وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR-ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR-ANNABA

جامعة باجي مختار - عنابة

Année 2008

Faculté de l'Ingénieur

Département d'Informatique

MEMOIRE

Présenté en vue de l'obtention du diplôme de **Magistère**

THEME

CONCEPTION D'UN SYSTEME ACOUSTICO-ANATOMIQUE POUR L'IDENTIFICATION DU LOCUTEUR : ARCHITECTURE ET PARAMETRISATION

Option :

Texte, Image et Parole

Par :

M^me^ Feriel DEBBECHE-GUERID

Directrice de mémoire :	Nacira	GHOUALMI-ZINE	Maître de Conférences	U. ANNABA

Devant le jury :

Président :	Mohamed	BENMOHAMED	Professeur	U. CONSTANTINE
Examinateurs :	Abdallah	BOUKERAM	Maître de Conférences	U. SETIF
	Nora	BOUNOUR	Maître de Conférences	U. ANNABA

Résumé

De nos jours, l'identification des utilisateurs est devenue une véritable nécessité. Dans cette optique, l'identification des personnes sur la base de leur voix est une technique biométrique prometteuse et compétitive en même temps pour plusieurs raisons. Elle est, en particulier, le seul moyen biométrique pour de nombreux types d'application.

La voix présente cependant un vrai challenge pour les chercheurs, dû à sa variabilité intrinsèque et extrinsèque.

Ce travail s'inscrit dans le cadre général de la Reconnaissance Automatique du Locuteur, domaine très fertile. Plus particulièrement, l'Identification Automatique du Locuteur en mode indépendant du texte et dans le cas où nous disposons de très peu de données d'apprentissage.

L'approche de modélisation que nous utilisons est une approche intéressante qui consiste à représenter le Locuteur non plus de façon absolue mais relativement à un ensemble de Locuteurs bien appris. Chaque Locuteur est représenté par sa localisation dans un espace optimal de voix propres (Eigen Voices).

Dans ce projet, nous présentons un système d'Identification Automatique du Locuteur (IAL) où l'espace de représentation est construit par l'Analyse en Composantes Principales (ACP). La localisation des nouveaux Locuteurs dans cet espace est faite par projection orthogonale.

Notre système, dédié aux applications de très haute sécurité, repose essentiellement sur une idée originale qui consiste à exploiter, à côté des paramètres acoustiques pour la paramétrisation du Locuteur, des paramètres anatomiques.

Ainsi, nous proposons une architecture acoustico-anatomique pour l'identification du Locuteur. Cette architecture se base principalement sur un algorithme de fusion de données hétérogènes.

Etant donné que les données que nous fusionnons sont entachées d'imprécision et d'incertitude, nous utilisons les principes de la théorie des possibilités pour les fusionner.

Cette fusion qui nous ramène à une nouvelle paramétrisation du Locuteur, i. e. un vecteur acoustico-anatomique représentatif de ce dernier, a pour but de renforcer le taux d'identification et d'apporter une dimension biométrique au système.

Mots clés :

Identification Automatique du Locuteur, Biométrie, Voix Propres, Analyse en Composantes Principales, Théorie des Possibilités, Architecture Acoustico-Anatomique, Algorithme de fusion, Paramétrisation du Locuteur, Vecteur Acoustico-Anatomique, Sécurité.

