties sont reliées aux autres unités du réseau, et sont donc non ''visible'' par des systèmes extérieurs. Ces unités cachées servent à coder, de façon interne au système, la structure des formes présentées à l'entrée [Ben 92].

Ce type de neurone calcule d'abord une fonction d'entrée :

$$a_i = \sum_j Wij * Xj \qquad (4.1)$$

Puis son activité est donnée par :

$$S_i = f(a_i) = f\left(\sum_j Wij * Xj\right) \qquad (4..2)$$

Où f peut être du type (voir figure 4.5) :

- Fonction identité (purelin)
- Fonction à seuil (hardlim)
- Fonction sigmoïde (logsig)

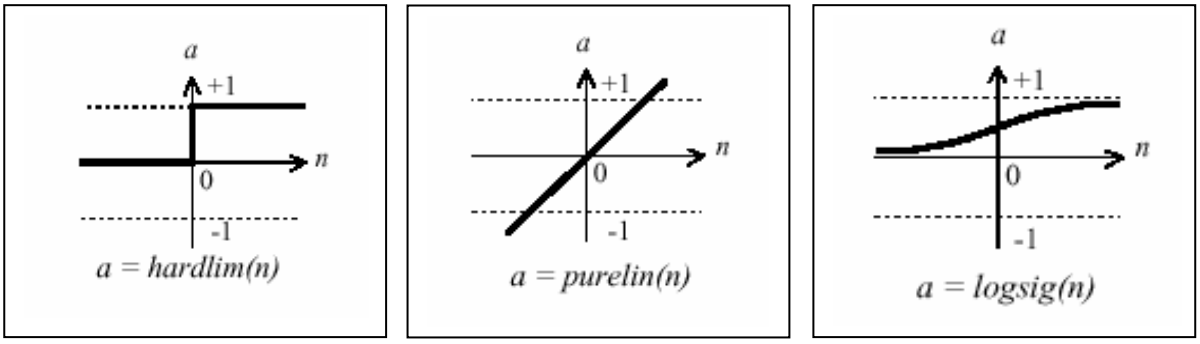


Figure 4.5: Quelques types de fonctions de transfert

La fonction sigmoïde possède les particularités suivantes [Mas 92] :

- Elle est continue sur l'ensemble des valeurs réelles.
- Sa dérivée est toujours positive.
- Elle est bornée : les intervalles des valeurs de sortie sont généralement] 0,1[ou]-1,1[.

IV.3. Construction des réseaux de neurones

Un réseau de neurone peut être représenté par un graphe direct composé d'un ensemble de nœuds ou éléments processeurs, fortement interconnectés par des liens orientés ou connexions.

Les réseaux de neurones peuvent se distinguer par : l'architecture et le mode d'apprentissage.

IV.3.1. L'architecture

IV.3.1.1. Les réseaux monocouche

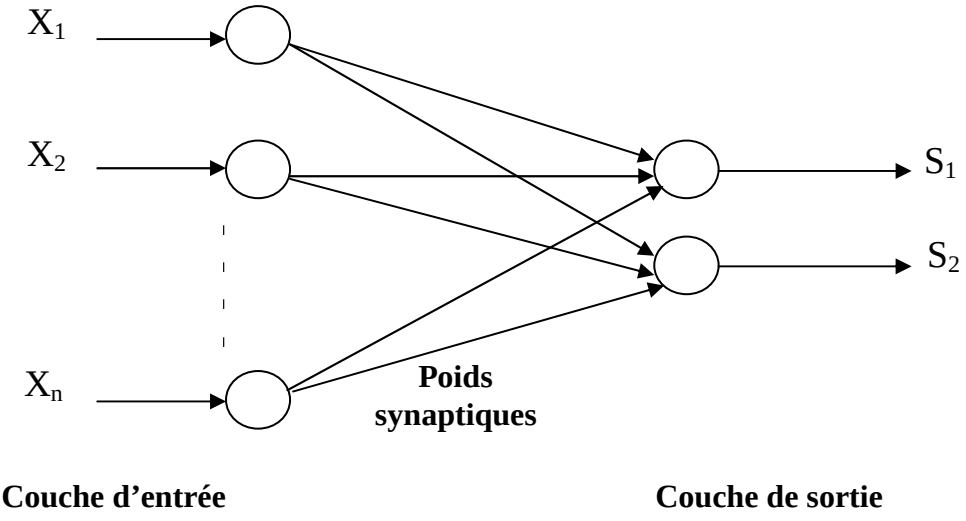


Figure 4.6: Architecture d'un réseau monocouche à deux sorties

L'ensemble des unités d'entrée est connecté à l'ensemble des unités de sortie par une couche de connexions modifiables. Cette architecture ne comporte pas de cellules cachées tel que donné par la figure ci-dessus.

IV.3.1.2. Les réseaux multicouche

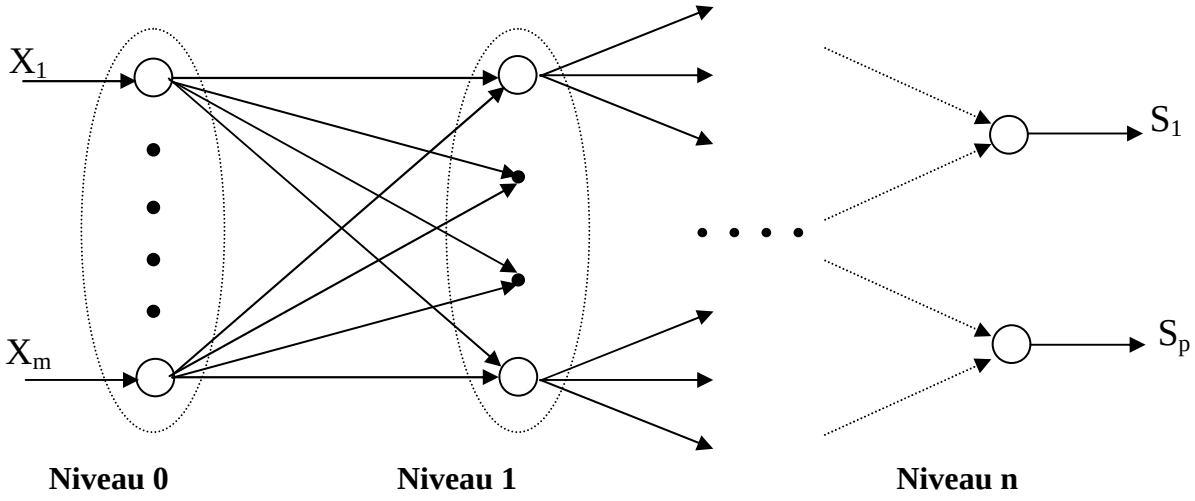


Figure 4.7 : Architecture d'un réseau multicouche à (n+1) niveaux

Un réseau multicouche est un réseau de neurones structuré en cascade de groupes de neurones (ou couches). Chaque couche est connectée à la suivante comme le montre la figure 4.7.

Cependant, il existe plusieurs façons de relier les unités d'un réseau, la plus simple est la connectivité totale : chaque unité est connectée à toutes les unités de la couche suivante.

Les unités de la même couche ne sont pas connectées entre elles. La première couche s'appelle alors couche d'entrée, la dernière : couche de sortie et les autres sont dites couches cachées.

IV.3.2. Mode d'apprentissage

D'habitude, l'apprentissage naturel est l'ensemble des méthodes permettant d'établir chez les êtres vivants des connexions entre certains stimulus et certaines réponses. C'est dans cette optique que l'on parle d'apprentissage neuronal.

Le choix d'un ensemble de poids synaptiques est un problème difficile en raison de leur nombre et de la complexité de leur comportement global dans le réseau. Pour cela, la majorité des réseaux neuromimétiques disposent d'une règle d'apprentissage qui leur permet d'adapter leurs poids automatiquement. En générale, l'apprentissage se fait sur une période relativement longue, durant laquelle les vecteurs d'entrées peuvent être représentés au réseau un grand nombre de fois chacun.

Il existe deux types classiques d'apprentissage : l'apprentissage supervisé et l'apprentissage non supervisé.

IV.3.2.1. Apprentissage supervisé

Le but de l'apprentissage supervisé est d'inculper un comportement de référence au réseau. On suppose alors qu'à chaque patron d'entrée est associée une sortie désirée qui spécifie les valeurs de sortie. L'apprentissage se déroule de la façon suivante :

Un patron est présenté aux neurones d'entrée, puis l'activation est propagée à travers le réseau, et enfin, la réponse des neurones de sortie est alors comparée aux valeurs désirées. Ceci détermine l'erreur du réseau pour le patron donné. Il s'agit alors de répartir cette erreur à chaque poids du réseau en fonction de la part qu'il a joué dans la production de la réponse erronée, on procède alors à une modification des poids qui vise à réduire l'erreur ainsi calculée.

L'ensemble d'apprentissage est donc constitué de paires de vecteurs {(Xi,Yi)}où : i=1,..N

N : représente le nombre d'exemples d'apprentissage.

Xi : représente le vecteur d'entrée.

Yi : représente le vecteur de la sortie désirée.

La procédure d'apprentissage consiste à :

- Appliquer le vecteur d'entrée
- Calculer le vecteur de sortie
- Comparer le vecteur de sortie au vecteur de la sortie désirée. La différence qui est l'erreur, est répliquée au réseau en modifiant les poids suivant un algorithme qui tend à minimiser cette erreur [Fre 92].

IV.3.2.2. Apprentissage non supervisé

Parallèlement et de façon complémentaire au développement des techniques d'apprentissage supervisé, ont été développées des techniques dites non supervisées. Elles visent à faire apprendre certaines informations sur des données non étiquetées. Il existe de nombreux cas où on ne possède pas d'information sur les classes de l'ensemble d'apprentissage. Ce manque de connaissance peut avoir plusieurs causes : manque d'information sur les données, volume d'information trop important pour pouvoir être étiqueté à la main [Ben 92], etc.

L'apprentissage non supervisé est également appelé apprentissage sans professeur. Dans cette catégorie d'apprentissage, la règle d'entraînement n'est pas fonction du comportement de sortie du réseau, mais plutôt du comportement local des neurones [Dav 90].

IV.4. Modèles de réseaux

IV.4.1. Le Perceptron

Un perceptron est un réseau de cellules composé de plusieurs modules, disposés 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 :

$$Y_j(X) = H\left[\sum_i W_i f_i(X)\right] \qquad (4.3)$$

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 $C_1, C_2,\ldots, C_p$.

L'apprentissage du perceptron :

L'apprentissage est supervisé et permet l'adaptation des poids des connexions de la couche de sortie.

Considérons, alors, un problème de classification de p modèles dans un espace de représentation égal à R^n.

Soit $E = \{(Xi, Yi)\}$ l'ensemble d'apprentissage où :

$Xi \in R^n$ vecteur d'entrée.

$Yi \in \{1, -1\}^p$ vecteur de sortie désirée.

Par convention si Xi appartient à la classe ou modèle j, tel que $1 \leq j \leq p$, alors on souhaite lui associer un vecteur de sortie $Yi = (Y_{i1}, Y_{i2},\ldots, Y_{ip})^t$ tel que :

$$Y_{ik} = \begin{cases} +1 \ldots\ldots\ldots si \;\; k = j \\ -1 \ldots\ldots\ldots si \;\; k \neq j \end{cases}$$

L'apprentissage est effectué selon les étapes suivantes :

Etape 1 : Initialiser aléatoirement les poids des connexions de la couche de sortie.

Etape 2 : choisir aléatoirement un couple (Xi, Yi) de l'ensemble E.

Présenter Xi en entrée sur la rétine, puis calculer la sortie du réseau à l'instant t.

$Y(Xi, t) = (Y_1(Xi), Y_2(Xi),\ldots, Y_p(Xi))^t$

Etape 3 : Pour tout automate K, pour lequel la sortie effective est différente de la sortie désirée ($Y_k(Xi, t) \neq Y_{ik}$) modifier les poids par la formule :

$$W_k(t+1) = W_k(t) + \varepsilon * \delta_{ik} * f_k(Xi) \qquad (4.4)$$

où : ε = constante positive.

$\delta_{i\,k} = Y_{i\,k} - Y_k(Xi, t)$ = signal d'erreur sur le $k^{ième}$ automate de sortie pour l'entrée Xi à l'étape t.

$f_k(Xi)$: sortie de la $k^{ième}$ cellule associative.

Etape 4 : Incrémenter t : t = t+1

Si condition d'arrêt non remplie, aller à l'étape 2.

Sinon fin.

IV.4.2. 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].

IV.4.3. L'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.

IV.4.3.1. Modèle du réseau à RPG

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 [Ben 92] :

$$f(x) = \frac{1}{1+e^{-x}} \qquad (4.5)$$

..

Remarque : le système neuronal artificiel ainsi constitue est appelé BPN (Back Propagation Network).

IV.4.3.2. Apprentissage

L'apprentissage des réseaux à 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 précédente 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 rétropropagé 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. [Fre 92], [Dav 90], [Hér 94].

IV.4.3.3. Formalisation

Les étapes de la règle de retro-propagation du gradient sont données en annexe (à la fin de cette thèse).

IV.4.3.4. Aspects pratiques de l'algorithme

IV.4.3.4.1. Choix de l'architecture du Réseau

La question qui se pose ici est : combien doit-il y'avoir de couches cachées et combien doit-il y'avoir de neurones sur les couches cachées ?

L'idée générale pour dimensionner correctement un réseau en fonction du problème à résoudre, est de se fixer un maximum de cellules dans la couches cachée. Si le réseau ne converge pas vers une solution satisfaisante, il est possible, alors, que plus de couches cachées soient nécessaires.

Ainsi, il est recommandé d'utiliser, au début, une seule couche cachée. Si cette dernière ne donne pas de résultats satisfaisants, il serait souhaitable d'essayer une seconde couche cachée, etc.

IV.4.3.4.2. Choix de l'ensemble d'apprentissage

Pour que le système soit efficace, pour une classe particulière, il faut que le réseau à RPG soit suffisamment entraîné sur les valeurs d'entrée de cette classe, sinon l'identification des membres de cette classe sera peu fiable. Il faut s'assurer donc que les données d'entraînement couvrent tout l'espace des entrées attendues.

IV.4.3.4.3. Valeurs initiales des poids

pour les poids :

Les poids sont initialisés à de petites valeurs aléatoires entre –0,5 et +0,5 [Fre 92].

pour les biais :

Le biais devra être traité comme tout autre poids qui est connecté à une unité, et qui aurait toujours une entrée égale à 1 [Fre 92].

IV.4.3.4.4. Le pas d'apprentissage

Le pas d'apprentissage doit être choisi avec soin car celui ci a un effet important sur la performance du réseau. Généralement, il doit être un petit nombre comprit entre 0.05 et 0.25 pour assurer que le réseau converge vers une solution donnée. Une petite valeur signifie que le réseau devra faire un grand nombre d'itérations avant de converger.

Il est possible de l'incrémenter durant l'apprentissage quand l'erreur décroît, et cela pour augmenter la vitesse de convergence, mais s'il devient trop grand, le réseau risquera de rebondir très loin du minimum.

IV.4.3.4.5. 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'attend jamais zéro. La méthode consiste, alors, à arrêter l'apprentissage dès que **Ep** devient inférieure à un seuil **Emin** donné. Notons que **Emin** ne doit pas être trop petit (un grand nombre d'itération).

IV.4.4. Architecture à poids partagés (TDNN)

L'idée de prendre en compte la notion du temps dans un réseau multicouche a été initialement introduite par Lang et Hinton en 1988 [Lan 88] sous le nom de TDNN (Time Delay Neural Network).

Les architectures TDNN ont été fréquemment utilisées en reconnaissance de la parole. Ce type de réseau est un cas particulier des MLP avec des connexions locales et des contraintes d'égalité entre les poids du réseau. Ce sont des architectures adaptées à la nature dynamique de la parole. Les cellules du TDNN apprennent à extraire du signal de parole les caractéristiques nécessaires pour résoudre la tache posée.

Ce type d'architecture effectue en fait un dépliage temporel de l'algorithme de rétropropagation qui s'applique directement sur le réseau déplié. Cependant, certaines connexions du réseau déplié correspondent à la même connexion du réseau temporel. Elles doivent donc conserver un poids identique. Il suffit pour cela de leur donner la même valeur initiale, et de les ajuster avec la somme des gradients de toutes les connexions partageant le même poids.

Le gain en temps de calcul n'est pas le principal avantage de ce type d'architecture : imposer aux cellules cachées des intervalles de constance plus longs leur permet de résister à de petites distorsions temporelles [Ben 92].

IV.4.5. Learning Vector Quantization (LVQ)

Les algorithmes de quantification vectorielle avec apprentissage ont été proposés par T. Kohonen [Koh 88]. Ils constituent une adaptation des 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].

Les algorithmes de quantification vectorielle sont des méthodes développées avant tout dans un but pratique et pouvant être mis en œuvre sur des problèmes réels. Ils 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. Ils permettent de réaliser des modules de décision efficaces.

Ces types de réseaux de neurones ont été conçus comme un cas particulier et une variation de l'algorithme des cartes topologiques [Koh 88]. De plus ils peuvent être intégrés dans des architectures de réseaux complexes à plusieurs modules. Il est donc intéressant de les présenter dans ce cadre.

IV.4.5.1. 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.

T. Kohonen a présenté deux variantes de l'algorithme, appelées LVQl et LVQ2 [Koh 84]. Elles diffèrent par la façon d'adapter, pendant la phase d'apprentissage, les vecteurs de référence. Les deux algorithmes sont initialisés par un apprentissage non supervisé type k-moyennes qui fournit une première partition de l'espace de départ. Ces algorithmes n'utilisent donc pas l'information de classe qui est disponible sur l'ensemble d'apprentissage. Cette information est utilisée ensuite dans une seconde étape où chaque groupe de cette partition est étiqueté par une classe, habituellement ce sera celle qui est majoritaire dans le groupe. L'apprentissage consiste ensuite à modifier cette partition en déplaçant les vecteurs de référence initiaux des différents groupes de façon à diminuer le taux de mauvaise classification par rapport à l'étiquetage précédent. Pour cela LVQl utilise toutes les formes de l'ensemble d'apprentissage alors que la LVQ2 n'utilise que les formes qui sont à la frontière de deux groupes. Les algorithmes sont itératifs et demandent plusieurs présentations successives de la base d'exemples, jusqu'à ce que ce critère d'arrêt soit satisfait. En général on s'arrêtera quand le taux de classification n'est plus amélioré.

Nous décrivons maintenant plus précisément ces deux algorithmes :

La première phase est identique pour les deux versions LVQl et LVQ2. On a donc, après cette étape d'initialisation, identifié un certain nombre de vecteurs de référence et étiqueté les groupes de formes correspondants par un indice de classe.

IV.4.5.1.1. 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.4.5.1.2. 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 car 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.

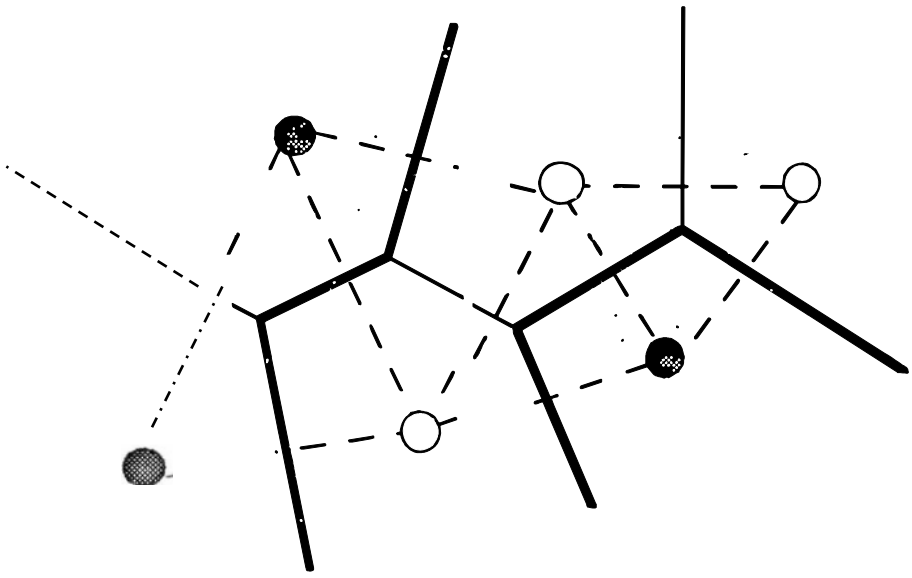


Figure 4.8 : Frontière de séparation entre deux classes dans le cas de la LVQ
Les traits en gras représentent des frontières entre des classes différentes.