ملخص

في وقتنا الحاضر أصبح التعرف على هوية المستعمل ضرورة حقيقية.
من هذه الزاوية، التعرف على الأشخاص إنطلاقا من صوتهم يعتبر تقنية بيومترية جد واعدة وتنافسية في نفس الوقت وذلك لعدة أسباب . إنها بالخصوص الطريقة البيومترية الوحيدة لأنواع عديدة من التطبيقات.
لكن الصوت يطرح تحدي حقيقي للباحثين، وذلك لتغيراته الداخلية والخارجية.
يندرج هذا العمل في إطارالتعرف الآلي على المتكلم وهو مجال خصب جدا .بالخصوص التعرف الآلي على المتكلم بالطريقة المنفصلة عن النص وفي حالة الحصول على معلومات ضئيلة حول التمرين.
منهجية النمذجة التي نستعملها هي طريقة مهمة تحتوي على تمثيل المتكلم، لا بطريقة مطلقة بل ارتباطا بعدد متمرن من المتكلمين .يتم تمثيل كل متكلم عن طريق تحديده في فضاء أمثل من أصوات إيقن.
في هذا المشروع نقدم نظام للتعرف الآلي على المتكلم أين يكون فضاء التمثيل مبنيا بطريقة التحليل الرئيسي للمكونات. يتم تحديد المتكلمين الجدد في هذا الفضاء عن طريق الإسقاط المتعامد.
نظامنا هذا الموجه الى تطبيقات ذات امن عالي يتركز أساسا على فكرة جديدة تتمثل في إستعمال عناصر تشريحية إلى جانب العناصر السمعية من أجل تمثيل المتكلم.
بما أن هذه الخصائص غير دقيقة وغير مضمونة نستعمل مبائ نظرية الإمكانية من أجل مزجهم. هذا المزج الذي يقودنا الى تمثيل جديد للمتكلم هدفه تقوية نسبة التعرف و إعطاء النظام بعد بيومتري.

كلمات مفاتيح :

التعرف الآلي على المتكلم، بيومترية، أصوات إيقن، تحليل المكونات الأساسية، نظرية الإمكانية، نظام سمعي تشريحي، خوارزم المزج، نمذجة المتكلم، عمود سمعي تشريحي، أمن.

Abstract

Nowadays, users' identification has become a real necessity. In this vision, people's identification on the basis of their voice is a very promising and competitive biometric technique at the same time for many reasons. It is in particular the only biometric means for numerous types of application. However, the voice presents a real challenge to researchers, due to its intrinsic and extrinsic variability.

This work comes among the general framework of Speaker's Automatic Recognition, very fertile field. More particularly the Speaker's Automatic Identification in independent to text mode and in the case where we have a few data of training.

The modelling approach that we use is an interesting approach that consists of representing the speaker, not in an absolute way, but relatively to a well learned set of speakers. Every speaker is represented by his localisation in an optimal space of Eigen voices

In this project, we present a system of Automatic Speaker Identification (ASI) where the representation space is built by the Principal Component Analysis (PCA). The localisation of new speakers in this space is done by orthogonal projection.

Our system which is dedicated to high security applications is based mainly on an original idea that consists of using the acoustical parameters besides anatomical parameters for the speaker parameterisation.

Hence we propose an acoustical anatomical architecture for the identification of the speaker. This architecture is based mainly on the algorithm of fusion of heterogeneous data.

While these data being vitiated by imprecision and uncertainty, we use the principles of the possibility theory to fusion them.

This fusion that leads us to a new Speaker parameterisation, i.e an acoustical anatomical vector, has an aim to reinforce the identification rate and to bring a biometrical dimension to the system

Keywords :

Automatic Speaker Identification, Biometrics, Eigen Voices, Principal Component Analysis , Possibility Theory, Acoustical anatomical Architecture, Fusion Algorithm, Speaker Parametrisation, Acoustical Anatomical Vector , Security.

Dédicace

De par ces quelques lignes écourtées,
Je transmets un message bien édité
A ma mère qui a tant sacrifié
Pour qu'un jour elle me voie émancipée ;
A mon père toujours occupé
Mais de par ses conseils il ne m'a jamais privée ;
Au reste de la famille bien adorée ;
Mon frère, également mes deux sœurs qui ont su m'épauler ;
A la mémoire de mon grand-père duquel j'ai appris que le savoir et la science doivent être vénérés !
Je termine par un être cher, mon bien aimé,
Mon mari à qui je voue respect et fidélité,
Car de bonheur, d'abnégation et de bienveillance il m'a toujours comblée.

Fériel.

REMERCIEMENTS

الحمد لله

Le travail présenté dans ce mémoire a été réalisé au Laboratoire LRI à l'Université Mokhtar Badji sous le précieux suivi de Madame le Docteur ***N. Zine-Ghoualmi.***

Qu'il me soit permis de lui témoigner toute ma gratitude pour m'avoir si bien suivie dans cette étude, pour m'avoir encadrée et pour m'avoir attribué un sujet fort intéressant et d'actualité qui a nourri en moi un esprit scientifique et de recherche.

Aussi, je tiens à lui exprimer ma reconnaissance la plus vive pour ses valeureuses directives et ses critiques très constructives, pour sa rigueur et la main de fer avec laquelle elle a piloté ce travail ne laissant rien au hasard et contrôlant le moindre détail.

Pour m'avoir honorée de sa présence au sein de ce jury, et pour avoir acquiescé sans hésiter d'en assurer la présidence, Monsieur ***M. BenMohammed*** *est prié d'agréer l'expression de mes sincères remerciements.*

Monsieur ***A. Boukeram*** *a d'emblée accepté de siéger à ce jury et d'en faire ainsi partie ; j'en suis honorée et fortement comblée et je le prie de croire en ma reconnaissance infinie.*

Pour l'intérêt qu'elle a bien voulu porter à l'ensemble de ce travail, Madame ***N. Bounour****, est priée de trouver sans faille l'expression de ma reconnaissance pour toujours.*

Une pensée distinctive s'adresse à mon ***mari*** *qui n'a ménagé aucun effort pour m'aider et m'apporter du réconfort.*

Je suis particulièrement reconnaissante envers mon ami de toujours, Monsieur ***Yassine Hammouche****, pour chaque instant passé ensemble en tant que binôme et pour l'aide précieuse et l'appui qu'il m'a toujours fournis.*

Un grand merci pour Monsieur ***Azzedine Boussaada*** *pour m'avoir fait confiance et, de faire partie de son équipe, m'a donné la chance ; aussi pour tous ses conseils et encouragements, sans cesse et en permanence.*

Une large part de ces remerciements est réservée à mon cher père, mon grand frère ***Hichem*** *et* ***Imène*** *mon adorable petite sœur ; à mes tantes, mes oncles et ma belle-famille pour le soutien moral qu'ils n'ont jamais cessé de m'apporter.*

Pour avoir pris la peine de lire et de corriger ce manuscrit, ma très chère sœur ***Hanane*** *est priée de bien vouloir trouver à travers ces modestes lignes mes sentiments sincères de reconnaissance et d'estime.*

Ne sachant les citer un par un et sans exception aucune, à tout un chacun et toute une chacune dont l'aide et l'assistance m'ont été des plus opportunes.

Et à la fin, je ne saurais manquer d'exprimer toute ma gratitude envers la personne la plus chère au monde qui n'a jamais cessé de me pousser à aller de l'avant et qui a tout sacrifié pour moi et pour ses enfants et qui n'est autre que…ma très chère ***Maman …Maman*** *que je ne remercierai jamais assez même infiniment.*

Table des Figures

Liste des Tableaux

Liste des Abréviations

Abréviation	Signification
ACP…….....	Analyse en Composantes Principales.
ALD…….....	Analyse Linéaire Discriminante.
DTW……...	Dynamic Time Warping.
CHM……..	Communication Homme Mchine.
EM……..	Expectation Maximization.
FAR……..	False Acceptance Rate.
FFR……...	False Rejection Rate.
GMM…….....	Gaussian Mixture Model.
HMM……..	Hidden Markov Model.
IAL……..	Identification Automatique du Locuteur.
IC ……	Identification Correcte.
II …….	Identification Incorrecte.
JDL …….....	Joint Directors of Laboratories.
LPCC……..	Linear Predictive Cepstrum Coefficient.
MAP……	Maximum A Posteriori.
MFCC…….	Mel Frequency Cepstrum Coefficient.
MRI…….	Magnetic Resonance Imaging.
MSSO…….	Méthodes Statistiques du Second Ordre.
RAL ……...	Reconnaissance Automatique du Locuteur.
RAP…….	Reconnaissance Automatique de la Parole.
RMP ……	Regression-Based Model Prediction.
RNA……	Réseaux de Neurones Artificiels.
RSW ……..	Reference Speaker Weighting.
UBM……...	Universal Background Model.
VAL……	Vérification Automatique du Locuteur.
VQ……..	Vector Quantization.