L'algorithme consiste alors à éloigner la référence de la "mauvaise" classe et de rapprocher celle de la "bonne" classe (par rapport à l'exemple) dans le cas où l'exemple tombe à

l'intérieur de la fenêtre définie précédemment. La reconnaissance est identique à celle de LVQl.

IV.4.5.1.3. Les Algorithmes de la LVQ

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

Dans ce qui suit, chaque point représente un détail (ou étape) important dans la technique LVQ.

- 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.
  - Etiqueter 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.

Algorithme

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 quand le taux de classification devient satisfaisant.

Les deux algorithmes diffèrent uniquement par l'étape 3.

- **Dans LVQl elle devient**

Modifier le vecteur de référence $w^*(t)$ par:

$w^*(t+1) = w^*(t) + \alpha(t)\,(x - w^*(t))$ **si** C = C* (4.6)

(la référence et l'exemple sont de même classe)

et

$w^*(t+1) = w^*(t) - \alpha(t)\,(x - w^*(t))$ **si** C ≠ C* (4.7)

(la référence et l'exemple sont de classes différentes)

mais

Dans tous les autres cas, faire: $w^k(t+1) = w^k(t)$ (4.8)

$\alpha(t)$ 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:

$\alpha(t) = 0.1\ (1 - t/M)$ (4.9)

où *M* est le nombre maximum d'itérations après lequel l'apprentissage est arrêté.

- **Et dans LVQ2 elle devient :**

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, alors déplacer les vecteurs de référence w(t) et w**(t) par :*

$w^*(t+1) = w^*(t) - \alpha(t)\ (x - w^*(t))$ (4.10)

$w^{**}(t+1) = w^{**}(t) + \alpha(t)\ (x - w^{**}(t))$ (4.11)

mais

Dans tous les autres cas, faire : $w^k(t+1) = w^k(t)$ (4.12)

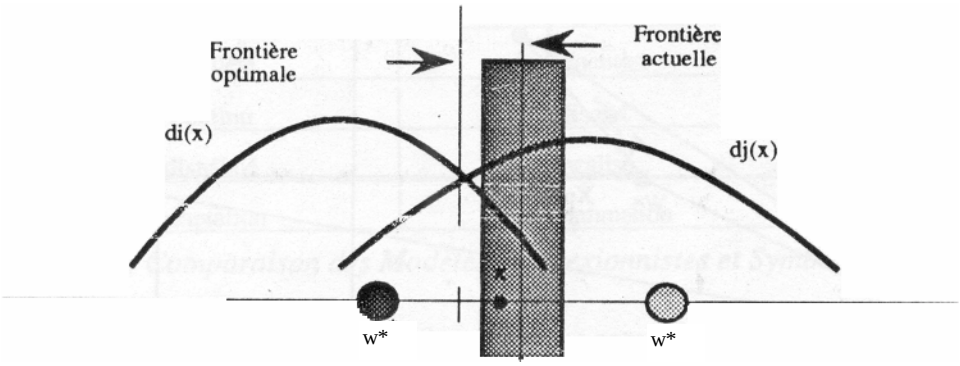


Figure 4.9 : Mécanisme d'adaptation dans le cas de LVQ2. La frontière initiale est déplacée jusqu'à optimisation du système.

Remarque : *Dans les équations précédentes, nous avons utilisé les notations suivantes.*

x : un vecteur de la base d'apprentissage

w et w**: vecteurs de référence*

di(x) et dj(x) : distributions des classes Ci et Cj

IV.4.5.1.4. Comment Définir la Fenêtre LVQ2 ?

Dans un espace de dimension n, pour tester si l'exemple x est dans la fenêtre LVQ2, on peut procéder de la façon suivante :

- *on calcule (x - w*) et (w** - w*)*
- *on calcule la projection de (x - w*) sur (w** - w*) : xp*
- *la valeur algébrique de xp est donnée par :*

$$xp = \frac{(x - w^*)(w^{**} - w^*)}{\|w^{**} - w^*\|} \qquad (4.13)$$

- *on considère une fenêtre de largeur* $L = \beta \|w^{**} - w^*\|$, *avec* $0 < \beta < 1$ (4.14)
- *x est dans la fenêtre L VQ2 si :*

$$\left| \frac{\|w^{**} - w^*\|}{2} - xp \right| \leq \beta \frac{\|w^{**} - w^*\|}{2} \qquad (4.15)$$

soit

$$\left| \frac{\|w^{**} - w^*\|}{2} - \frac{(x - w^*)(w^{**} - w^*)}{\|w^{**} - w^*\|} \right| \leq \beta \frac{\|w^{**} - w^*\|}{2} \qquad (4.16)$$

La figure 4.10 illustre cette méthode avec une représentation dans l'espace R^2.

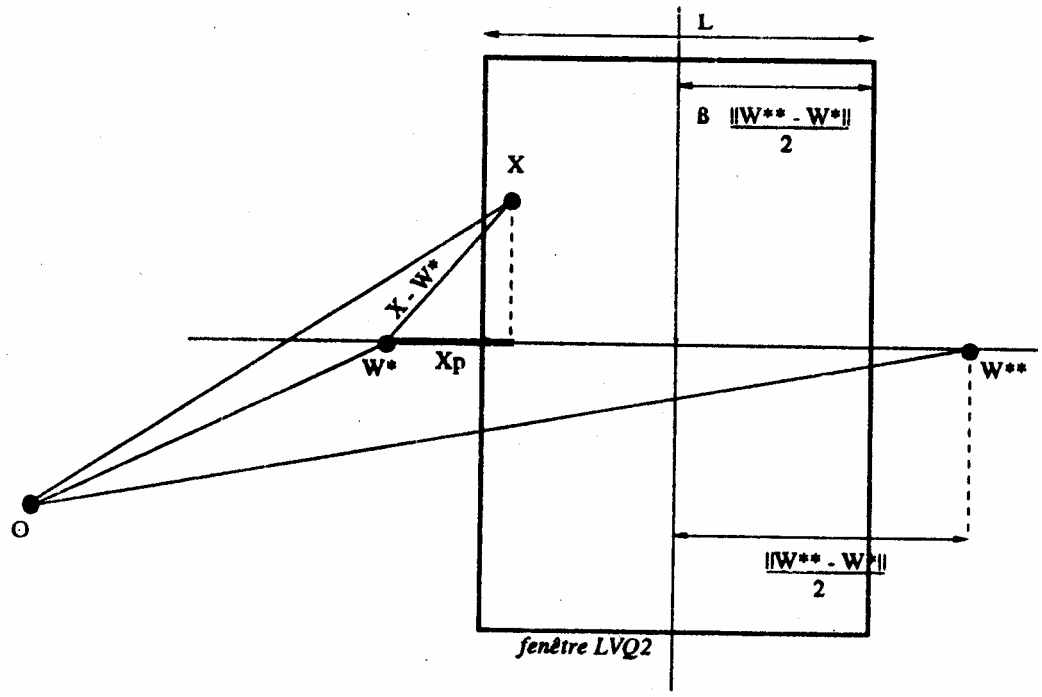


Figure 4.10 : Fenêtre d'adaptation de la LVQ2

IV.5. Conclusion

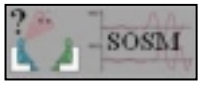
Après avoir présenté l'approche connexionniste en expliquant le mécanisme des réseaux de neurones et l'excellente possibilité de classification de ceux-ci, on peut conclure que l'intérêt porté aujourd'hui aux réseaux de neurones tient sa justification dans les propriétés suivantes :

- **La capacité d'adaptation :** elle permet au réseau de tenir compte de nouvelles contraintes (à conditions que celles-ci ne soient pas trop importantes).
- **La capacité de généralisation :** elle se traduit par la capacité à apprendre et à retrouver à partir d'un ensemble d'exemples une fonction qui permette d'assurer les liens entre ces exemples.
- **Le parallélisme :** cette notion se situe à la base de l'architecture des réseaux de neurones considérée comme un ensemble d'entités élémentaires qui travaillent simultanément, ce qui permet une grande rapidité de calcul.

En contrepartie, certains problèmes gênants se posent , comme par exemple :

- **La complexité :** ce qui les rend lourds.
- **Le temps d'apprentissage :** ce temps (qui est trop important) est nécessaire au bon apprentissage du système.
- **La non évolutivité :** les systèmes connexionnistes ne sont pas évolutifs si des modifications assez importantes sont imposées au système (nouvelles classes ajoutées).

En tout cas, le meilleur moyen d'estimer objectivement leurs performances est de les tester en pratique, dans des applications réelles et dans des conditions difficiles.



- Chapitre V -

Identification du Locuteur par la Méthode SOSM

- Recherche de la Résolution Spectrale Optimale -

N.B. *Ce chapitre concerne la première série d'expériences effectuées en RAL.*

V.1. Introduction

Ce chapitre décrit les expériences d'identification du locuteur, par la méthode SOSM, en exposant les différents résultats obtenus et en commentant ces résultats d'un point de vue objectif. Les tests de reconnaissance ont été faits en milieu sain et en milieu bruité. De plus, deux bandes spectrales ont été expérimentées : la bande large de 0-8 kHz et la bande téléphonique de 300-3400 Hz.

La méthode SOSM (*Second Order Statistical Measures)* est étudiée selon plusieurs critères :

- la résolution spectrale variant de 12 canaux à 60 canaux,
- le Rapport Signal sur Bruit (RSB) variant de 0 dB à son maximum correspondant à un environnement non bruité (ici trois types de bruit sont utilisés),
- 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).

Ces paramètres sont pris comme facteurs variables de l'environnement d'identification.

Les différentes expériences, montrent que cette méthode se comporte différemment selon le cas étudié. En plus des constats qualitatifs 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 0-8 kHz et celle de 48 canaux / 8 kHz 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 dans la synthèse générale de ce chapitre.

V.2. Résolution spectrale optimale

Une importante question pouvant se poser en reconnaissance (robuste) du locuteur est : Est-il possible d'améliorer la qualité de la reconnaissance du locuteur par un choix judicieux et optimal de la dimension des caractéristiques spectrales du locuteur ?

L'aspect physique de cette optimisation, dans notre étude, est lié à un compromis de paramétrisation :

Enveloppe formantique *(information basse fréquence)* **/** ***harmoniques*** *(information haute fréquence)*

Mais si le problème d'extraction de l'enveloppe formantique a été résolu par la paramétrisation cepstrale et Mel-cepstrale [Rey 94], il n'en est pas de même avec la difficulté rencontrée lors de l'extraction des informations de haute fréquence, en particulier si la parole est bruitée [Say 98] [Say 98b] ou filtrée (bande téléphonique).

Vu que la résolution spectrale à 12 ou 24 canaux est largement suffisante pour extraire l'information contenue dans l'enveloppe formantique et que lorsque nous utilisons des dimensions très grandes nous nous heurtons à deux problèmes : la complexité de calcul et la non inversibilité de la matrice de covariance, alors nous avons intérêt à réduire cette résolution. Cependant, vu que l'information contenue dans les hautes fréquences est complémentaire [Bon 97], il sera nécessaire de rajouter cette dernière par une élévation de la résolution spectrale d'où la nécessité de trouver un compromis optimal pour l'identification du locuteur.

Durant cette étude on a pu montrer l'importance de la haute résolution spectrale dans l'identification du locuteur en trouvant le nombre optimal de canaux (banc de filtres) à adopter en reconnaissance du locuteur par la méthode SOSM : les essais expérimentaux ont montré que la haute résolution spectrale permet de mieux distinguer des locuteurs de voix proches. Les résultats étayent effectivement cette hypothèse, avec une amélioration du taux de reconnaissance, par exemple, de 11% sur la bande téléphonique dès qu'on passe de 24 canaux à 48 canaux.

V.3. Base de Données des Locuteurs

La base de données des locuteurs est extraite de la base TIMIT [Fis 86] et FTIMIT [Mag 96] (FTIMIT correspond à TIMIT dont lequel on ne préserve que les coefficients de caractéristiques spectrales contenues dans la bande : 300-3400Hz). Le nombre de locuteurs est de 37, avec un accent anglais nord-américain ; 22 sont des locuteurs masculins et 15 sont féminins. La durée approximative des énoncés est de 9 s pour l'apprentissage et de 7 s pour le test. Les enregistrements sont faits avec un microphone de haute qualité, sur 16 bits et avec une fréquence d'échantillonnage de 16 kHz.

Trois types de bruits sont indépendamment utilisés (bruit blanc, bruit de voiture et bruit de chahut) et ajoutés à la base de données de l'apprentissage et de test à 5 différents RSBs.

Chaque base de données subit 5 fois le traitement global destiné à extraire les caractéristiques acoustiques et statistiques des locuteurs, car nous utilisons 5 résolutions spectrales différentes allant de 12 à 60 canaux dans le banc de filtres. Deux types de distances sont testés : la µG et la µGc. Cette différence implique 2 apprentissages différents et 2 tests différents. Au total la consistance du traitement global équivaut (en terme d'opérations de calcul) au traitement d'une base de données de **11.100** locuteurs.

V.4. Méthode SOSM

Dans cette méthode on fait l'hypothèse simplificatrice que les caractéristiques spectrales du locuteur suivent une distribution gaussienne stationnaire au second ordre [Bim 95].

Dans l'algorithme de la SOSM, on procède, tout d'abord, à l'extraction des MFSC (ou Mel énergies), ensuite pour chaque prononciation on fabrique le vecteur moyenne x et la matrice de covariance X ; ainsi il existe une moyenne x pour chaque locuteur et une covariance X pour chaque locuteur. Le couple (x,X) représente la référence statistique d'ordre 2 pour le locuteur ***'x'*** utilisé dans le dictionnaire des références.

En phase de test, une modélisation similaire de la phrase de test générera le couple (y,Y) représentant la modélisation statistique d'ordre 2 pour le locuteur inconnu ***'y'***.

Le test d'identification est basé, alors, sur la distance minimale (plus proche voisin) au sens de la métrique statistique d'ordre 2 appelée μ_G (généralement on utilise la $\mu_{G0.5}$). Voir figure 5.1.

Cette distance (définie au chapitre III) est donnée par :

$$\mu_{G0.5}(x,y) = \frac{1}{2}\left[a + \frac{1}{h} + \frac{1}{p}\delta^T \left(X^{-1} + Y^{-1} \right)\delta \right] - 1 \qquad (5.1)$$

avec :

$$a = \frac{1}{p} tr\left(YX^{-1} \right) \qquad (5.2)$$

$$h = \frac{p}{tr\left(XY^{-1} \right)} \qquad (5.3)$$

$$\delta = y - x \qquad (5.4)$$

p étant la dimension du vecteur des caractéristiques acoustiques et "tr" dénote la trace d'une matrice. ***x*** représente la référence et ***y*** représente le locuteur à reconnaître.

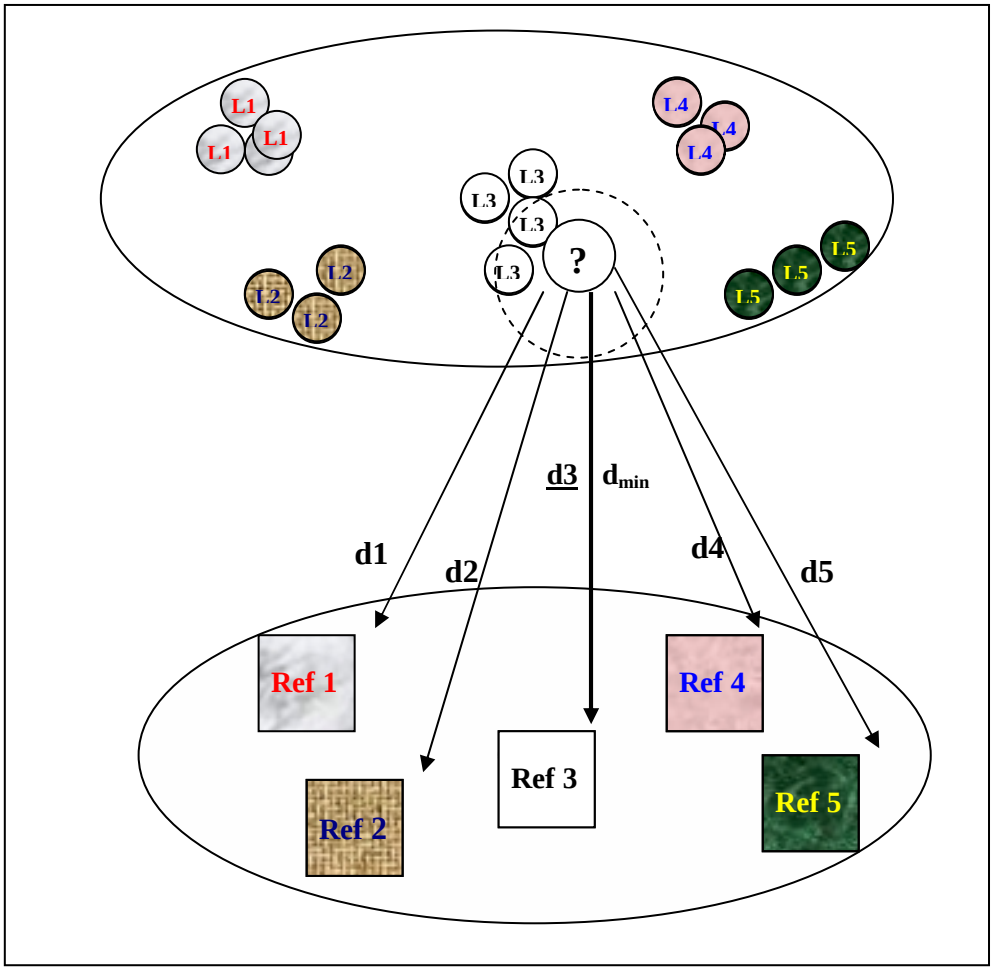


Figure 5.1 : Recherche de la distance minimale.

Ici le locuteur inconnu est identifié comme étant la référence du milieu (L3).

V.5. Distance minimale et seuillage de rejet (cas d'un ensemble ouvert)

Dans le cas où l'ensemble des locuteurs est ouvert ou dans le cas de la vérification [Dod 98], l'acceptation d'un locuteur nécessite que la distance minimale (μ_G) trouvée soit inférieure à un seuil de rejet représentant le maximum de distance autorisé ; ceci permet d'éviter les fausses identifications et les impostures [Dod 98] :

Si μ_G > Seuil, alors rejeter la décision d'identification.

Ce seuil (*"Seuil"*) peut être déterminé empiriquement, par les formules 5.5 ou 5.6.

$$Seuil = \frac{\min(Interdist) + \max(Intradist)}{2} \qquad (5.5)$$

La formule 10 est utilisée dans le cas d'un seuil commun à tous les locuteurs.

Cependant, une technique plus intéressante et plus performante consiste à calculer un seuil optimal pour chaque locuteur comme le montre la formule 5.6 suivante.

$$Seuil(i) = \frac{\min(Interdist(i)) + \max(Intradist(i))}{2} \qquad (5.6)$$

où *Interdist* représente les distances inter-locuteur sur la base de données d'apprentissage et *Intradist* représente les distances intra-locuteur sur la base de données d'apprentissage. De même, *Interdist(i)* représente les distances inter-locuteur par rapport au locuteur "i" et *Intradist(i)* représente les distances intra-locuteur par rapport au locuteur "i", toujours sur la base de données d'apprentissage.

La variable *"i"* symbolise un locuteur de l'ensemble des locuteurs de référence.

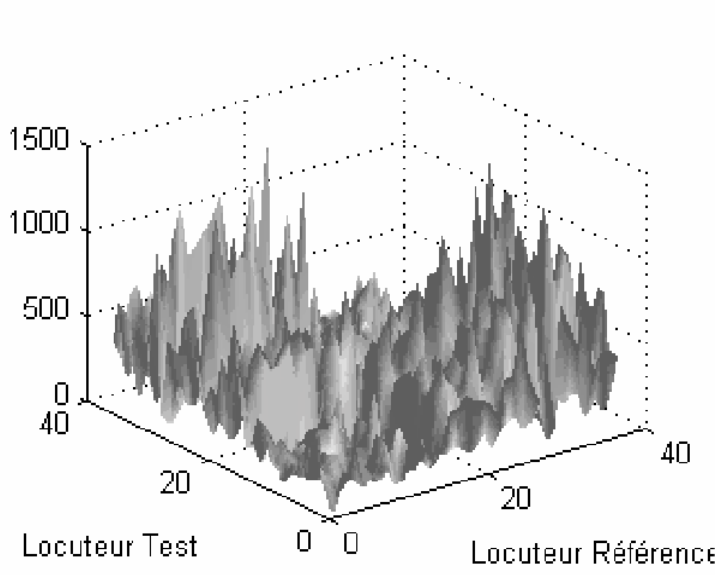


Fig. 5.2 : Distances inter-locuteurs trouvées lors du test de reconnaissance (l'axe vertical mesure les distances entre les deux ensembles de locuteurs).

V.6. Compromis "Basse Résolution / Haute Résolution Spectrale".

Une importante question pouvant se poser lors du choix de la dimension optimale des caractéristiques acoustiques du locuteur est la suivante : Existe-il un rapport entre la résolution spectrale (dimension des coefficients MFSC) et la fiabilité de modélisation des paramètres acoustiques essentiels du signal de parole ?

Pour ce faire, on commencera à étudier les spectres tridimensionnels obtenus expérimentalement avec les différentes résolutions spectrales afin de trouver un lien avec la fidélité de modélisation de l'enveloppe spectrale et la fidélité de modélisation des harmoniques. Voir les figures 5.3 à 5.7.

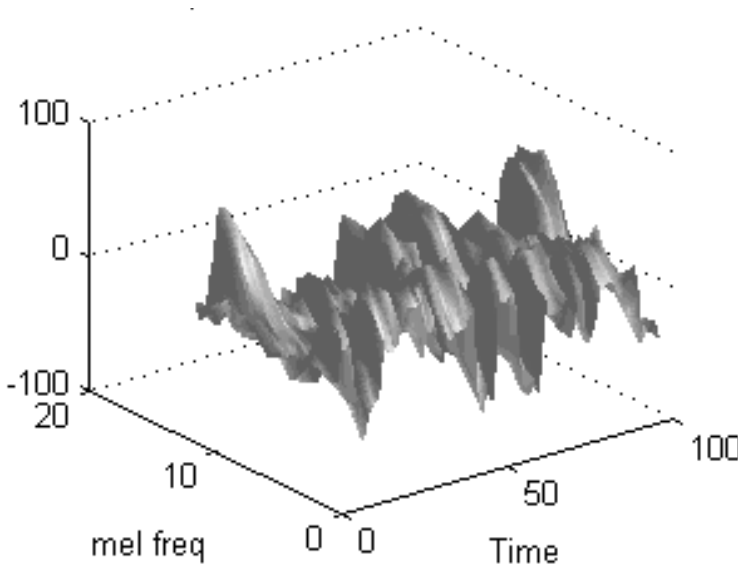


Fig. 5.3 : Représentation en 3D du spectre (échelle Mel) à l'aide de 12 canaux

❖ Sur la figure 5.3, soit à 12 MFSC, nous pouvons voir que nous ne préservons qu'une image grossière de l'enveloppe spectrale.

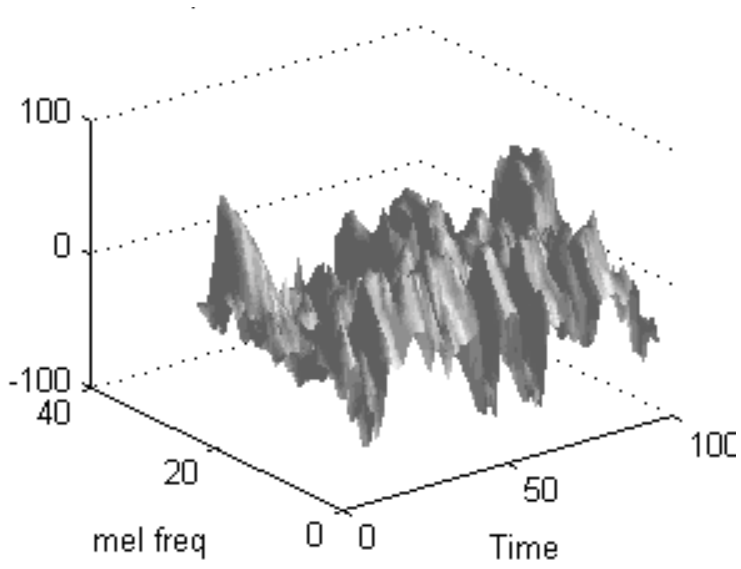


Fig. 5.4 : Représentation en 3D du spectre (échelle Mel) à l'aide de 24 canaux.

❖ Sur la figure 5.4, soit à 24 MFSC, nous modélisons très bien la structure de l'enveloppe spectrale mais aucun harmonique n'est encore bien visible.

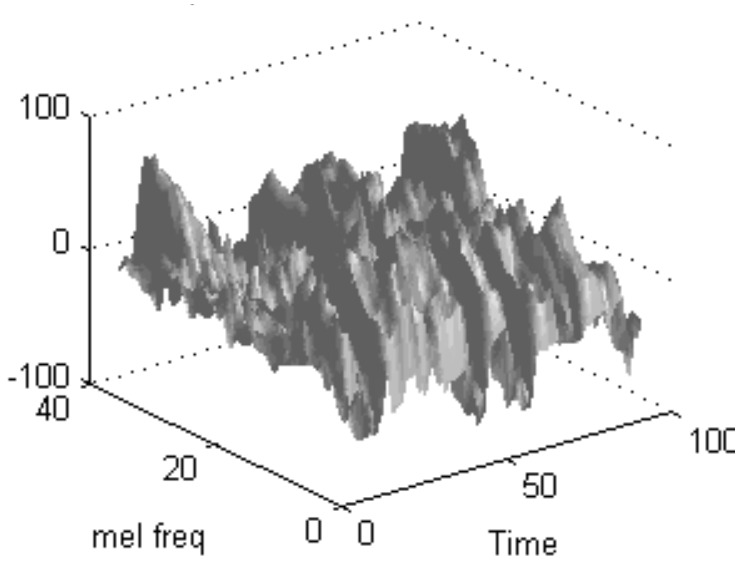


Fig. 5.5 : Représentation en 3D du spectre (échelle Mel) à l'aide de 36 canaux.

❖ Sur la figure 5.5, soit à 36 MFSC, nous modélisons assez bien la structure de l'enveloppe spectrale, en visualisant quelques importantes harmoniques visibles dans l'illustration 3D.

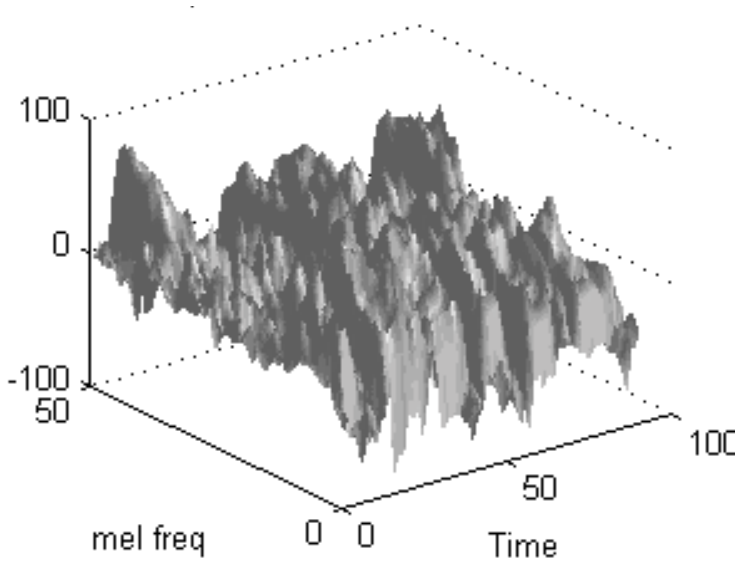


Fig. 5.6 : Représentation en 3D du spectre (échelle Mel) à l'aide de 48 canaux.

❖ Sur la figure 5.6, soit à 48 MFSC, nous modélisons moins bien la structure de l'enveloppe spectrale, mais en visualisant beaucoup mieux les harmoniques principales du spectre, bien visibles dans l'illustration 3D.

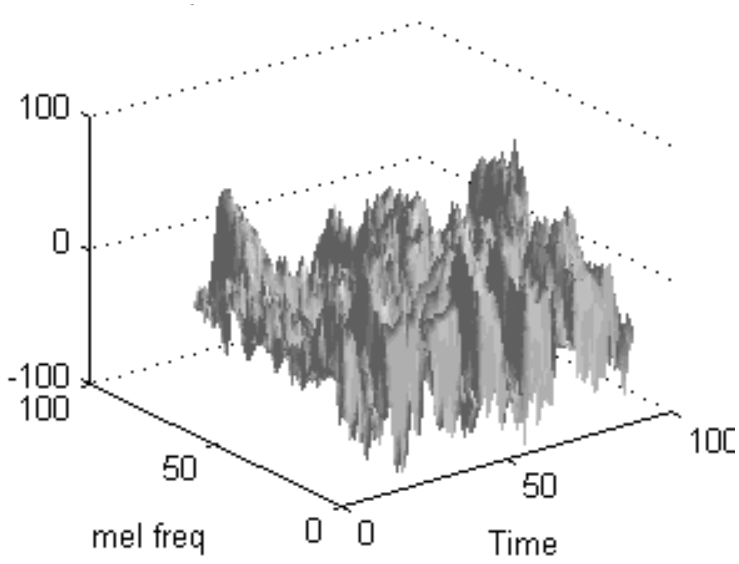


Fig. 5.7 : Représentation en 3D du spectre (échelle Mel) à l'aide de 60 canaux.

- Sur la figure 5.7, soit à 60 MFSC, nous perdons la structure de l'enveloppe spectrale en visualisant, par contre, toutes les harmoniques dans un contraste très clair, presque similaire à un spectre d'énergie classique.

Par cette étude nous constatons que la résolution spectrale permet d'ajuster la fiabilité de modélisation de l'enveloppe formantique et des harmoniques. Ainsi, vu que l'information contenue dans l'enveloppe formantique est importante pour identifier un locuteur, nous avons intérêt à réduire le nombre des filtres Mel, mais comme des études précédentes faites sur l'identification en sous-bandes spectrales [Bon 97] ont prouvé que les informations de haute fréquence apportent toujours une amélioration appréciable en identification du locuteur, il sera profitable alors d'introduire l'effet des harmoniques dans la paramétrisation MFSC si nous voulons obtenir une qualité d'identification supérieure à celle obtenue avec la basse résolution induisant une modélisation restreinte aux seules informations de basse fréquence ; ce qui se traduit par une élévation incontestable du nombre de filtres Mel. Toutefois il ne faut pas oublier que lorsque l'on utilise des dimensions très grandes nous nous heurtons à deux problèmes : la complexité de calcul et le mauvais conditionnement de la matrice de covariance.

Ainsi, théoriquement, la reconnaissance du locuteur nécessiterait un compromis entre la basse résolution spectrale et la haute résolution spectrale.

Une deuxième question se pose, alors : quel est le nombre optimal des coefficients Mel-spectraux pour cette tâche ?

Ce problème nous a incité à tester différentes résolutions spectrales allant du modèle classique à 12 canaux jusqu'au modèle complexe à 60 canaux.

Les résultats de cette étude sont représentés sur la figure 5.8.

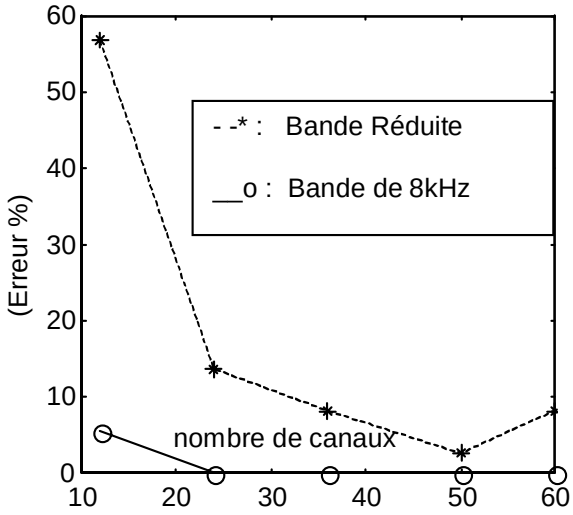


Figure 5.8 : Taux de fausses identifications pour le cas non bruité.

V.7. Cas non bruité

L'étude suivante consiste à identifier les 37 locuteurs de TIMIT [Fis 86] par la méthode 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 de [300-3400 Hz]. Les dimensions des MFSC varient, alors, de 12 à 60.

V.7.1. Résultats de cette étude

Les taux d'identification sont exposés sur la figure 5.8. 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 300-3400Hz,** 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 réduite.

V.7.2. Discussion et conclusion

Vu que l'information contenue dans l'enveloppe formantique est importante pour identifier un locuteur, on a intérêt à réduire la résolution spectrale ; cependant, vu que l'information contenue dans les harmoniques est complémentaire, il sera intéressant d'introduire l'effet de ces dernières par une élévation de la résolution spectrale d'où la nécessité de trouver un compromis optimal pour l'identification. Durant cette étude on a pu montrer l'importance de la haute résolution spectrale en trouvant le nombre optimal de filtres Mel à adopter en reconnaissance du locuteur quand le milieu est non bruité.

Ainsi, relativement à la figure 5.8, le nombre optimal de filtres Mels, en environnement non bruité, est compris indifféremment entre 24 et 60 pour la bande 0-8000 Hz et il est d'environ 48 canaux pour la bande téléphonique.

N.B. Le terme « bande téléphonique », ici, signifie que le signal original subit un filtrage du type téléphonique [300-3400 Hz]. Ce n'est pas un signal téléphonique réel qui est utilisé.

V.8. Cas bruité

La deuxième étude consiste à identifier les 37 locuteurs de TIMIT par la méthode SOSM, en environnement bruité. Les bruits utilisés sont décrits ci-dessous, avec une pondération variable correspondant à un RSB 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 [300-3400 Hz]. Les dimensions des MFSC varient, alors, de 12 à 60.

V.8.1. Bruits utilisés

Nous avons utilisé trois types de bruits pour essayer de diversifier le milieu environnant. Ce sont :

- le bruit blanc gaussien centré et stationnaire au deuxième ordre noté BBG [Cou87].
- 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 une sommation de 3 bruits réels de voiture enregistrés par un microphone de haute qualité dans une route à moyenne circulation et au format 16 bits/16 kHz.

Le bruitage est pondéré par le RSB choisi, allant de 0 dB jusqu'à 24 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.

Dans nos expériences, le bruitage est pondéré par le RSB imposé ; ainsi l'énergie moyenne du signal de parole est d'abord estimée ensuite une pondération du bruit normalisé (dont l'énergie moyenne est égale à l'unité) est effectuée avant de l'additionner au signal de parole, et ceci se fait dans la base de données d'apprentissage et de test aussi.

V.8.2. Observation des résultats obtenus

Les résultats de cette deuxième étude (environnement bruité) sont exposés sur les figures 5.9 à 5.14. Ainsi plusieurs constats peuvent être faits :

- 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 (voir fig. 5.9, 5.10 et 5.11).

- Sur la bande téléphonique 300-3400 Hz, 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 (voir fig. 5.12 à 5.14).

- Sur la bande 0-8 kHz, le bruit le plus gênant est le chahut suivi du BBG et enfin du bruit de voiture qui parait non gênant, relativement aux deux précédents (voir fig. 5.9, 5.10 et 5.11).

- Sur la bande téléphonique le bruit le plus gênant est le chahut suivi des deux autres types (voir fig. 5.12, 5.13 et 5.14).

- 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 (voir fig. 5.9, 5.10 et 5.11).

- Sur la bande téléphonique, 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 (voir fig. 5.12, 5.13 et 5.14).

V.8.3. Discussion et conclusion

La première conclusion à tirer est que l’approche statistique est très performante tant en milieu sain qu’en milieu bruité (à faible bruitage).

La deuxième conclusion à tirer est que la haute résolution spectrale apporte une grande quantité d'informations propre au locuteur et aide à le distinguer des autres, surtout en environnement bruité. Ainsi la dimension « 60 canaux » apparaît comme une dimension très favorable aux systèmes robustes de reconnaissance du locuteur en environnement bruité. Cependant sur la bande réduite l'optimal serait un compromis entre 48 et 60 canaux.

La trisième conclusion à tirer est 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. Alors que le chahut, si bien filtré par notre système nerveux, s'avère être extrêmement gênant pour identifier un locuteur, plus gênant encore que le bruit blanc. La cause la plus probable, de cette défaillance face au chahut, est que ce dernier est de même nature que le signal de parole et ainsi donc les caractéristiques spectrales les plus pertinentes, pour chaque locuteur, deviennent fortement altérées.

N.B. Le terme « bande téléphonique », ici, signifie que le signal original subit un filtrage du type téléphonique [300-3400 Hz]. Ce n’est pas un signal téléphonique réel qui est utilisé.

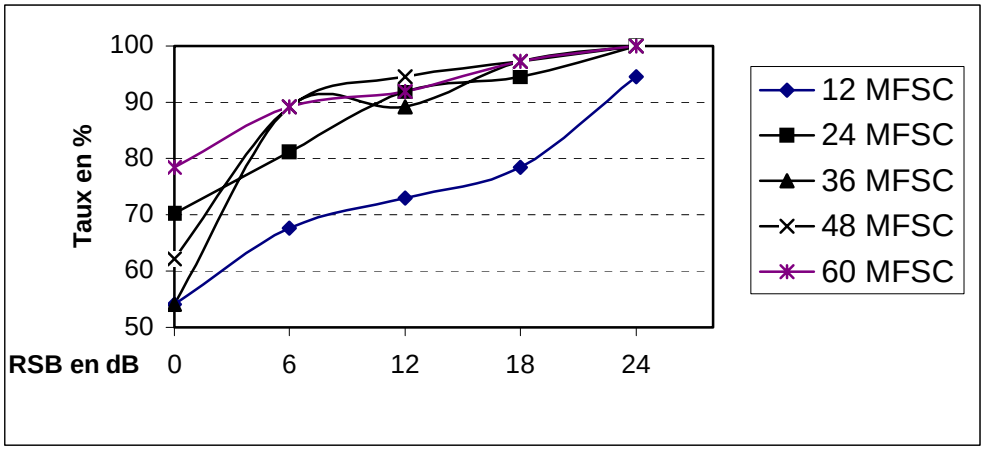


Fig. 5.9 : Taux de reconnaissance en environnement bruité par le BBG, sur la bande 0-8 kHz.

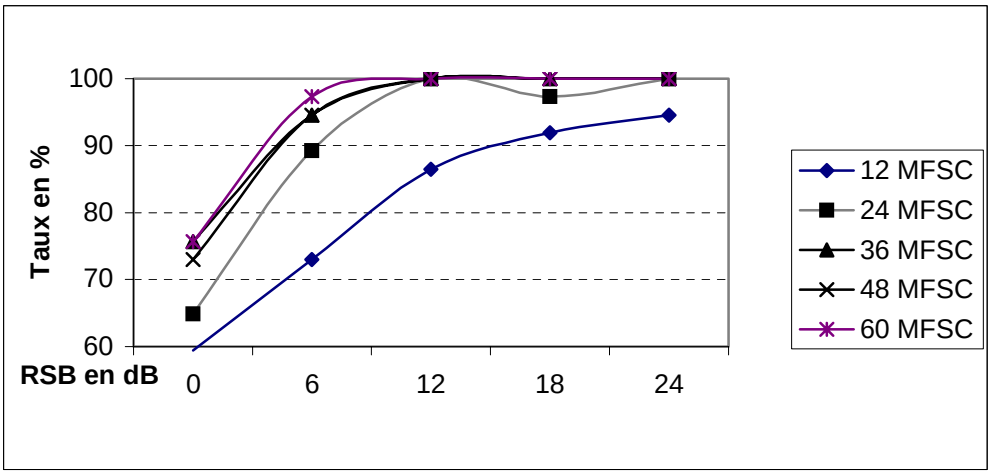


Fig. 5.10 : Taux de reconnaissance en environnement bruité par le chahut, sur la bande 0-8 kHz.