Table des Matières

La sécurité est depuis tout temps la préoccupation majeure des individus. Que ce soit pour les biens, les personnes ou les données, les techniques de sécurité ont connu une véritable avancée.

Passant de ce que *l'on possède* (clef, badge, etc.) ou de ce que *l'on sait* (mot de passe, etc.), la tendance actuelle de sécurité se base sur ce qu'*on est.*
Cette approche, nommée *Biométrie*, a apporté simplicité et confort aux utilisateurs.

Dans ce travail, nous nous intéressons à la biométrie vocale et plus exactement à l'Identification Automatique du Locuteur (IAL) en mode indépendant du texte. Il s'agit de reconnaître une personne à partir de sa voix.

Nous visons des applications de très haute sécurité où la population concernée est restreinte.

Cette thèse est donc essentiellement dédiée à la conception d'un système d'Identification Automatique du Locuteur et à la Paramétrisation de ce dernier.
Dans cette optique, nous proposons une paramétrisation basée sur une fusion de données hétérogènes. En effet, nous exploitons à côté du signal acoustique les caractéristiques anatomiques de la sphère ORL du Locuteur, plus exactement la longueur et l'épaisseur de ses cordes vocales [Debbeche et al., 2007b], [Debbeche et al., 2008a], [Debbeche et al., 2008b].

Le domaine de la fusion de données connaît ces dernières années une forte évolution, rapide et foisonnante. De nombreuses méthodes ont été proposées mais le choix de l'une d'entre elles est aussi crucial que difficile.

Etant donné que les paramètres que nous utilisons sont entachés d'imprécision et d'incertitude et que nous travaillons sur le cas où nous disposons de peu de données, nous avons opté pour une fusion via les principes de la théorie des possibilités [Dubois et al., 1994].

Ainsi, nous proposons un algorithme de fusion de données qui nous permet d'obtenir un vecteur représentatif du Locuteur.

Après avoir effectué la paramétrisation des Locuteurs, nous passons à l'étape de modélisation.

Pour cette dernière, de nombreuses approches ont été proposées dans la littérature : approche vectorielle, connexionniste, statistique, etc.

De ce large panel, seule l'approche statistique demeure au premier plan des systèmes de Reconnaissance Automatique du Locuteur des récentes années. Offrant d'excellentes performances, elle est généralement considérée comme l'état de l'art dans le domaine (en particulier la méthode GMM).

Malheureusement, dans ces méthodes-là, les performances se dégradent considérablement si les données d'apprentissage sont insuffisantes.

Pour cette raison et vu que nous travaillons sur peu de données, nous avons choisi d'utiliser l'approche relative ou bien de voix propres qui est apparue pour pallier à ce problème. Elle offre une modélisation non plus absolue mais relative à un ensemble de Locuteurs bien appris, [Merlin et al., 1999], [Kuhn et al., 1998a], [Kuhn et al., 1998b], [Kuhn et al., 1998c],[Kuhn et al., 1999], [Mami et al., 2002], [Nguyen et al., 1999].

Chaque Locuteur est représenté par sa localisation dans un espace optimal de voisins.

Dans ce sens, les systèmes se divisent en trois modules. Dans le premier, un espace de représentation des Locuteurs est construit ; le deuxième étant consacré à la localisation des nouveaux Locuteurs dans cet espace. Le test d'identification est effectué dans le dernier module.

Dans cette thèse, nous avons utilisé l'ACP (Analyse en Composantes Principales) pour construire l'espace représentatif. La deuxième phase consiste à placer chaque Locuteur dans l'espace précédemment construit.

Pour ce faire, nous utilisons la projection orthogonale pour localiser les Locuteurs.

Par conséquent dans le troisième module, les Locuteurs sont représentés par des points dans l'espace ; donc nous évaluons la proximité spatiale entre eux par une distance entre leurs vecteurs de coordonnées.

Ainsi, cet ouvrage s'articule autour de quatre chapitres.

Dans le premier, nous introduisons le domaine de la sécurité en exposant la biométrie avec toutes ses techniques, ses avantages et inconvénients tout en mettant l'accent sur la biométrie vocale.

Le deuxième chapitre présente un rappel sur le principe de la reconnaissance automatique du Locuteur.