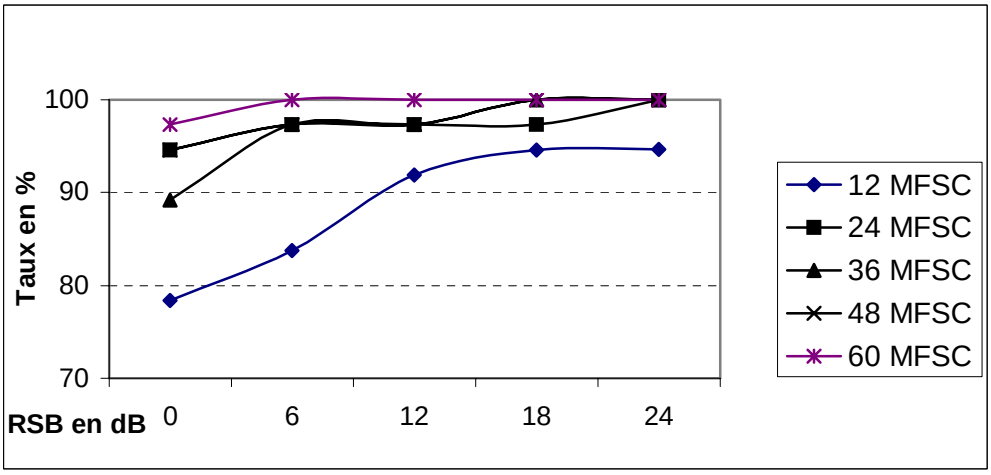


Fig. 5.11 : Taux de reconnaissance en environnement bruité par le bruit de voiture, sur la bande 0-8 kHz.

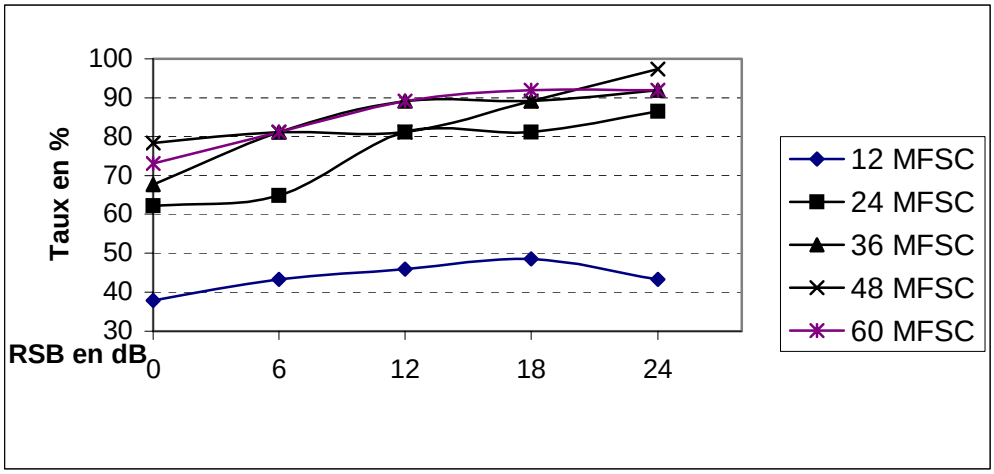


Fig. 5.12 : Taux de reconnaissance en environnement bruité par le BBG, sur la bande téléphonique.

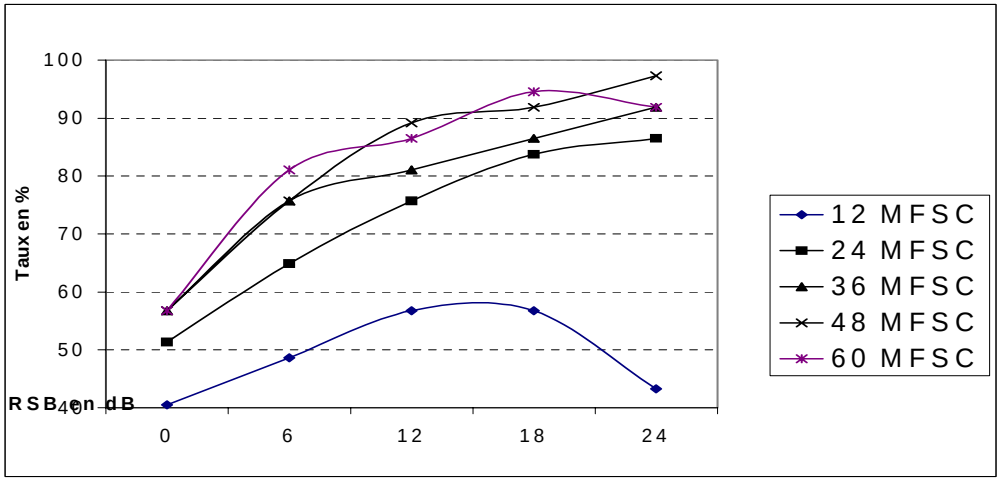


Fig. 5.13 : Taux de reconnaissance en environnement bruité par le chahut, sur la bande téléphonique.

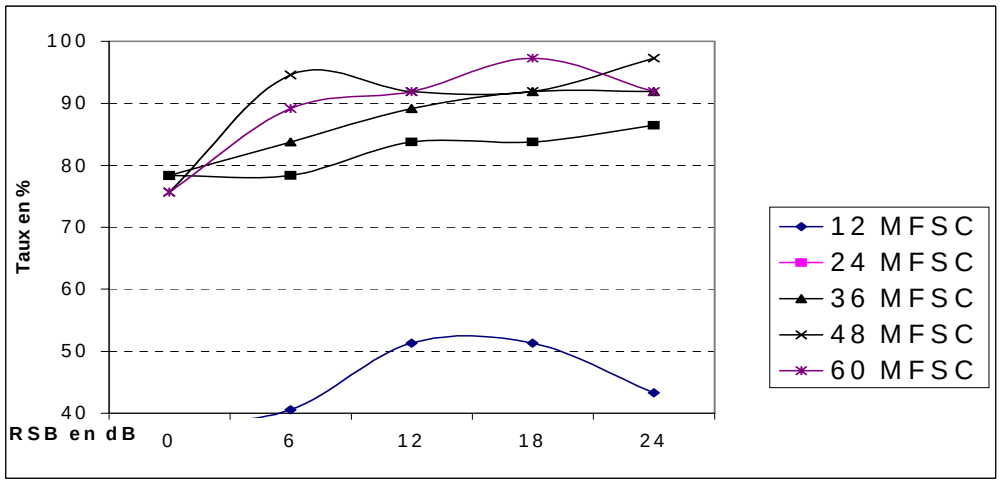


Fig. 5.14 : Taux de reconnaissance en environnement bruité par le bruit de voiture, sur la bande téléphonique.

V.9. 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 performante 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.

En milieu fortement bruité, cette méthode présente quelques problèmes ; c'est pour cette raison que nous avons essayé d'améliorer ses performances en jouant sur la résolution spectrale (taille des MFSC).

Généralement, les dimensions de basse résolution, qui sont souvent utilisées pour les coefficients Mel-spectraux, sont de l'ordre de 12 à 24 coefficients. Cette basse résolution spectrale a deux avantages : la simplification des opérations de calcul et la bonne représentation de l'enveloppe formantique du spectre. De plus, lorsque nous utilisons des dimensions très grandes nous nous heurtons aux problèmes suivants : la complexité de calcul et le mauvais conditionnement de la matrice de covariance.

Cependant, vu que les informations de haute fréquence peuvent aider à mieux distinguer les locuteurs, il est intéressant d'introduire ces dernières par une élévation de la résolution spectrale ; ce qui implique un compromis.

Durant cette étude nous avons pu montrer l'importance de la haute résolution spectrale en IAL, en trouvant le nombre optimal de filtres Mel 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 [300-3400Hz] 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 bruité. 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 du 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 divers tests d'identifications décrits dans ce chapitre.

- Chapitre VI -

Identification du Locuteur par la Méthode LVQ

- Proposition d'une Nouvelle Métrique Hybride -

N.B. *Ce chapitre concerne la deuxième série d'expériences effectuées en RAL.*

VI.1. Introduction

Dans cette étude, nous utilisons l'approche connexionniste du type LVQ, sous différentes formes (LVQ1 et MLVQ1) et avec différents types de caractéristiques acoustiques dans une application d'identification du locuteur.

Cependant, dans cette méthode nous nous heurtons souvent au problème d'incompatibilité des paramètres (caractéristiques acoustiques), dans le cas où ceux-ci ne sont pas homogènes. Comme exemple, nous pouvons citer l'association des MFSC avec Fo (caractéristiques hétérogènes).

Pour contourner ce problème, nous avons proposé une nouvelle métrique associative (que nous avons nommée ODHEF) dans le cas des paramètres acoustiques de natures différentes. Celle-ci nous a permis d'associer des paramètres hétérogènes dans le calcul d'une distance interlocuteur. Ainsi des caractéristiques très significatifs (telles que Fo) ont pu être associées aux paramètres spectraux (ou cepstraux) classiques, en rehaussant considérablement le taux de bonne reconnaissance trouvé par LVQ.

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

VI.2. Métrique ODHEF (**nouvelle métrique**)

Généralement, si nous avons des caractéristiques de même nature nous pouvons utiliser une métrique uniforme (par exemple L^2), mais dans le cas où ces caractéristiques sont de natures différentes, il sera incorrecte et illogique d'utiliser cette métrique classique. D'où la nécessité d'adopter une métrique associative multi-variable qui prenne en compte cette hétérogénéité et qui soit bien adaptée à la reconnaissance du locuteur.

Pour ce faire, nous définissons ci-dessous quelques notions importantes en caractérisation du locuteur (telles que la variabilité).

VI.2.1. Expression mathématique de la variabilité

Nous donnons d'abord quelques définitions de la variabilité (pour un paramètre donné) :

- **Variabilité Inter Locuteur** : C'est la variance des moyennes des différents locuteurs.
- **Variabilité Intra Locuteur** : C'est la moyenne des variances des différents locuteurs (ici, pour chaque locuteur, une variance est calculée).

- **Coefficient de Wolf « CW »** : C'est le rapport de la **Variabilité Inter Locuteur** sur la **Variabilité Intra Locuteur** [Wol 72].

$$CW(x) = \frac{Var_{\text{int er\_loc}}(x)}{Var_{\text{int ra\_loc}}(x)} \qquad (6.1)$$

avec x représentant une caractéristique acoustique quelconque.

VI.2.2. Signification physique de la variabilité

Voici quelques particularités de la variabilité :

La Variabilité Inter-Locuteur est grande : indique qu'on peut, facilement, séparer les locuteurs par leurs caractéristiques.

La Variabilité Intra-Locuteur est petite : indique que chaque locuteur peut être représenté par une référence qui représente très bien ce locuteur.

Le Coefficient de Wolf est grand : indique, alors selon (6.1), que le paramètre choisi est pertinent en identification du locuteur et devrait donner un bon score de reconnaissance.

Exemple : Mesure de CW pour Fo moyen

Les calculs de la variabilité de Fo moyen (Fo_{moy}) ont montré que son coefficient de Wolf $CW(Fo_{moy})$ vaut 229.

Nous considérons que ce taux est très élevé relativement à 7 autres paramètres étudiés simultanément avec Fo_{moy}.

VI.2.3 Principe de la distance ODHEF (cas de 3 paramètres hétérogènes)

Nous avons essayé d'élaborer une métrique, adaptée à notre application de RAL, pour des caractéristiques de natures différentes.
On considère ici 3 paramètres, seulement pour simplifier, mais l'extension à N paramètres différents est parfaitement possible.

Ainsi, 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 :

$$\boxed{Dist^2(X, Xr) = \sum_{i=1}^{3} (Fiab_i)(Norm_i)(Xr_i - X_i)^2} \qquad (6.2)$$

X étant un vecteurs représentant la prononciation d'un locuteur quelconque et Xr correspond à un vecteur représentant la prononciation d'un locuteur de référence, par exemple.

Les composantes, de caractéristiques différentes, du vecteur X sont X1, X2 et X3 (de même Xr1, Xr2, Xr3 sont les composantes du vecteur Xr).

- Le 1er terme, après la somme dans la formule 6.2, 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 premier paramètre.

$$Fiab_i = \left(\frac{\sigma^2_{Inter}(X_i)}{\sigma^2_{Intra}(X_i)}\right) \tag{6.3}$$

- 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 très 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.

$$Norm_i = \left(\frac{1}{E(E(X_i))}\right)^2 \tag{6.4}$$

- Le 3ème terme, dans la formule 6.2, est la distance euclidienne classique.

Par conséquent, l'expression générale de cette nouvelle distance sera :

$$Dist^2(X, Xr) = \sum_{i=1}^{3} \left(\frac{\sigma^2_{Inter}(X_i)}{\sigma^2_{Intra}(X_i)}\right)\left(\frac{1}{E(E(X_i))}\right)^2 (Xr_i - X_i)^2 \tag{6.5}$$

Ce qui peut se simplifier en :

$$Dist^2(X, Xr) = \sum_{i=1}^{3} \left[\left(\frac{\sigma_{Inter}(X_i)}{\sigma_{Intra}(X_i)}\right)\left(\frac{1}{E(E(X_i))}\right)(Xr_i - X_i)\right]^2 \tag{6.6}$$

ou mieux encore :

$$\sum_{i=1}^{3} \left[\left(\frac{\sigma_{Inter}(X_i)}{\sigma_{Intra}(X_i)}\right)\left(\frac{1}{E(E(X_i))}\right)Xr_i - \left(\frac{\sigma_{Inter}(X_i)}{\sigma_{Intra}(X_i)}\right)\left(\frac{1}{E(E(X_i))}\right)X_i\right]^2 \tag{6.7}$$

soit :

$$Dist^2(X, Xr) = \sum_{i=1}^{3} [Pond(i)Xr_i - Pond(i)X_i]^2 \tag{6.8}$$

où Pond(i) représente le coefficient de pondération normalisé du paramètre "i" visible dans la formule 6.7.

En introduisant la variable pondérée Y (égale au produit du coefficient de pondération 'Pond' par la vecteur X), l'expression précédente devient :

$$Dist^2(X, Xr) = \sum_{i=1}^{3} (Yr_i - Y_i)^2 \tag{6.9}$$

Ce qui se ramène à un calcul de distance classique du type L^2, après transformation des vecteurs X en vecteurs Y.

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

Par conséquent, on peut utiliser la métrique L² classique à condition de transformer les vecteurs X en des vecteurs adaptés Y par une transformation de pondération définie par :

$$Y = \begin{pmatrix} Pond(1) & 0 & 0 \\ 0 & Pond(2) & 0 \\ 0 & 0 & Pond(3) \end{pmatrix} X \qquad (6.10)$$

Remarque1 : Cette démonstration à 3 variables peut aisément être étendue à N variables différentes.

Remarque 2 : Dans le cas où les N variables sont de même nature, alors

Pond(1) = Pond(2) = ... =Pond(N),

donc la matrice de transformation devient proportionnelle à la matrice identité. C'est alors une métrique du type L² classique.

Nous avons appelé cette métrique : "**Optimized Distance for Heterogeneous Features**" ou **ODHEF** (*en Français : Distance Optimisée pour des Caractéristiques Hétérogènes*).

VI.3. Base de données utilisée lors des expériences d'identification

Pour les expériences d'identification, nous avons choisi une base de données parlée régie par les conditions suivantes :

Langue : Arabe standard.

Choix des phonèmes : Après une longue étude sur les phonèmes les plus typiques au locuteur, on a choisi des phrases riches en "Kaf" (ك), "Rra" (ر), "Aïn" (ع) et "Dta" (ط).

Tableau 6.1 : Phonèmes souvent utilisés dans nos phrases.

Désignation du phonème	Ressemblance latine	Symbole arabe
Kaf	K	ك
Rra	R espagnol	ر
Aïn	Aucune	ع
Dta	Aucune	ط

Locuteurs : 6 adultes masculins et 5 adultes féminins soit : 11 locuteurs adultes de voix normales.

Ensembles de locuteurs : Nous avons 2 ensembles de locuteurs : un ensemble fermé constitué de 11 locuteurs, et un ensemble ouvert constitué de ces 11 locuteurs et de 4 autres imposteurs (figures 6.1 et 6.2).

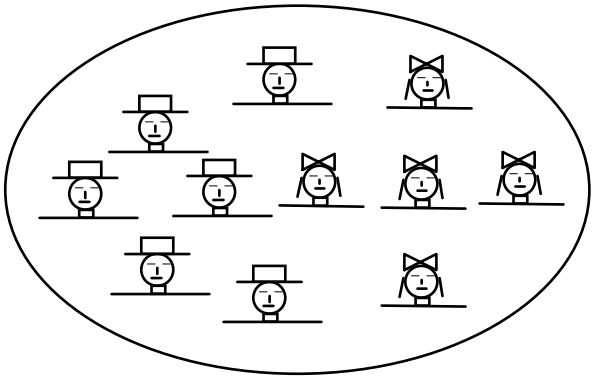


Figure 6.1 : Locuteurs de l'ensemble fermé (en tout, 11 locuteurs adhérents).

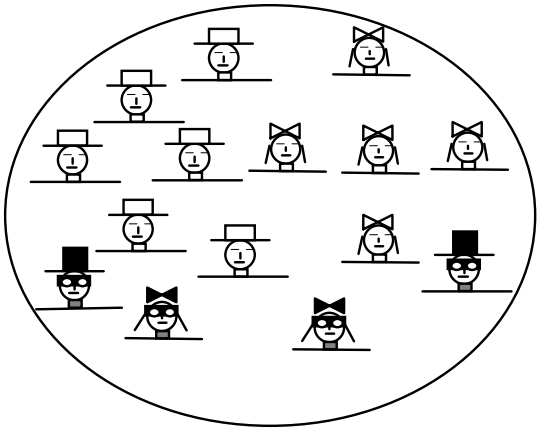


Figure 6.2 : Locuteurs de l'ensemble ouvert (en tout, 11 locuteurs adhérents et 4 imposteurs). *Les locuteurs masqués sont des imposteurs.*

Durée moyenne d'une phrase : 2.7 secondes à un rythme normal.

Types de phrases : 4 phrases textuelles différentes : 2 phrases concaténées en phrase A et deux autres concaténées en phrase T (A étant différente de T).

Durée moyenne d'une prononciation (A ou T) : 5.5 secondes.

Nombre de répétitions d'une prononciation : la prononciation A est répétée 10 fois et la prononciation T est répétée 4 fois.

Qualité d'enregistrement : Bonne à moyenne qualité (RSB moyen).

Acquisition du signal de parole : 16 bits à 16 kHz.

VI.4. Méthodes de reconnaissance du locuteur à base de LVQ1

Nous avons testé 3 méthodes différentes, toutes sont basées sur la LVQ (LVQ1 ou MLVQ1), pour tenter d'identifier chacun des locuteurs du test.

VI.4.1. Nouvelle Méthode Prosodique

Dans cette nouvelle méthode, nous avons réalisé un système d'identification du locuteur purement prosodique, basé sur Fo, la durée et l'énergie.

VI.4.1.1. Description

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 $(Fo_{moy}, D_{orig}, E_{bf})^t$.

La description de ces caractéristiques est donnée ci-dessous.

- **Fo_{moy}** est la moyenne du pitch, en Hz, sur toute la phrase. Son choix est justifié par le grand coefficient de Wolf qu' on obtient avec Fo_{moy} : **$CW(Fo_{moy})=229.$**

- **D_{orig}** est la durée en ms d'un mot ORIGINAL mis au milieu d'une phrase. Le mot original est un mot nouveau ou étranger au dialecte de la population des locuteurs. Car nous avons remarqué expérimentalement que la durée d'un mot prononcé pour la 1ère fois était très typique à l'état psychologique de l'individu et que chacun essaie de lui adapter une durée typique : Un bébé qui prononce un certain mot, pour la 1ère fois, lui attribue une durée originale qui sera altérée, par la suite, par l'influence des parents. Son coefficient de Wolf est : **$CW(D_{orig})=26.9.$**

- **E_{bf}** est l'énergie basse fréquence à la sortie du 3ème ou du 4ème canal du banc de filtres (à 24 canaux) en dB. Le choix de l'énergie basse fréquence est justifiée par les résultats de BONASTRE en [Bon97]. L' ordre 3 et 4 ont été retenus après test expérimental. Son coefficient de Wolf est : **$CW(E_{bf})=13.88$**.