Les étapes du système de reconnaissance sont exposées en détail avec un état de l'art sur les méthodes de modélisation utilisées. Une description anatomique détaillée du Locuteur est aussi passée en revue.

Le troisième chapitre est consacré à la fusion de données. Tout d'abord, nous dressons une étude assez exhaustive sur le domaine de la fusion ainsi que sur les différentes approches qui y sont utilisées. Nous effectuons ensuite une discussion sur le choix de l'une d'entre elles.

Le système Acoustico-Anatomique, système proposé, fera l'objet du quatrième chapitre.

Nous présentons en premier lieu, l'architecture générale de notre système suivie de la paramétrisation proposée. Nous détaillerons par la suite les étapes de l'algorithme proposé pour la fusion des données acoustiques et anatomiques.

Enfin, nous décrirons l'approche de modélisation empruntée.

Nous clorons ce présent travail par une conclusion et des perspectives.

CHAPITRE I

LA BIOMETRIE

C'est une chose étrange à quel point la sécurité de la conscience donne la sécurité du reste.
[Victor Hugo]
Extrait de
Les Misérables

1.1 INTRODUCTION

Depuis la nuit des temps, le besoin de sécurité s'est fait ressentir. La sécurité peut se voir comme étant l'état d'esprit d'une personne qui se sent tranquille et confiante. C'est le sentiment, bien ou mal fondé, d'être à l'abri de tout danger et risque ; il associe calme, confiance, quiétude, sérénité, tranquillité, assurance et sûreté [Wikipédia].

Avec la croissance incommensurable des données et le besoin de les échanger, tous domaines confondus, les techniques qui ont été proposées pour les sécuriser sont nombreuses.

L'avènement de l'informatique a beaucoup contribué dans la façon de recueillir, traiter, analyser, stocker et partager ces données. Comme l'informatique est en train de changer de manière assez profonde, elle est devenue ubiquitaire.

Au départ confinée aux ordinateurs, elle est en train d'investir les objets de la vie courante et elle devient ainsi de plus en plus diffuse et distribuée dans de multiples domaines et fonctionnalités qui sont amenés à coopérer.

De tout temps, les technologies nouvelles ont stimulé l'imagination des aigrefins ; de ce fait, les chercheurs s'investissent de plus en plus dans les techniques de sécurité, à savoir sécurité de l'information et sécurité des individus. La détermination de l'identité de ces derniers à travers des moyens simples, fiables, efficaces et peu dispendieux est devenue un problème crucial. Traditionnellement, il existe deux manières pour y procéder.

La première est basée essentiellement sur *ce que l'on sait*, à savoir mot de passe, code, etc. La deuxième méthode s'appuie sur *ce que l'on a*, par exemple un badge, une clef, etc.

Ces deux méthodes présentent des inconvénients majeurs ; dans le premier cas, un mot de passe ou bien un code peuvent être oubliés par leur porteur ou devinés par une autre personne. Dans le second cas, un badge ou une clé peuvent être perdus, volés ou copiés par des personnes mal intentionnées.

L'inconvénient commun aux deux méthodes est que l'on identifie un objet et non la personne elle-même. Pour pallier à ce problème, la nouvelle tendance qui apporte simplicité et confort aux utilisateurs consiste à identifier une personne à partir de *ce qu'on est.*

Cette méthode est connue sous le nom de ***« BIOMETRIE ».***

1.2 BIOMETRIE

1.2.1 Définition

La biométrie comme son étymologie l'indique est composée des deux mots : vie et mesure.
De par cette étymologie, un système de contrôle biométrique peut être défini comme étant un système automatique de mesure basé sur la reconnaissance de caractéristiques propres à l'individu [Clusif, 2003].
Dans la langue française, plusieurs acceptions du mot biométrie sont données :

« Étude mathématique, surtout statistique, des phénomènes biologiques » Dictionnaire Hachette.
« Science qui étudie à l'aide des mathématiques (statistiques et probabilités) les variations biologiques à l'intérieur d'un groupe déterminé » Dictionnaire Robert.