Nous pouvons voir sur la figure 6.3 la représentation des 11 locuteurs avec ces 3 paramètres.

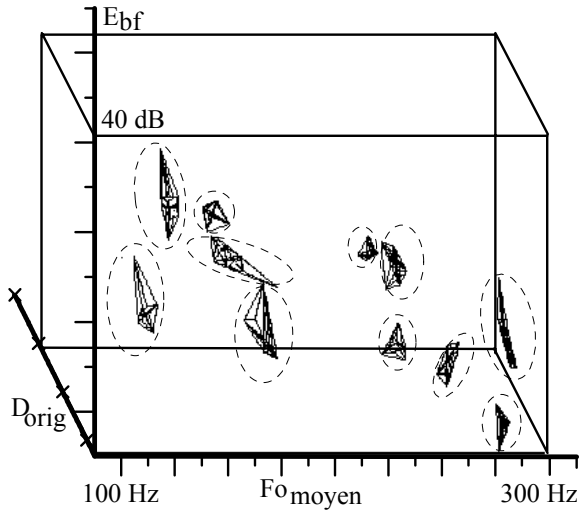


Fig. 6.3 Représentation des 11 locuteurs avec les 3 dimensions prosodiques Fo_{moy}, D_{orig} et E_{bf}.

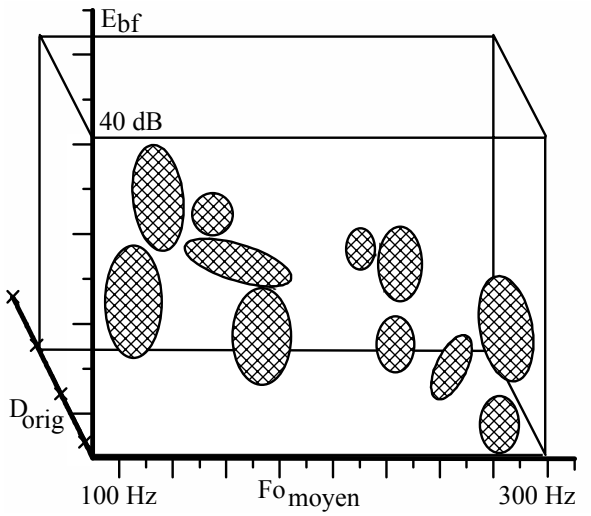


Fig. 6.4 Représentation des régions des 11 locuteurs avec les 3 dimensions prosodiques Fo_{moy}, D_{orig} et E_{bf}. Nous pouvons distinguer 6 hommes à gauche et 5 femmes à droite.

Nous pouvons 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.4.1.2. Algorithme

On utilise dans cette méthode la LVQ1 et la MLVQ1 (Modified LVQ1). Le temps d'apprentissage est de 30 secondes et la distance utilisée est la distance euclidienne multivariable ODHEF (voir section VI.2.).

Les calculs statistiques de la matrice de pondération de la métrique ont donné les résultats suivants :

$Pond(1)=Pond(Fo_{moy})=0.24$.

$Pond(2)=Pond(D_{orig})=0.21$.

$Pond(3)=Pond(E_{bf})=0.54$.

VI.4.1.3. Description de la MLVQ1

La MLVQ1 ou modified-LVQ1 est une variante de la LVQ1 à laquelle nous avons apporté certaines modifications. Les tests expérimentaux ont montré, que dans certains cas, la MLVQ1 donnait de meilleurs résultats que la LVQ1 classique. Mais ceci n'est pas toujours vérifié en pratique.

L'algorithme de la MLVQ1 (ou LVQ1 modifiée) se résume ainsi :

> Supposons que l'on a à l'entrée du réseau de neurones un vecteur X et que le neurone de sortie de la famille de Xr est égal à 1 (i.e. X identifié comme étant de la même famille que Xr).

Ainsi :

- Si X est réellement de la même famille que Xr

  alors on ne fait rien ;

- et Si X n'est pas de la même famille que Xr

  alors on éloigne la référence Xr selon l formule suivante :

  $$\mathbf{Xr(t+1)=Xr(t)-\alpha(t).(\{X(t)-Xr(t)\}} \quad (6.11)$$

  où la fonction $\alpha(t)$ est une fonction décroissante du temps "t" avec $0< \alpha(t)$.

VI.4.1.4. 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 6.2. L'identification par cette méthode se fait correctement, en donnant des taux appréciables.

Tableau 6.2 : Taux de reconnaissance par la nouvelle méthode prosodiques.

Algorithme	Taux de Reconnaissance %	
	Ensemble fermé (11 locut.)	Ensemble ouvert (15 locut.)
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.4.2. Méthode Cepstrale

Dans cette méthode, 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 calcule le 1er Vecteur Propre [Ben 95] car les premiers vecteurs propres sont les plus représentatifs, et nous gardons, alors, seulement les caractéristiques suivantes :

MFCC$_{moy}$**,** et **MFCC**$_{vp1}$**.**

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

VI.4.2.1. Algorithme

L'algorithme de cette méthode se compose de la phase d'apprentissage basée sur la LVQ1 et de la phase de test de reconnaissance basée sur la technique du plus proche voisin.

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. Voir « Algorithme de la LVQ1 ».

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:

$$\text{Dist}_{\text{ODHEF}} = \sqrt{\alpha_1 d^2_1 + \alpha_2 d^2_2} \qquad (6.12)$$

où d^2_1=distance relative à MFCCmoy et d^2_2= distance relative à MFCCvp1,

Les coefficients de pondération α_i valent :

$$\alpha_1 \approx 0.5 \text{ et } \alpha 2 \approx 0.5.$$

VI.4.2.2 Résultats

Les tests effectués, avec cette méthode, sur un ensemble de 11 locuteurs ont donné un taux de bonne reconnaissance dépendant du texte égal à 79%. Voir tableau 6.3.

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.

Les résultats obtenus par cette méthode sont insuffisants. Cela est dû au fait qu'on a remplacé la matrice de covariance par le premier vecteur propre; ce qui est une approximation simplificatrice entraînant des erreurs d'estimation.

Tableau 6.3 : Taux de Reconnaissance par la méthode cepstrale

Algorithme	Taux de bonne Reconnaissance
LVQ1	79%
MLVQ1	79%

VI.4.3 Méthode Spectrale

Dans cette méthode, 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 $MFSC_{moy}$ qui sera le seul élément représentatif du locuteur dans ce cas : dans cette perspective nous supposons que la variabilité de la matrice de covariance est négligeable.

Nous préservons alors seulement le vecteur moyenne : $\mathbf{MFSC}_{moy}$ (de dimension 24x1) et chaque locuteur sera représenté par ce vecteur.

VI.4.3.1 Algorithme

L'algorithme de cette méthode se compose de la phase d'apprentissage basée sur la LVQ1 et de la phase de test de reconnaissance basée sur la technique du plus proche voisin.

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 $\mathbf{MFSC}_{moy}$ 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**.

VI.4.3.2 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 6.4.

Par ailleurs, des tests d'identification en ensemble ouvert ont aussi été faits (présence d'imposteurs) et un seuil d'acceptation a donc été imposé avant confirmation du résultat d'identification :

Si **dist(X,Xref)** < Seuil alors **X ⇔ Xref** (Acceptation) sinon Rejet.

Le taux de reconnaissance en ensemble ouvert est aussi de 100%. Les résultats montrent que cette méthode est assez performante sur l'ensemble des 11 locuteurs, cependant les résultats d'identifications devraient être moins appréciables si le nombre des locuteurs était plus important, car dans cette approche on suppose la matrice de covariance négligeable, alors

qu'en pratique beaucoup de locuteurs possèdent des vecteurs moyennes très similaires et ne sont discernables que par leurs covariances.

Tableau 6.4 : Taux de Reconnaissance par la méthode spectrale.

Méthode	**Taux de Reconnaissance %**		
	Dépendant du texte	Indépendant du texte	
		Ensemble fermé	Ensemble ouvert
LVQ1	99	100	100

VI.5. Conclusion

Nous avons proposé trois méthodes connexionnistes basées sur la LVQ1 pour l'identification automatique du locuteur. Pour ce faire, nous avons mis au point une nouvelle métrique (ODHEF) capable d'associer des caractéristiques hétérogènes dans un système LVQ et qui a très bien fonctionné.

Nous avons pu déduire, des résultats précédents, une estimation sur les performances de l'approche LVQ1 appliquée sur différents types de caractéristiques acoustiques.

Globalement, cette approche donne 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.

Toutefois, rien n'empêche d'utiliser ces types d'approches dans le cas de petites bases de données. Comme, on peut aussi les utiliser dans des applications hybrides (statistico-neuronales) pour rehausser la performance des méthodes statistiques.

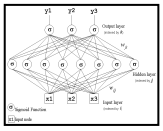
En résumé :

- la méthode prosodique (*que nous venons de mettre au point récemment*), semble donner des résultats précis avec un taux de l'ordre de 100% de bonne reconnaissance, mais les paramètres utilisés ici sont assez difficiles à extraire.

- la méthode cepstrale a manqué de performance 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 performante : taux de reconnaissance proche de 100%, cependant les résultats d'identifications devraient être moins appréciables si le nombre des locuteurs était plus important, car dans cette approche on suppose la matrice de covariance négligeable, alors qu'en pratique beaucoup de locuteurs possèdent des vecteurs moyennes très similaires et ne sont discernables que par leurs covariances.

Nous tenons à faire remarquer, enfin, que 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 est très intéressante.



- Chapitre VII -

Vérification Automatique du Locuteur par MLP

N.B. *Ce chapitre concerne la troisième (et dernière) série d'expériences effectuées en RAL.*

VII.1. Introduction

L'objectif de cette étude (connexionniste) est la réalisation d'un système de vérification automatique du locuteur (VAL) en utilisant des réseaux de neurones mono-locuteur à base de MLP (Multi-Layer Perceptron).
Nous allons présenter, dans ce chapitre, la procédure qu'on a élaborée pour réaliser cet 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.

VII.2. Base de données

Nous avons utilisé, pour les tests de VAL, une base de données universelle (TIMIT) composée, en tout, de 420 locuteurs (à partir de la quelle on a utilisé 37 locuteurs) et comprenant 8 dialectes différents. Chaque locuteur prononce 10 phrases, dont huit sont complètement différentes d'un locuteur à l'autre. La parole est enregistrée à travers un microphone de très haute qualité, échantillonnée à 16 kHz et codée sur 16 bits.
Pour tester notre système de VAL, nous avons travaillé sur un sous-ensemble de cette base, composé de 15 femmes et de 22 hommes. Le dialecte choisi est appelé « *New England* ».

VII.3. Réseau de neurones utilisé

Nous avons 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 (voir chapitre IV). 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. Le réseau conçu est du type mono-locuteur : chaque locuteur possède un réseau (matrice de poids) propre à lui.

VII.4. Technique d'apprentissage et de test

Les techniques d'apprentissage et de test, utilisées dans nos expériences, sont résumée s sur le synoptique de la figure 7.1 suivante.

Le module d'apprentissage (pour un locuteur donné) fait présenter au MLP une série d'exemples différents appartenant au locuteur désiré (que l'on veut référencer) et à d'autres locuteurs différents. A chaque fois, on vérifie la sortie du MLP et on réactualise les poids du réseau grâce à l'algorithme de la RPG. A la fin de l'apprentissage on sauvegarde les poids obtenus : ceux-ci représentent, alors, la référence MLP du locuteur désiré.

Le module de test, par contre, fait affecter au MLP les poids (déjà sauvegardés) correspondant au locuteur proclamé. Il calcule ensuite la sortie du réseau qui décidera de l'acceptation ou du refus de l'identité proclamée.

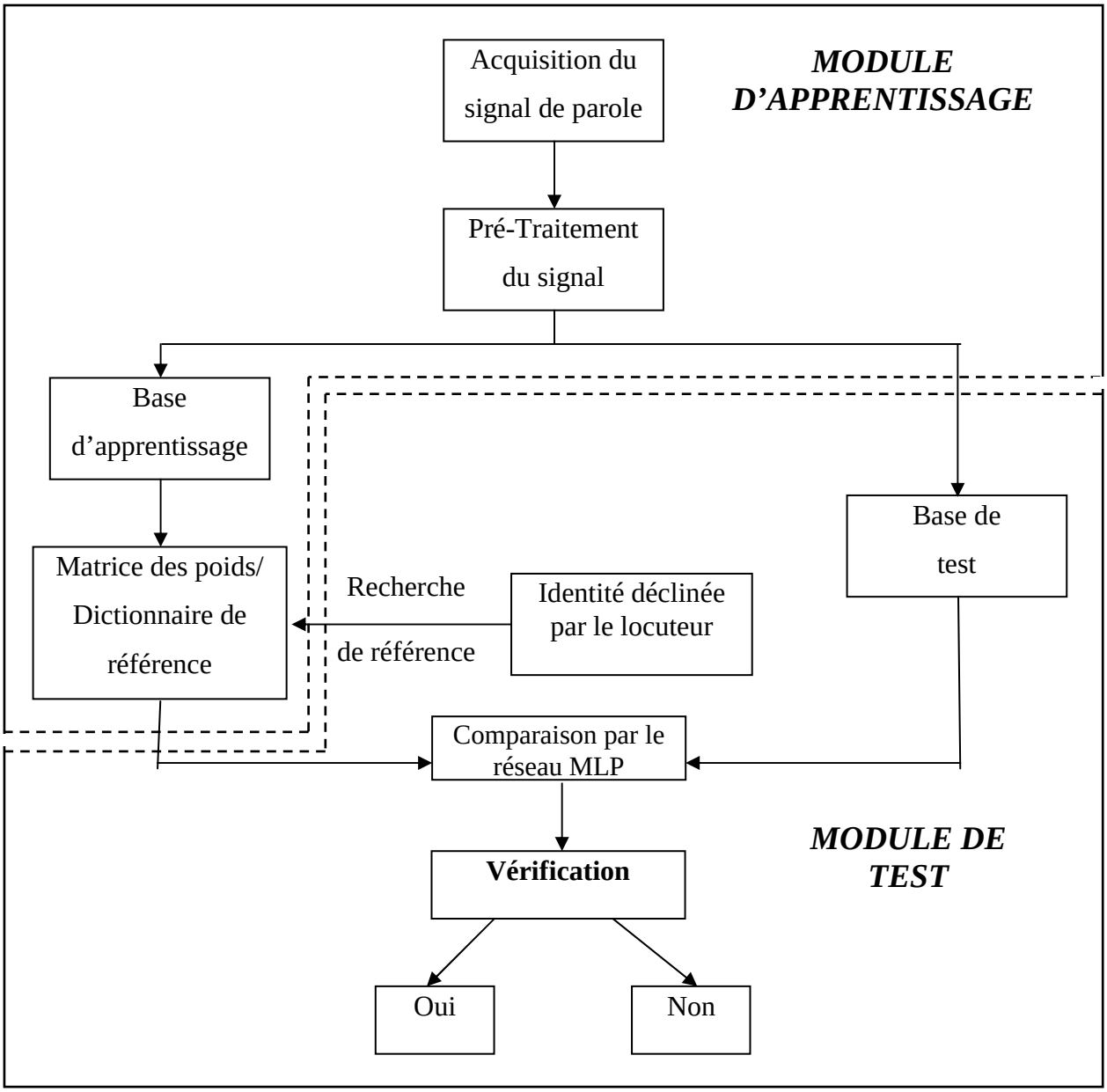


Figure 7.1 : Synoptique général de la technique d'apprentissage et de test.

VII.4.1. Acquisition du signal de parole

En ce qui concerne l'acquisition du signal de parole, nous avons utilisé un sous-ensemble de la base TIMIT qui contient 37 locuteurs (22 hommes et 15 femmes). Ainsi, dans notre cas

l'acquisition a été faite par les concepteurs de cette base de données [Fis 86] et avec les conditions décrites précédemment.

VII.4.2 Pré-Traitement du signal

Chaque signal issu d'une phrase prononcée par l'un des locuteurs est analysé de la façon suivante :

- Segmentation du signal de parole en trame de 512 échantillons (soit 32 ms).
- Application d'une fenêtre rectangulaire à chaque trame avec un recouvrement de 50% soit : un pas de glissement de 256 échantillons (soit 16 ms).
- Application d'une FFT sur cette fenêtre pour l'évaluation du spectre.
- Filtrage spectral par 24 bancs de filtres à l'échelle Mel.
- Calcul et déduction des coefficients MFSC comme c'est décrit au chapitre V.

VII.4.3 Dimension du réseau de neurones

Le nombre de neurones de la couche d'entrée est un vecteur de taille fixe qui devrait représenter la taille des coefficients MFSC. Or pour chaque fenêtre du signal, un vecteur de taille M (M=24) est produit. Ainsi, si nous avons 1000 fenêtres, alors on devrait avoir 24000 neurones à la couche d'entrée, ce qui est irréalisable en pratique.
Pour cette raison on a procédé à une normalisation (en éliminant le facteur temps) des entrées en utilisant :

- Le vecteur moyenne,
- la diagonale de la covariance extraite des MFSC,
- les deux premiers vecteurs propres de cette covariance.

Ainsi, la dimension de la couche d'entrée ne dépendra que des caractéristiques choisies à l'entrée du réseau : par exemple, si l'on choisit les 2 premiers vecteurs propres comme entrée du réseau alors le nombre de neurones d'entrée sera 2x24 soit 48 neurones.

Rappelons que ces trois paramètres sont calculés statistiquement de la manière suivante : Soit $x : \{ x_j \}$ une séquence de N vecteurs de dimension M résultant de l'analyse acoustique d'un signal de parole prononcé par le locuteur x (coefficients MFSC, par exemple). La moyenne $\bar{x}$ et la matrice de covariance Cov sont alors définies par :

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j \quad (7.1) \qquad Cov_X = \frac{1}{N}\sum_{j=1}^{N}(x_j - \bar{x})(x_j - \bar{x})^T \quad (7.2)$$

Le vecteur diagonal est directement déduit de la matrice de covariance.

Pour le calcul des deux premiers vecteurs propres, on calcule tout d'abord les valeurs propres λ_i de la covariance, puis on calcule leurs vecteurs propres correspondants V_i à partir de l'équation suivante : $(\lambda_i I - Cov)V_i = 0$ (7.3)

Pour la taille de la couche de sortie, celle-ci devra être suffisamment grande pour pouvoir coder tous les locuteurs référencés. Dans nos expériences, nous avons 37 locuteurs numérotés de 1 à 37, ce qui nécessite 6 bits au moins :

Exemple de codage des locuteurs par les neurones de sortie

- 000000 : imposteur
- 000001 : locuteur 1
- 000010 : locuteur 2
- 000011 : locuteur 3
- ……
- 100101 : locuteur 37.

Quant à la taille de la couche cachée celle-ci devra être estimée et optimisée expérimentalement.

VII.4.4 Spécifications du réseau de neurones et protocole d'apprentissage

On a utilisé un réseau de trois couches : une couche d'entrée, une couche cachée, et une couche de sortie.

La couche d'entrée est constituée de M neurones (où M correspond à la taille du paramètre d'entrée choisi).

Le nombre de neurones de la couche cachée sera variable, afin d'étudier les capacités d'apprentissage du réseau en vue de son optimisation.

La couche de sortie est constituée de 6 neurones (on a codé la sortie désirée sur 6 bits, ce nombre va assurer que les 37 réseaux seront codés différemment en utilisant le code binaire, et chaque code va représenter un mot de passe par locuteur).

La fonction d'activation des neurones est la fonction sigmoïde suivante :

$$f(x) = \frac{1}{1+e^{-x}} \qquad (7.4)$$

dont l'allure est schématisée par la figure suivante.

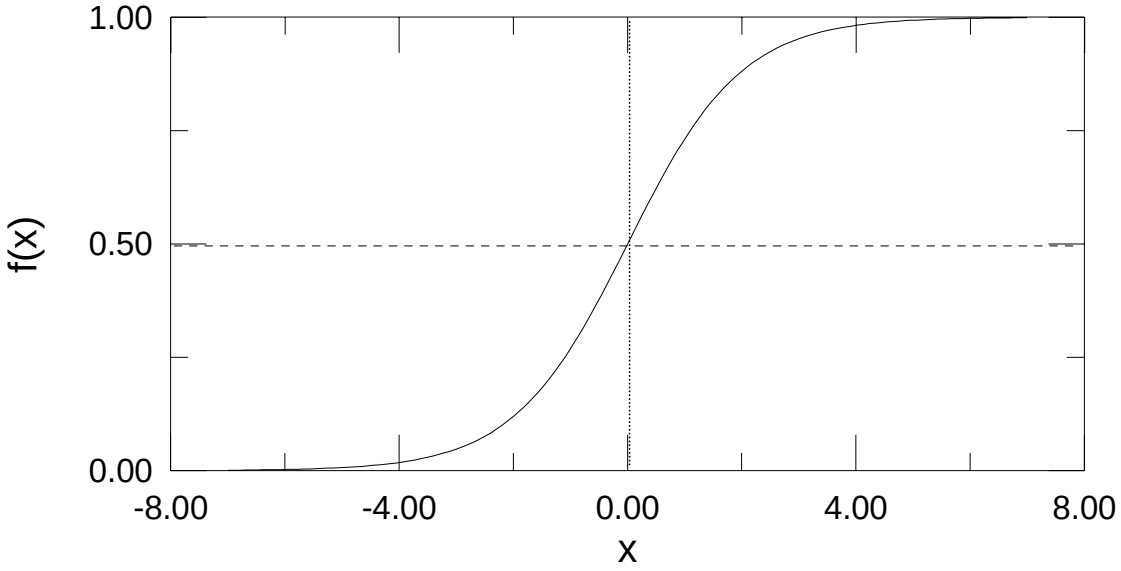


Figure 7.2 : Allure de la fonction sigmoïde.

L'apprentissage est initialisé avec les paramètres suivants (après tests expérimentaux) :

- Pas d'apprentissage : Pa = 0.2
- Erreur d'arrêt désirée : Err = 10^{-3}
- Nombres d'itération : Iter = 100.000
- 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 la 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 : i.e. dès que Ep est inférieur à l'erreur voulue Err (voir figure 7.3).

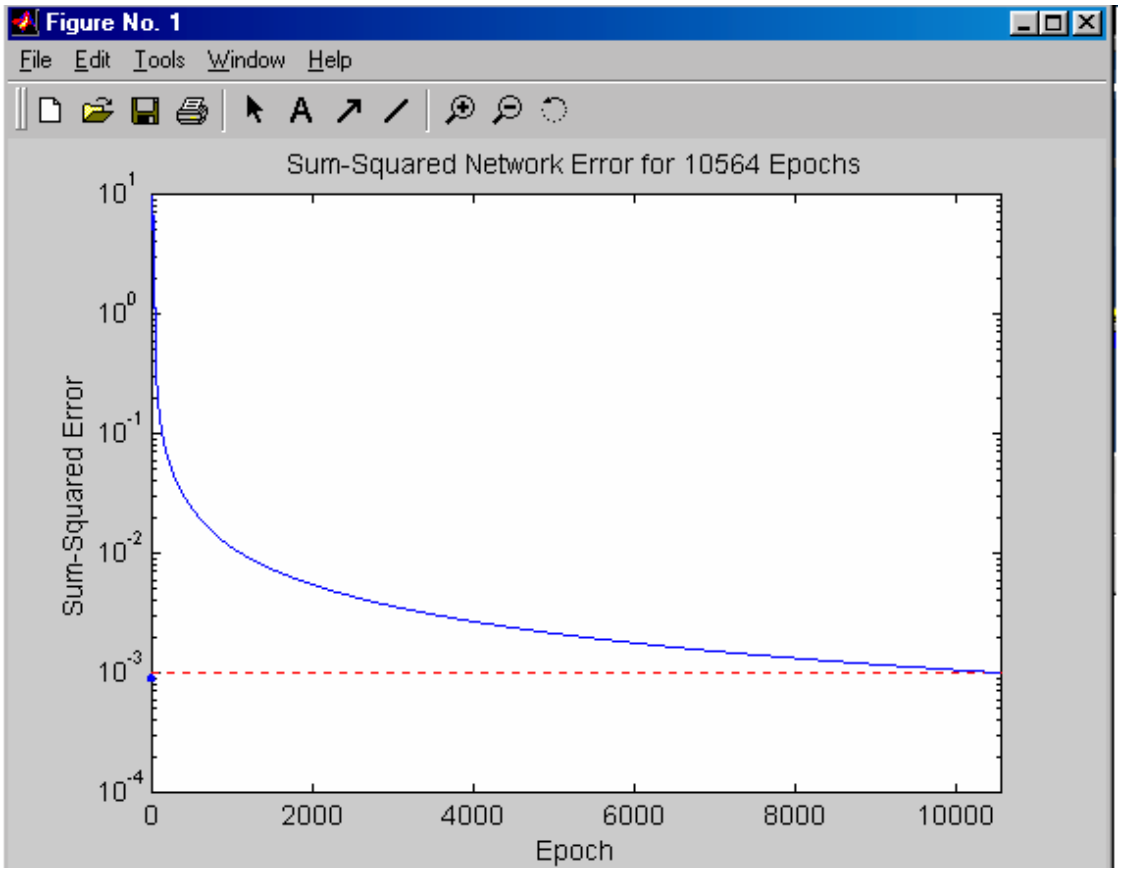


Figure 7.3 : Evolution de l'erreur quadratique en fonction du nombre d'itération au cours de l'apprentissage. Ici on voit que la convergence est atteinte après 10.564 itérations.

La base de test (différente de celle d'apprentissage) correspond à 5 phrases prononcées par chaque locuteur. Ainsi, en tout, nous aurons 185 phrases utilisées juste pour les tests.

Lors de la phase de vérification, on va charger les paramètres de référence de chaque locuteur proclamé : il s'agit des poids et des biais sauvegardés à la fin de son apprentissage.

VII.4.5. Description des algorithmes utilisés

Nous avons jugé utile de détailler certains algorithmes utilisés dans nos programmes, ceux-ci sont illustrés sur les figures suivantes.

La figure 7.4 présente le principe de l'apprentissage du réseau MLP et la figure 7.5 présente le principe de la vérification du locuteur par MLP.

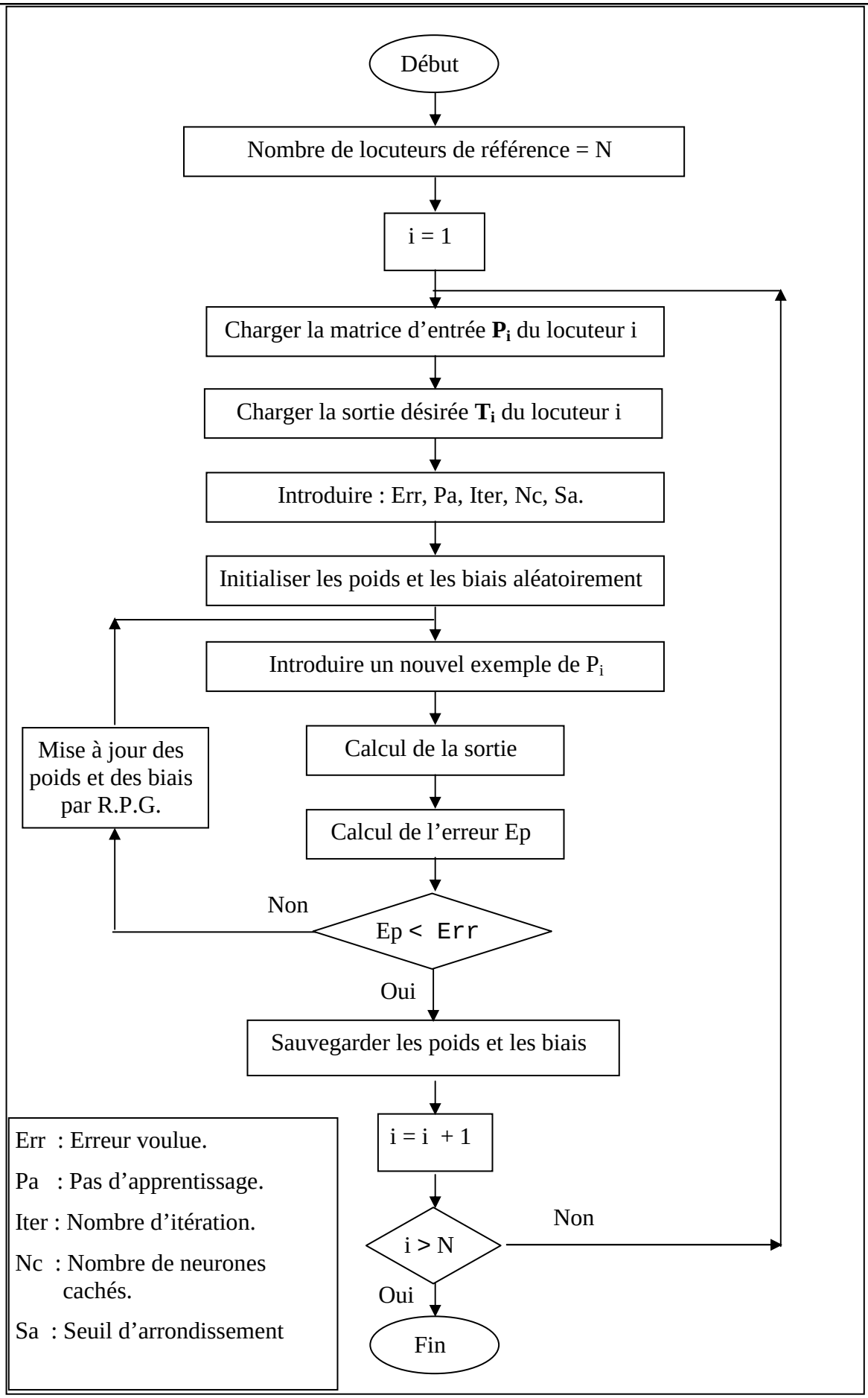


Fig. 7.4 : Principe d'apprentissage d'un MLP mono-locuteur

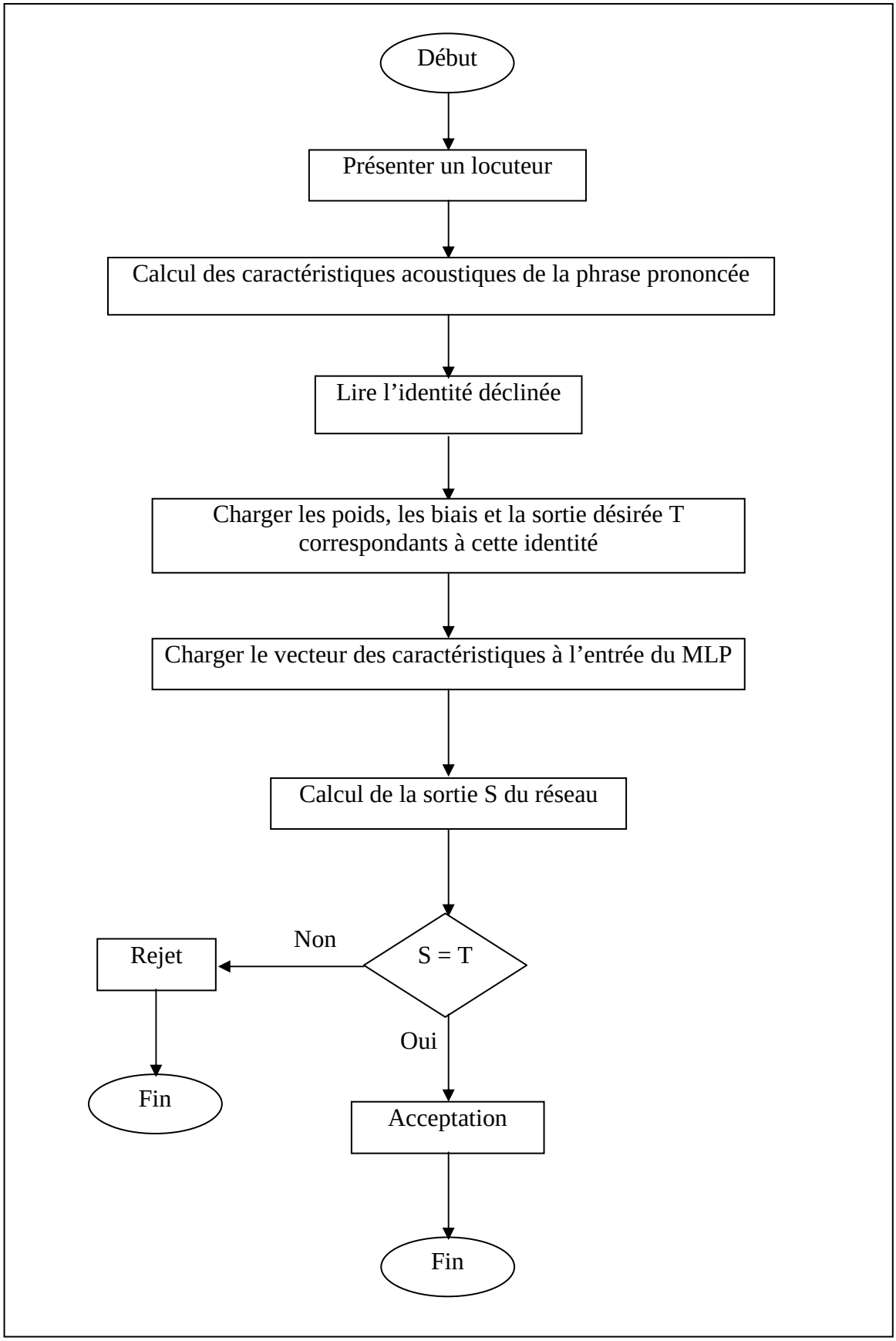


Fig. 7.5 : Principe de la vérification du locuteur par MLP mono-locuteur

VII.5. Résultats des expériences en vérification automatique du locuteur

Dans ce qui suit, nous procédons à des expériences de VAL (vérification automatique du locuteur) par le biais du MLP mono-locuteur élaboré.
Mais, tout d'abord, nous devons définir quelques taux d'erreurs souvent employés en VAL.

On définit :

- Le Taux de Bonne Acceptation (TBA %) par :

$$\text{Taux de Bonne Acceptation} = \frac{\text{Nombre de phrases bien reconnues}}{\text{Nombre total de phrases de test}} \times 100 \qquad (7.5)$$

- Le Taux de Faux Rejet (TFR %) par :

$$\text{Taux de Faux Rejet} = \frac{\text{Nombre de phrases rejetées par erreur}}{\text{Nombre total de phrases de test}} \times 100 \qquad (7.6)$$

Il existe, en vérification, un autre terme important ; c'est :

- **Le Taux de Fausse Acceptation** : *lorsqu'un locuteur prend l'identité d'un autre locuteur.*

  Pour le calculer, on présente toutes les phrases des locuteurs (sauf celles du locuteur propriétaire du réseau) à l'entrée de chaque réseau, et on calcule le nombre de phrases acceptées par erreur. Ceci qui va donner le Taux de Fausse Acceptation (TFA) (voir formule suivante).

$$TFA_{Locuteur} = \frac{\text{nombre de phrases acceptées par erreur dans le cas de ce locuteur}}{\text{nombre total des phrases} - \text{celles prononcées par ce locuteur}} \qquad (7.7)$$

*Dans notre cas : Le nombre total de phrases pendant le test est 5 * 37 = 185 et le nombre total de phrases pendant l'apprentissage est 10 * 37 = 370.*

Durant nos expériences de VAL, nous avons utilisé plusieurs types de paramètres d'entrée (du réseau) : nous avons utilisé le vecteur moyenne, la diagonale de la covariance et les deux vecteurs propres de la covariance, séparément. Par ailleurs, le nombre de neurones cachés a aussi été varié dans le but de trouver la meilleure configuration possible.

VII.5.1. 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 comme entrée du réseau et en variant le nombre de neurones cachés de 20 à 35 neurones.

Tableau 7.1 : 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 7.1, 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 car les meilleurs taux ont été obtenus pour des configurations différentes : pour le TBA c'est en utilisant 20 neurones cachés et pour TFA c'est 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.5.2. 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 comme entrée du réseau et en variant le nombre de neurones cachés de 20 à 35 neurones.

Tableau 7.2 : 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 7.2, 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. De même, le taux moyen de fausse acceptation varie entre 25,54% et 27,81% dont le meilleur taux est obtenue en utilisant 35 neurones cachés.

On peut conclure que pour ce type d'entrée - c'est à dire 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 (selon les conditions d'expérience).

VII.5.3. 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 comme entrée du réseau et en variant le nombre de neurones cachés de 20 à 35 neurones.

Tableau 7.3 : 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 7.3, 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. De même, le taux moyen de fausse acceptation varie entre 11.70% et 25.91% dont le meilleur taux est obtenue en utilisant 35 neurones cachés.

On peut conclure que pour ce type d'entrée - c'est à dire en utilisant le vecteur moyenne comme paramètre d'entrée du réseau - les meilleurs résultats sont obtenus en utilisant 35 neurones (selon les conditions d'expérience).

VII.6. Discussion

Après avoir fait modifier les paramètres d'entrées des réseaux (les deux premiers vecteurs propres, la diagonale de la covariance et le vecteur moyenne) et après avoir fait varier le nombre de neurones cachés on a aboutit aux résultats suivants (voir figure 7.8).

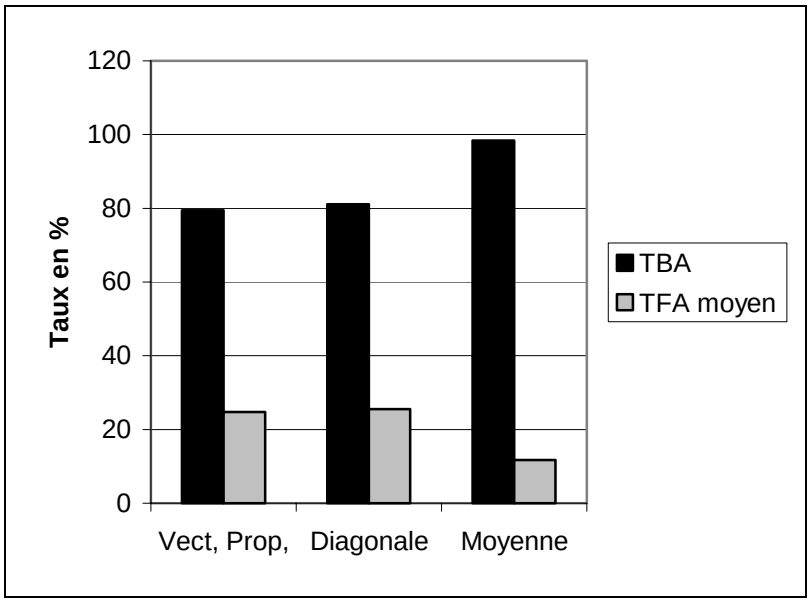


Fig. 7.8 : Taux de Bonne Acceptation (TBA) et Taux de Fausse Acceptation moyen (TFA_{moy}).

- 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% ; cela 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% toujours avec 35 neurones cachés.

- Pour le vecteur moyenne : on a obtenu un taux de bonne acceptation égal à 98,38% et un taux moyen de fausse acceptation égal à 11,70% ; cela en utilisant toujours 35 neurones cachés.

D'après ces chiffres on voit que les meilleurs scores sont obtenus en utilisant le vecteur moyenne, ce qui implique que ce dernier est plus précis 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 : théoriquement la performance qui pourrait en découler ainsi serait nettement meilleure.
Concernant la dimension de la couche cachée, nous pouvons conclure que les meilleures performances sont obtenues avec des dimensions élevées et que la performance sera d'autant plus grande que le nombre des neurones cachées est important, mais ceci au prix d'une complexité plus grande et un temps d'apprentissage nettement plus important.