De ces définitions, la biométrie peut être vue comme étant une technique globale qui repose sur l'analyse mathématique des caractéristiques biologiques d'une personne, destinée à déterminer son identité de manière irréfutable. Il peut y avoir plusieurs types de caractéristiques physiques, les unes plus fiables que d'autres, mais toutes doivent être infalsifiables et uniques pour pouvoir être représentatives d'un et un seul individu.
La biométrie n'est pas une discipline récente, son histoire remonte à plusieurs siècles. [Wikipedia]. Elle émane de l'anthropométrie, méthode scientifique développée par Alphonse Bertillon (dans le cadre de ses fonctions de chef du service de l'identité judiciaire à la préfecture de police de Paris, 1882) permettant l'identification de criminels d'après leurs mesures physiologiques.
L'intérêt à la biométrie s'est accentué ; c'est un champ de recherche qui se voit de plus en plus attrayant et qui croît de façon exponentielle. Cela est dû aux évènements du 11 septembre 2001 aux Etats-Unis. Elle répond à une exigence très actuelle de sécurité partagée aussi bien par les individus et les entreprises que par les Etats.

1.2.2 Techniques biométriques

On recense plus d'une dizaine de technologies biométriques classées en trois catégories : les premières reposent sur l'analyse morphologique ; les secondes sur l'analyse comportementale

et les troisièmes enfin sur l'analyse de traces biologiques [Biometrie], [Mahmoudi, 2000], [Sécurite], [Savoirs].

1.2.2.1 Morphologie

Il existe plusieurs caractéristiques physiques qui se révèlent être uniques pour un individu ; celles exploitées jusqu'à nos jours par les systèmes biométriques sont :

- Empreintes digitales
- Géométrie de la main
- Iris
- Rétine
- Visage
- Configuration des veines

1.2.2.2 Comportement

Outre les caractéristiques physiques, un individu possède également plusieurs éléments liés à son comportement qui lui sont propres :

- Dynamique des frappes au clavier
- Voix
- Dynamique des signatures
- Démarche

1.2.2.3 Biologie

Comme caractéristiques biologiques, on peut citer :

- Odeur
- ADN
- Salive
- Etc.

1.2.2.4 Comparatif

Une comparaison entre les différentes techniques biométriques s'avère intéressante. Elle permet de choisir une technologie en fonction des contraintes liées à l'application.

Une compagnie américaine (*a New York based integration and consulting firm*), *l'International Biometric Group* a procédé à une comparaison sur la base de 4 critères (Figure 1.1) :

- L'intrusivité : Ce critère décrit dans quelle mesure l'utilisateur perçoit la technique comme intrusive.
- La précision : C'est l'efficacité de la méthode (capacité à identifier un individu).
- Le coût : Coût de la technologie utilisée (lecteurs, capteurs, etc.).
- L'effort : Effort requis pour l'utilisateur lors de la mesure.

Une autre comparaison (avantages / inconvénients) a été faite, cette fois-ci, par la France par le biais du Club de la Sécurité des Systèmes d'Information Français (le CLUSIF) [Clusif, 2003]. Voir Tableau 1.1.