VII.7. Conclusion préliminaire

En vérification (VAL), il semble logique de pénaliser l'erreur de fausse acceptation plutôt que l'erreur de faux rejet. Comme exemple de fausse acceptation on peut citer un voleur qui arriverait à accéder à un domicile étranger via un verrou vocal. Et comme exemple de faux rejet, on peut citer un propriétaire qui n'arrive pas à accéder à son domicile personnel lorsqu'il utilise son verrou vocal. Nous voyons bien que le premier cas d'exemple est nettement plus grave, même si le deuxième exemple est assez gênant.
En effet, la vérification automatique du locuteur doit avant tout réaliser une protection du système dont elle commande l'accès.

Dans notre étude, il ressort que la configuration utilisant le vecteur moyenne permet d'obtenir un taux d'erreur (erreur de reconnaissance) 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 (11.70%), même si l'erreur de reconnaissance est très faible. Pour cela on a pensé à améliorer encore ce résultat en jouant sur la couche de sortie (section suivante).

VII.8. Effet de la Variation du seuil d'arrondissement (cas du vecteur moyenne)

VII.8.1. Introduction

Cette section est consacrée à l'amélioration des résultats obtenus précédemment, afin d'arriver à un meilleur fonctionnement de notre système de vérification.

Pour ce faire, on a fait varier le seuil d'arrondissement (*le seuil d'arrondissement est le seuil avec lequel la sortie analogique du réseau de neurone est discrétisée soit à 0 ou bien à 1*), au niveau de la couche de sortie, et on a refait l'apprentissage des réseaux en utilisant le vecteur moyenne comme vecteur d'entrée.

VII.8.2. Paramètres initiales d'apprentissage

L'apprentissage est initialisé avec les paramètres suivants :

- Paramètres d'entrée : vecteur moyenne.

- Pas d'apprentissage : 0.2.
- Erreur d'arrêt désirée : 10^{-3}.
- Nombre d'itérations maximal : 100.000.
- Nombre de neurones cachés : 35.
- Plage de variation du seuil d'arrondissement : Sa $\in$ [0.5 0.95].

Rappelons que la structure du réseau est la même que précédemment, ainsi que la base de données d'apprentissage et de test.

VII.8.3 Résultats obtenus et interprétation

De nouvelles expériences de VAL ont été faites en faisant varier le seuil d'arrondissement, tout en observant l'évolution de la performance du réseau. Les résultats obtenus sont exposés dans le tableau 7.4 et le graphe 7.9 suivants.

Tableau 7.4 : Taux de vérification obtenus en variant le seuil d'arrondissement Sa.

	Sa = 0.5	Sa = 0.6	Sa = 0.7	Sa = 0.8	Sa = 0.9	Sa =0.95
Taux de bonne acceptation	98.38 %	98.38 %	98.38 %	98.38 %	98.38 %	93.51 %
Taux de faux rejet	1.62 %	1.62 %	1.62 %	1.62 %	1.62 %	6.49 %
Taux moyen de fausse acceptation	11.70 %	10.24 %	8.83 %	7.70 %	6.02 %	4.63%

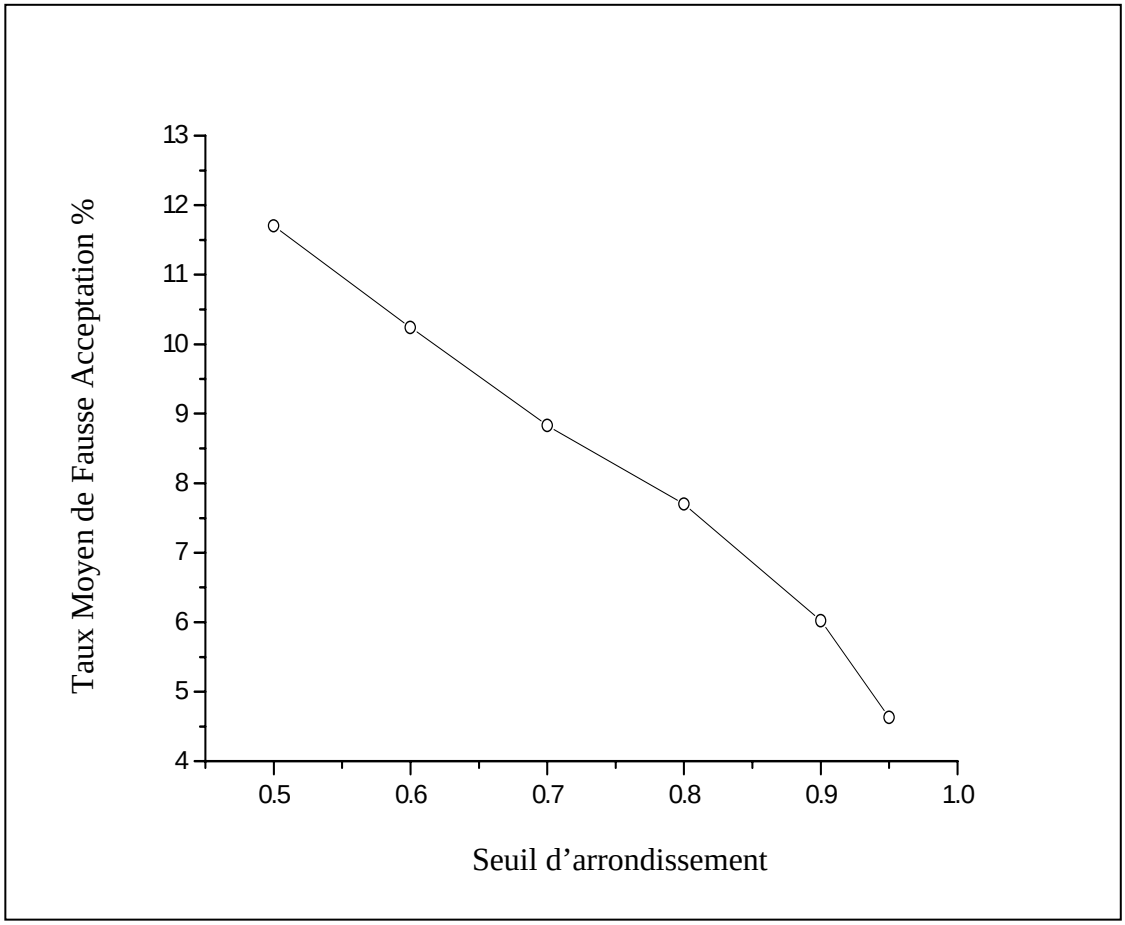


Fig. 7.9 : Evolution du Taux de Fausse Acceptation en fonction du seuil d'arrondissement.

D'après le tableau 7.4 et la figure 7.9 on peut déduire qu'en faisant varier le seuil d'arrondissement de 0,5 jusqu'à 0,95 le système évolue vers un état plus sécurisé.
Nous remarquons que pour les 5 premiers cas, c'est à dire qu'en faisant varier le seuil d'arrondissement de 0,5 à 0,9 le taux de bonne acceptation reste constant (98,38%) et le taux moyen de fausse acceptation diminue de 11,70% jusqu'à 6,02%. Ceci implique que le meilleur seuil, vu ce compromis, est de 0.9.
En augmentant, légèrement encore, le seuil d'arrondissement - c'est à dire jusqu'à 0,95 - le taux moyen de fausse acceptation diminue jusqu'à 4,63% (ce résultat est intéressant car il renforce la sécurité du système de vérification), tandis que le taux de bonne acceptation chute de 98,38% à 93,51% (ce qui n'est pas très gênant, car en vérification il est logique de pénaliser l'erreur de fausse acceptation plutôt que l'erreur de faux rejet).

L'explication de ce résultat vient du fait qu'en augmentant le seuil d'arrondissement, le système de vérification arrive à mieux fermer le champ d'acceptation, comme le montre la figure 7.10 suivante, en réduisant ainsi l'erreur de fausse acceptation.

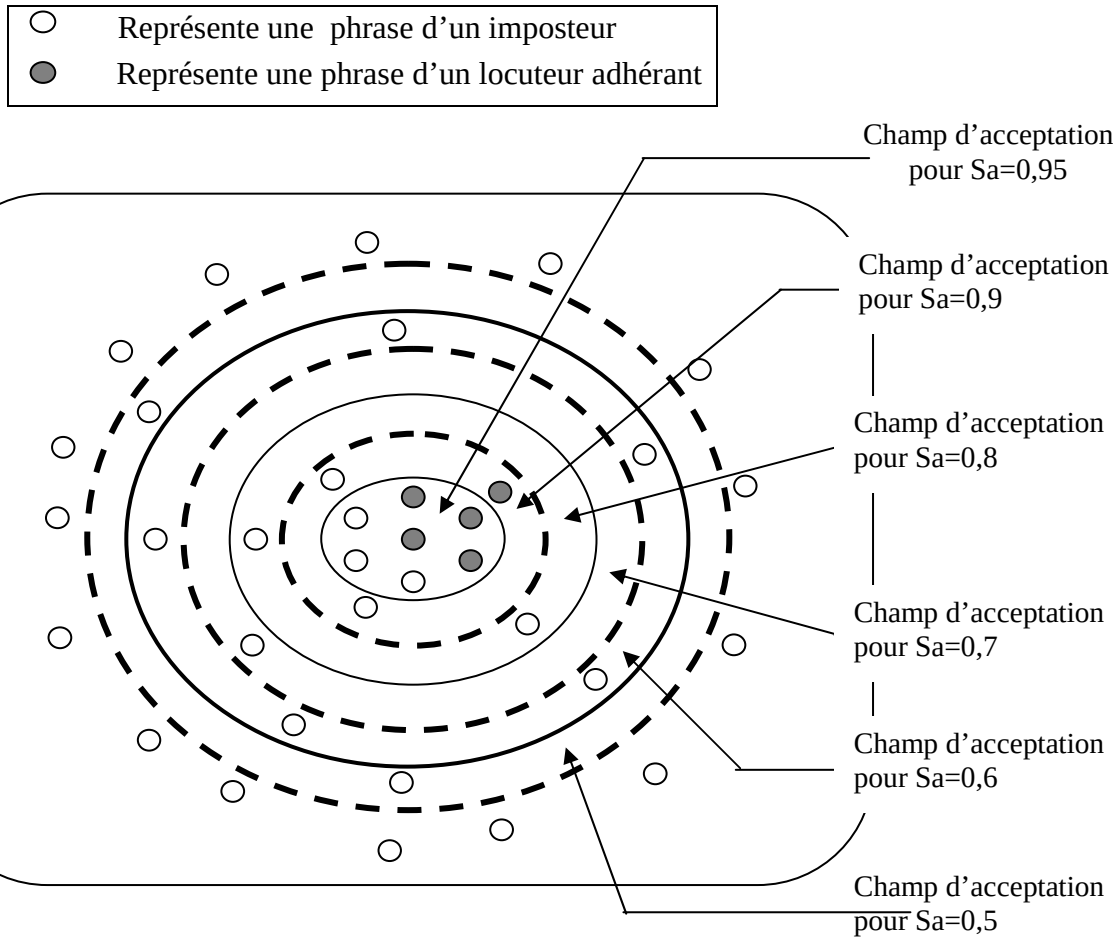


Figure 6.10 : Variation du champ d'acceptation en fonction du seuil d'arrondissement.

Cette figure représente le nombre de phrases acceptées en fonction du seuil d'arrondissement du locuteur (cet exemple correspond à un traitement de 35 phrases).

Pour mieux illustrer aussi ce phénomène, on peut donner l'exemple pratique suivant :

Supposons que l'on présente une phrase du locuteur 6, par exemple, à l'entrée du réseau appartenant au locuteur 11 (qui comportera, en cas d'acceptation, la sortie désirée :

$$T = (0\ 0\ 1\ 0\ 1\ 1)^t)$$

et supposons, par exemple, que la sortie (réelle) ainsi calculée par le réseau est :

$$(0.4\ \ 0.3\ \ 0.51\ \ 0.35\ \ 0.7\ \ 0.57)^t,$$

- alors si (par exemple) le seuil d'arrondissement Sa est fixé à 0.5, on aura comme sortie finale :

  $$(0\ 0\ 1\ 0\ 1\ 1)^t,$$

  donc ce réseau va accepter le locuteur 6 comme étant le locuteur 11. Ce qui correspond à une fausse acceptation.

- mais si le seuil d'arrondissement Sa est fixé à 0.6, on aura comme sortie finale :

  $$(0\ 0\ 0\ 0\ 1\ 0)^t,$$

  et ainsi donc, le réseau n'acceptera pas le locuteur 6 ; il le rejettera par conséquent.

Cet exemple élémentaire a montré que le fait d'augmenter le seuil d'arrondissement (de 0.5 à 0.6) a bien permis d'améliorer la sécurisation du système (dans ce cas).

VII.8.4. Conclusion

Dans cette nouvelle étape, nous avons essayé d'améliorer la performance du réseau de neurones en jouants sur les paramètres de sortie de ce dernier. Cette nouvelle étude nous a permis de montrer que le seuil d'arrondissement a un grand effet sur les performances des systèmes de vérification du locuteur (VAL).
Ainsi, nous avons pu constater, dans le cas de notre modèle, qu'il existe un seuil d'arrondissement critique qui nous permet d'assurer une bonne sécurisation du système de vérification. Ce seuil doit être très grand pour permettre d'atteindre un faible TFA (assurant une bonne protection contre les impostures). Cependant, ce seuil critique est majoré aussi par le taux de bonne acceptation qui présente une limite maximale définie par le TBA (assurant une bonne identification des locuteurs), car à partir de cette limite la probabilité qu'un locuteur soit mal reconnu augmente. Un tel compromis impose un seuil optimal critique (*Sa = 0.9 dans notre cas*) assurant, en même temps, aussi bien la sécurité que la bonne reconnaissance des systèmes de VAL.

Par conséquent, le choix des paramètres et des seuils initiaux du réseau possède un rôle très important en vérification automatique du locuteur. Il s'en suit qu'un choix judicieux doit être fait avant toute tentative de VAL.

Finalement, nous avons pu concevoir durant cette étude un système de VAL à base de MLP mono-locuteur précis et assez sécurisé, présentant un taux de bonne vérification pouvant atteindre 98.38% et un taux moyen de fausse acceptation pouvant atteindre (dans les mêmes conditions) 6.02%. Nous pensons que ces taux sont intéressants et que ce système de VAL peut encore être amélioré si on rajoute d'autres types de caractéristiques à l'entrée du réseau.
Cependant, le seul inconvénient que nous évoquons ici, est la complexité du réseau et le nombre important d'exemples nécessaires au bon apprentissage de ce dernier.

- Conclusion Générale -

Dans le cadre de cette thèse, nous avons abordé un domaine de recherche non encore maîtrisé : La Reconnaissance Automatique du Locuteur (RAL). 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 mis en place plusieurs techniques pour la reconnaissance automatique du locuteur. Nous en citons :

- Une méthode statistique basée sur la SOSM pour l'Identification du locuteur : IAL.
- Trois méthodes connexionnistes basée sur la LVQ pour l'Identification du locuteur : IAL.
- Une méthode connexionniste basée sur le MLP pour la Vérification du locuteur : VAL.

Des tests expérimentaux ont été faits sur ces méthodes et nous avons pu noter que :

- La méthode statistique, basée sur les mesures de similarité du 2nd ordre, nécessite peu d'apprentissage. Cette méthode testée, pour l'identification automatique de 37 locuteurs différents, a donné d'excellents résultats (performance 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 l'identification de 11 locuteurs différents. 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 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%.

Par ailleurs, nous avons rajouté de nouvelles expériences, 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.
- Recherche de la résolution spectrale optimale pour la méthode statistique.
- Mise au point d'une nouvelle métrique adaptée aux caractéristiques hétérogènes pour les méthodes à base de LVQ.
- Conception d'une nouvelle méthode prosodique pour l'IAL à base de LVQ.
- Proposition d'un réseau de neurone mono-locuteur à base de MLP pour la VAL.
- Amélioration de la performance du MLP en jouant sur ses paramètres de sortie.
- Mise en place d'un ensemble d'indications comparatives servant de conseils pour l'utilisation de ces méthodes en RAL.

Comparaison et Perspectives…

Il est difficile de comparer les différentes techniques proposées en reconnaissance du locuteur car les conditions générales de reconnaissance ne sont pas tout à fait les mêmes. Néanmoins, on peut citer certaines propriétés (avantages et inconvénients) de ces techniques.

Ainsi, leurs performances comparatives sont données dans le tableau ci-dessous.

Nous tenons à rappeler, seulement, que les 3 premières approches, affichées sur ce tableau, ont été testées durant le travail de recherche de cette thèse, tandis que les autres méthodes (GMM et HMM) sont des méthodes récemment testées par d'autres chercheurs du domaine.

Tableau illustrant les performances comparatives de certaines méthodes en RAL.

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

Du point de vue complexité et temps d'apprentissage, la SOSM demeure actuellement la méthode la moins complexe et la plus rapide pouvant être utilisée en RAL.

Du point de vue indépendance du texte, toutes les méthodes le sont sauf la LVQ (comme décrite dans nos expériences) 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 rajouté à l'ensemble des adhérents, le système n'est fonctionnel que si l'on refasse tout l'apprentissage). Les méthodes statistiques, par contre, sont évolutives, d'où un autre avantage de ces techniques.

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 récentes, dans ce sens, ont montré que cela est possible et qu'ainsi le taux de reconnaissance est bien amélioré.

Toutefois, l'amélioration apportée en performance 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 !

Quant aux approches statistiques, cette étude a bien montré leur efficacité, tant en milieu sain qu'en milieu corrompu (bruité ou téléphonique). En tout cas, 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.

- Introduction -

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 principal est de reconnaître une personne par l'analyse de sa voix.

Cette 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 (souvent simulé par un réseau de neurones) qui permettent de reconnaître simultanément ce qui est dit et celui qui parle…

Cette faculté de reconnaître un locuteur qui parle (par le biais du système nerveux), nous rend beaucoup de services en pratique : Par exemple, on n'est pas obligé de demander à un interlocuteur de s'identifier, quand celui-ci ne se situe pas dans notre champ de vision. Par suite, cela est très utile pour les personnes aveugles ainsi que pour les interlocuteurs engagés dans une conversation téléphonique ou encore pour les interlocuteurs masqués (un imposteur qui demanderait à quelqu'un de lui ouvrir la porte).
Notons aussi que cette faculté naturelle est très importante pour pouvoir suivre correctement plusieurs personnes qui parlent dans une conversation.

Le but actuel des chercheurs est de reconnaître automatiquement un locuteur par des techniques fiables du traitement du signal. Ainsi, la Reconnaissance Automatique du Locuteur (RAL) consiste à reconnaître par microphone, grâce à un système automatisé, un individu qui parle d'une manière naturelle.

Quels en sont les intérêts pratiques ?

La reconnaissance automatique du locuteur constitue un domaine de recherche de pointe comportant des applications très intéressantes. Globalement, on peut distinguer les applications suivantes :

- Contrôle d'accès à des processus physiques :
  - Zones sécurisées.
  - Protection industrielle.
- Accès privilégié à des informations :
  - Centres de renseignements.

  - Banques de données.

- Validation des transactions et opérations bancaires :
  - Transactions bancaires par téléphone.
- Affectation des tâches nominatives et confidentielles :
  - Utilisation d'appareils ne devant servir qu'à une seule personne.
- Applications domestiques :
  - Protection de domiciles, garages et téléphones par verrou électronique.
- Amélioration de la reconnaissance de la parole :
  - Adaptation d'un système de reconnaissance de la parole à plusieurs locuteurs.
- Téléconférence :
  - Accès réservé aux membres de la réunion.
  - Gestion et surveillance des canaux de communication.
- Criminologie :
  - Recherche de suspects : nous citons un exemple réel qui pourra concrétiser cet intérêt ; il s'agit du meurtre d'un ministre belge, il y' a quelques années de cela, où la RAL a pu prouver l'innocence d'un accusé après avoir comparé sa voix à celle du véritable meurtrier (enregistrements téléphoniques incompatibles).
- Détermination de l'état émotionnel du locuteur.
- Multimédia :
  - Accès à Internet et aux Emails.
- Télévision :
  - Recherche automatique de chaînes par identification des journalistes.
  - Affichage automatique des noms des journalistes.

Certes, les domaines d'intérêt de la RAL sont très nombreux (nous ne pouvons citer toutes ces applications ici), mais en pratique les systèmes proposés soufrent de quelques défaillances réelles.

Description de la RAL

Théoriquement, la reconnaissance du locuteur est un terme générique pour discriminer parmi plusieurs personnes en fonction de leurs voix. Les variations individuelles entre locuteurs ont deux origines essentielles. En premier lieu, les caractéristiques physiques de l'appareil de phonation influencent les formants, 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 et dans l'étendue des variations du pitch.

En RAL, on distingue deux spécialités : 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 (généralement) 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 la référence proclamée.

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).

Approches étudiées

Dans le cadre de cette thèse, nous avons abordé un domaine de recherche non encore maîtrisé. Nous avons ainsi proposé plusieurs types de méthodes de RAL émanant de deux visions différentes :

- une vision statistique

et

- une vision connexionniste.

Les méthodes élaborées sont les suivantes :

- Une méthode statistique :
  - basée sur la SOSM pour l'Identification du locuteur
- Deux méthodes connexionnistes ; il s'agit de :
  - trois techniques basée sur la LVQ pour l'Identification du locuteur
  - une technique basée sur le MLP pour la Vérification du locuteur

Comment est organisée la thèse ?

Pour ce qui est de la thèse, nous l'avons organisée de la manière suivante :

Le chapitre II décrit brièvement la problématique et l'état de l'art en reconnaissance du locuteur. Il explique, aussi, en détail les différentes spécialités existant en RAL.

Le chapitre III illustre l'approche statistique proposée pour l'identification automatique du locuteur.

Le chapitre IV introduit l'approche connexionniste et décrit les méthodes proposées pour l'identification automatique du locuteur.

Le chapitre V décrit les expériences d'identification IAL effectuées par la méthode SOSM..

Le chapitre VI décrit les expériences d'identification IAL effectuées par la méthode LVQ.

Le chapitre VII décrit les expériences de vérification VAL effectuées par la méthode MLP.

Des discussions relatives à ce travail de recherche, ainsi que diverses conclusions sont apportées au chapitre *« Conclusion »*.

Et finalement, des références bibliographiques ainsi que des annexes utiles sont mises à la disposition du lecteur (à la fin du manuscrit).

Numéro d'ordre : 02/2003 - E/EL

UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE HOUARI BOUMEDIENE

FACULTE D'ELECTRONIQUE ET D'INFORMATIQUE

Laboratoire de la Communication Parlée et du Traitement du Signal

---

Thèse présentée pour obtenir le GRADE de DOCTORAT d'état en électronique

Spécialité : S.S.T.R.

par : **Mr SAYOUD Halim**

Sujet :

Reconnaissance Automatique du Locuteur

-Approche Connexionniste-

Soutenue le 29.11.2003 devant le jury composé de :

▪ Mr ATTARI Mokhtar	Professeur	Président
▪ Mme BOUDRAA Malika	Maître de Conférences	Directrice de thèse
▪ Mr DJERADI Amar	Maître de Conférences	Examinateur
▪ Mr SMARA Youcef	Maître de Conférences	Examinateur
▪ Mme BELHADJ Aichouche	Maître de Conférences	Examinateur
▪ Mr BELOUCHRANI Adel	Maître de Conférences	Examinateur

Préambule au chapitre suivant

Après avoir introduit le domaine de la Reconnaissance Automatique du Locuteur, nous passerons au chapitre II, intitulé : « La Reconnaissance Automatique du Locuteur : Position du problème, Définition et Applications ».

Celui-ci expose en détail le domaine de la RAL en définissant les différentes spécialités qui en dépendent. Nous y trouveront également, l'état de l'art du domaine et les différentes tendances qui ont émergé ces dernières années.

Nous vous souhaitons une bonne lecture…

Préambule au chapitre suivant

Après avoir clarifié le domaine de la Reconnaissance Automatique du Locuteur, nous passerons au chapitre III, intitulé : « Approche Statistique».

Celui-ci expose en détail les techniques statistiques utilisées en reconnaissance du locuteur en donnant toutes les théories correspondantes.
L'approche statistique est l'approche de base en RAL , car la plupart des autres approches en dépendent.
Depuis que les modèles mono-gaussiens aient été découverts par F. Bimbot en 1995, les chercheurs n'ont guère arrêté de se pencher vers cette approche. Actuellement, les modèles statistiques GMM fondés par Reynolds sont considérés comme les techniques de pointe en RAL.

Nous vous souhaitons une bonne lecture…

Préambule au chapitre suivant

Après avoir exposé la théorie de l'approche statistique et décrit quelques méthodes d'identification statistique, nous passerons au chapitre IV, intitulé : « Approche Connexionniste».

Les réseaux de neurones ont été beaucoup utilisés en reconnaissance de la parole. Nous avons voulu voir leur performance en RAL.

Depuis plusieurs années, les neurobiologistes (en collaboration avec des mathématiciens) ont tenté de mettre au point un modèle mathématique de neurones qui simule assez bien le cerveau, pour mieux comprendre certains fonctionnements du système nerveux.

Ainsi plusieurs types de réseaux de neurones existent ; certains réseaux bien adaptés à la classification non linéaire (eg. RAL) ont, alors, été sélectionnés.

Nous vous souhaitons une bonne lecture…

Préambule au chapitre suivant

Après avoir vu, en détail, les différentes approches dans les chapitre III et IV, nous avons entamé 3 séries d'expériences en reconnaissance automatique du locuteur.

Dans ce chapitre IV, nous décrivons la 1ère série d'expérience faite ; il s'agit de : «L' Identification du Locuteur par la Méthode SOSM -Recherche de la Résolution Spectrale Optimale».

L'originalité de ces expériences tient à l'étude qui est faite dans un environnement très difficile et qui est analysée sous plusieurs paramètres de variabilité : résolution spectrale, bande passante, type de bruit, RSB, …

Les résultats obtenus dans ces expériences sont très intéressants et montrent clairement les hautes performances de la technique SOSM.

Nous vous souhaitons une bonne lecture…

Préambule au chapitre suivant

Après avoir observé les différents résultats d'identification (par la méthode SOSM) au chapitre V, nous décrivons à présent la 2ème série d'expérience faite ; il s'agit de « L'Identification du Locuteur par la Méthode connexionniste LVQ - avec Proposition d'une Nouvelle Métrique Hybride».

Cependant, dans cette méthode nous nous heurtons souvent au problème d'incompatibilité des paramètres (caractéristiques acoustiques), dans le cas où ceux-ci ne sont pas homogènes. Pour contourner ce problème, nous avons proposé une nouvelle métrique associative dans le cas des paramètres acoustiques de natures différentes. Celle-ci nous a permis d'associer des paramètres hétérogènes dans le calcul d'une distance interlocuteur.

Ainsi des caractéristiques très significatifs (telles que Fo) ont pu être associées aux paramètres spectraux (ou cepstraux) classiques, en rehaussant considérablement le taux de bonne reconnaissance trouvé par LVQ.

Les différents résultats obtenus dans cette méthode connexionniste sont exposés et critiqués dans ce chapitre VI.

Nous vous souhaitons une bonne lecture…

Préambule au chapitre suivant

Après avoir observé les différents résultats d'identification (par la méthode LVQ) au chapitre VI, nous décrivons à présent la 3ème série d'expérience faite ; il s'agit de « La Vérification Automatique du Locuteur par MLP».

La vérification du locuteur est basée sur l'identification du locuteur ; à laquelle on rajoute une proclamation d'identité dans le but de l'accepter ou de la refuser.

L'objectif de cette étude (connexionniste) est la réalisation d'un système de vérification automatique du locuteur (VAL) en utilisant des réseaux de neurones mono-locuteur à base de MLP (Multi-Layer Perceptron).

L'originalité de ces réseaux est qu'ils sont du type mono-locuteur : chaque locuteur adhérent possède un réseau de neurone propre à lui.

Les différents résultats obtenus dans cette 2ème méthode connexionniste sont exposés et critiqués dans ce chapitre VII.

Nous vous souhaitons une bonne lecture…

Préambule au chapitre suivant

Après avoir observé les différents résultats d'identification et de vérification aux chapitres V, VI et VII, nous essayerons de synthétiser une série de constats et de conclusions dans le prochain chapitre intitulé « Conclusion Générale ».

Ce chapitre est important, car il résume tous les travaux effectués durant cette thèse en offrant une certaine comparaison objective sur les performances des différentes approches étudiées.

Nous espérons que cette conclusion générale sera utile aux chercheurs intéressés par la RAL et qu'elle apportera un PLUS à ce domaine de recherche.

Nous vous souhaitons une bonne lecture…

- Références bibliographiques -

Références

[Art 93] T. Artieres, P. Gallinari, 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] B. Atal, Automatic recognition of speakers from their voices. IEEE transactions, volume 64 (4), pages 460-475, 1976.

[Auc 00] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, 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] T. Banziger, G. Klasmeyer, T. Johnstone, T. Kamceva, K. Scherer, 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] Y. Bennani, « Approches Connexionnistes Pour La Reconnaissance Automatique Du Locuteur : Modelisation Et Identification ». Thèse de Doctorat de l'université de Paris Sud.1992.

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

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

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

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

[Bim 98] F. Bimbot, H. Hutter, C. Jaboulet, J. Koolwaaij, J. Lindberg, J. Pierrot, 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] F. Bimbot, I. Magrin-Chagnolleau, L. Mathan, Second-order statistical measures for text-independent speaker identification. Speech Communication, volume 17 (1-2), pages 177-192, Août 1995.

[Bim 92] F. Bimbot, L. Mathan, A. De-Lima, G. Chollet, 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] L. Boe, 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] L. Boe, F. Bimbot, J. Bonastre, P. Dupont, De l'évaluation des systémes de vérification du locuteur à la mise en cause des expertises vocales en identification juridique. Revue Langues, volume 2 (4), Décembre 1999.

[Bon 97] J. Bonastre, L. Besacier , 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] J. Bonastre, P. Delacourt, C. Fredouille, S. Meignier, T. Merlin, C. Wellekens, Difiérentes stratégies pour le suivi du locuteur. Reconnaissance des Formes et Intelligence Artificielle (RFIA), pages 123-129, 2000 Paris (France).

[Bon 00b] J. Bonastre, P. Delacourt, C. Fredouille, T. Merlin, C. Wellekens, 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] I. Booth, M. Barlow, B. Watson, Enhancements to DTW and VQ decision algorithms for speaker recognition. Speech Communication, volume 13 (3-4), pages 427-433, Décembre 1993.

[Bov 98] L. Boves, 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] C. Champod, D. Meuwly, M. Weintraub, K. Sonmez, 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] D. Charlet, 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, “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] E. Davalo, P. 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] P. Delacourt, La segmentation et le regroupement par locuteurs pour l'indexation de documents audio. Thése de doctorat, Institut Eurecom, 2000, Nice (France).

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

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

[Dod 98] G. Doddington, 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, “An acoustic-phontic database”, JASA, suppl. A, Vol. 81 (S92) 1986.

[Fis 99] L. Fissore, F. Ravera, C. Vair, 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] A. Freeman, D. Skapura, « Neural Networks ». Algorithms, Applications and programming techniques. 1992.

[Fre, 94] S. Frederickson, L. Tarassenko, Radial basis functions for speaker identification. Workshop on Automatic Speaker Recognition, Identification, Verification, pages 107-110, Avril 1994, Martigny (Suisse).

[Fre 98] C. Fredouille, J. Bonastre , 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] C. Fredouille, J. Bonastre, T. Merlin, 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), 2000.

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

[Fur 77] S. Furui, 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] S. Furui, 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] S. Furui, An overview of speaker recognition technology. Workshop on Automatic Speaker Recognition, Identification, Verification, pages 1-9, Avril 1994, Martigny (Suisse).

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

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

[Gis 86] H. Gish, M. Krasner, W. Russel, J. Wolf, 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] Y. Grenier, 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).

[Gre 80] Y. Grenier, 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] C. Grifin, T. Matsui, S. Furui, 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] H. Hattori, 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] J. Hérault, C. Jutten, « Reseaux Neuronaux Et Traitement Du Signal ». Editions Hérmés, Paris, 1994.

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

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

[Hom 94] M. Homayounpour, G. Chollet, 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] B. Jacob, J. Mariéthoz, G. Gravier, F. Bimbot, 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] S. Johnson, Who spoke when ? - automatic segmentation and clustering for determining speaker turns. European Conference on Speech Communication and Technology (Eurospeech), Septembre 1999, Budapest (Hongrie).

[Kar 98] I. Karlsson, T. Banziger, J. Dankovicov, T. Johnstone, J. Lindberg, H. Melin, F. Nolan, K. Scherer, 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] J. Kharroubi, G. Chollet, 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] T. Kohonen, Self-Organization and Associative Memory. Springer Series in Information Sciences 8, Springer Verlag, New York, 1984.

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

[Kon 98] Y. Konig, L. Heck, M. Weintraub, K. Sonmez, 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] H. Kunzel, 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] C. Leggetter, P. Woodland, Maximum Likelihhod Linear Regression for speaker adaptation of continuous Hidden Markov Models. Computer Speech and Language, volume 9, pages 171-185, 1995.

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

[Mag 99] I. Magrin-Chagnolleau, G. Durou, 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] I. Magrin-Chagnolleau, J. Wilke, F. Bimbot, 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] A. Martin, M. Przybocki, 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] A. Martin, M. Przybocki, The DET curve in assessment of detection task performance. European Conference on Speech Communication and Technology (Eurospeech), pages 1895-1898, Septembre 1997, Rhôdes (Gréce).

[Mas 89] J. Mason, J. Oglesby, L. Xu, 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] T. Matsui, S. Furui, 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, 1994, Adélaide (Australie).

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

[Mon 92] C. Montacié, P. Deléglise, F. Bimbot, M. Caraty, 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).

[Nai 94] J. Naik, 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] J. Oglesby, What's in a number ? : moving beyond the Equal Error Rate. Speech Communication, volume 17 (1-2), pages 193-209, Août 1995.

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

[Ogl 91] J. Oglesby, J. Mason, 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] D. O'Shaughnessy, Speaker recognition. IEEE Transactions Acoustics, Speech, and Signal Processing (ASSP), pages 4-17, Octobre 1986.

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

[Pao 96] A. Paoloni, S. Ragazzini, G. Ravaioli, 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] J. Pierrot, Elaboration et validation d'approches en vérification du locu-teur. Thése de doctorat, Ecole Nationale Supérieure des Télécommunications (ENST), 1998, Paris (France).

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

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

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

[Rab 89] L. Rabiner, 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] D. Reynolds, Gaussian mixture modeling approach to text-independent speaker identification. Thése de doctorat, Georgia Institute of Technology, 1992, (USA).

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

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

[Rey 96] D. Reynolds, 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] D. Reynolds, Comparison of background normalization methods for text-independent speaker verification. European Conference on Speech Communication and Technology (Eurospeech), Septembre 1997, Rhôdes (Gréce).

[Rey 00] D. Reynolds, 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] A. Rosenberg, Automatic speaker verification, a review. Proceedings IEEE, volume 64 (4), pages 475-487, 1976.

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

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

[Say 94] H. Sayoud, M. Boudraa, B. Guerin, « Etude comparative de plusieurs détecteurs de F0 », pp IV-32à 36 , ICSS'1994 Alger 24-26 sept 1994.

[Say 94a] H. Sayoud, M. Boudraa, B. Guerin, « PDA à Ambiguïté Modifiée », pp IV-37 à 40 ICSS'1994, Alger 24-26 sept 1994.

[Say 95] H. Sayoud, M. Boudraa, B. Guerin, « Détecteurs de Pitch à convergence de fréquence » MCEA'1995, Grenoble septembre 1995.

[Say 96] H. Sayoud, M. Boudraa, B. Guerin, « Interprétation des erreurs de PDA » JTEA'1996, Tunis 8-9 nov. 1996.

[Say 97] H. Sayoud, M. Boudraa, B. Guerin, 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, M. K. Selmane, "Error Correction Algorithmses for PDAs", CESA'98 Tunisia April 1-4 1998, pp 337-341, Volume 4, 1998.

[Say 98a] H. Sayoud, M. K. Selmane, " Méthodes comparatives en reconnaisance du locuteur ", JTEA'98 Nabeul, Tunis, Tunisie 6 et 7 Nov. 1998.

[Say 98b] H. Sayoud, M. Boudraa, B. Guerin, « Post-Processing for PDA's », SSST'1998, Virginia 8-10 Mars 1998.

[Say 98c] H. Sayoud, M. K. Selmane, On the use of statistical ratio for classification in automatic speaker recognition, MCEA, Marrakech 17-19 sept 1998.

[Say 98d] H. Sayoud, M. K. Selmane, 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 98e] H. Sayoud, M. K. Selmane, Tests comparatifs de differentes methodes en identification du locuteur, NWSIP'98, 2 dec 1998, Sidi-bel- abbas, Algerie.

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

[Say 99a] H. Sayoud M. K. Selmane, Tentatives en reconnaissance du locuteur, SSA2'99, Blida 10-12 Mai 99, Algerie.

[Say 00] H. Sayoud, S. Ouamour, N. Kernouat, M. K. Selmane, Reconnaissance Automatique du Locuteur en Milieu Bruité. Juin 2000, pp. 345-348. JEP'2000, Aussois France.

[Say 01] H. Sayoud, S. Ouamour, M. Boudraa, Discrimination Parole / Non Parole en suivi de locuteur. 8-11 Septembre 2001, pp. : 130-132. RJC'2001 Mons Belgique.

[Say 01a] H. Sayoud, S. Ouamour, M. Boudraa, Suivi de locuteur par la statistique d'ordre 2. 8-11 Septembre 2001, pp.: 134-135. RJC'2001 Mons Belgique.

[Say 02] H. Sayoud, S. Ouamour, M. Boudraa, 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.

[Say 02a] H. Sayoud, S. Ouamour, M. Boudraa, Indexation des Documents Audio en vue d'un Filtrage par Locuteur. 24-27 juin 2002, pp 141-148. TALN'2002, Nancy France.

[Say 02b] H. Sayoud, S. Ouamour, M. Boudraa, Automatic Speaker Indexing in Corrupted Speech. 8-12 September 2002, Brésil, pp 877-882. ITS'2002, Natal Brésil.

[Say 02c] H. Sayoud, S. Ouamour, M. Boudraa, A Robust Method for Speaker Tracking. 27-29 Août 2002, Egypte, pp 187-192. ICCTA'2002, Alexandria Egypt.

[Say 02d] H. Sayoud, S. Ouamour, M. Boudraa, Speaker Indexing in a noisy environment. -Investigation of 3 types of noise-

[Say 03] H. Sayoud, S. Ouamour, M. Boudraa, Application des Mesures Statistiques pour la détection des points de changement des locuteurs. AMAM'2003, February 10-13, 2003 in Nice, France.

[Say 03a] H. Sayoud, S. Ouamour, M. Boudraa, Application of Statistical Measures in Speaker Identification. AMAM'2003, February 10-13, 2003 in Nice, France.

[Say 03b] H. Sayoud, S. Ouamour, M. Boudraa, Application of the MLVQ1 in Speaker Identification. NOLISP'03, 20-23 Mai 2003. Le Croisic, France.

[Say 03c] H. Sayoud, S. Ouamour, M. Boudraa, '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] K. Scherer, T. Johnstone, J. Sangsue, 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).

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

[Soo 88] F. Soong, A. Rosenberg, 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] F. Soong, A. Rosenberg, L. Rabiner, B. Juang, A vector quantization approach to speaker recognition. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 387-390, 1992, Tampa (USA).

[Van 96] S. Van-Vuuren, 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] K. Yu, J. Mason, J. Oglesby, Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation. IEE vision, image and signal processing, 1995, Berlin (Allemagne).

[Zaa 00] I. Zaaf, C. Smaili, « Systeme de reconnaissance de mots isoles basé sur deux approches : Reseaux de Neurones (ANN) et Programmation Dynamique (DTW) ». Université des sciences et de la technologie Houari Boumedienne. Institut d'informatique, 2000.