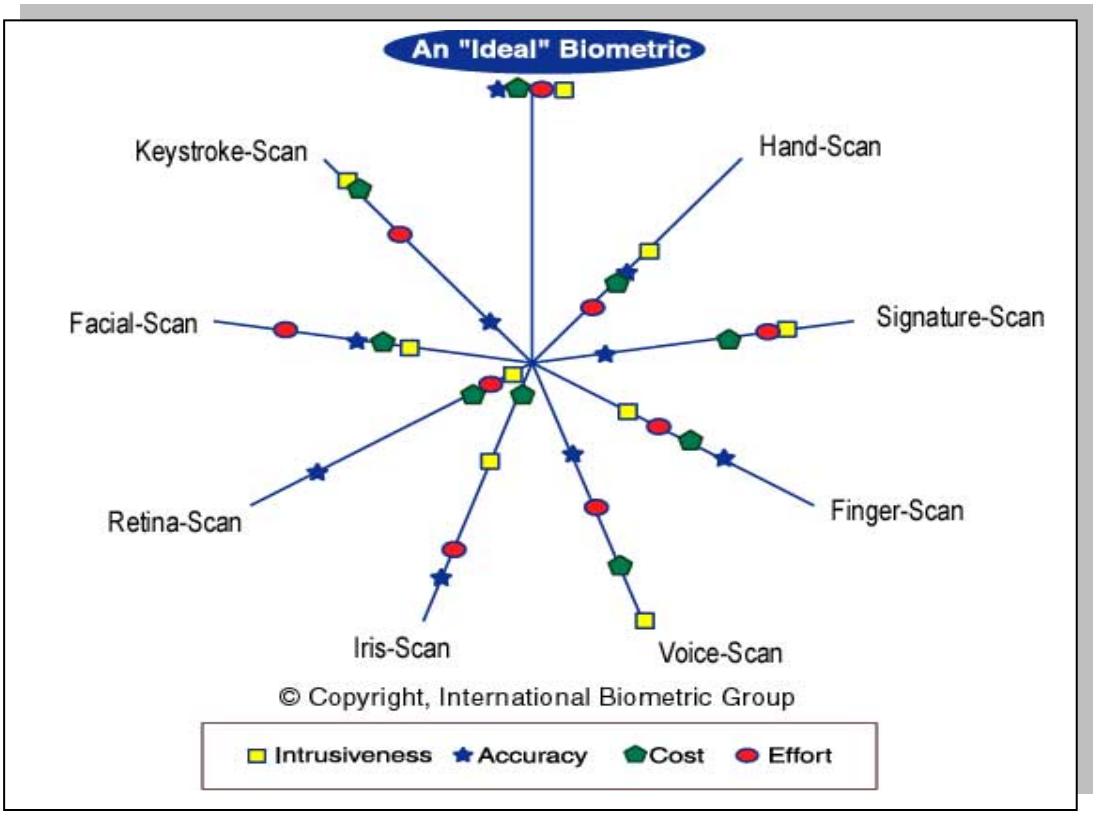

Figure 1.1 Comparaison entre les différentes techniques biométriques.

Tableau 1.1 Avantages et inconvénients des technologies biométriques.

Techniques	Avantages	Inconvénients
Empreintes digitales	Coût, ergonomie moyenne, facilité de mise en place, taille du capteur.	Qualité optimale des appareils de mesure (fiabilité), acceptabilité moyenne.
Forme de la main	Très ergonomique, bonne acceptabilité.	Système encombrant, coût, perturbation possible par des blessures et l'authentification des membres d'une même famille.
Visage	Coût, peu encombrant, bonne acceptabilité.	Jumeaux, psychologie, religion, déguisement, vulnérabilité aux attaques.
Rétine	Fiabilité, pérennité.	Coût, acceptabilité faible, installation difficile.
Iris	Fiabilité.	Acceptabilité très faible, contrainte d'éclairage.
Voix	Facilité.	Vulnérable aux attaques.
Signature	Ergonomie.	Dépendant de l'état émotionnel de la personne, fiabilité.
Frappe au clavier	Ergonomie.	Dépendant de l'état physique de la personne.

1.2.3 Panorama d'application

Les usages de la biométrie sont de deux ordres ; le premier est lié aux enjeux sécuritaires tandis que le second s'apparente à une biométrie de confort.

D'une façon générale, Trois secteurs peuvent être délimités :

- L'identification judicaire.
- La gestion des titres délivrés par la puissance publique (cartes d'identité, passeports informatisés).
- La gestion des accès physiques (locaux, service de recherche, site nucléaire, etc.) et logiques (contrôle d'accès à un ordinateur, login d'ouverture de sessions réseaux, accès distants, etc.)

1.2.4 Processus d'identification biométrique

Tout système biométrique est composé de deux grandes étapes ; l'enrôlement et le contrôle (Figure 1.2) [Biométrie].

L'enrôlement des personnes est la phase initiale de création du gabarit biométrique et de son stockage en liaison avec une identité déclarée. Les caractéristiques physiques sont transformées en un modèle représentatif de la personne et propre au système de reconnaissance. Cette étape n'est effectuée qu'une seule fois.

Le contrôle d'un autre côté, représente l'action de contrôler les données d'une personne afin de procéder à la vérification de son identité proclamée ou à son identification. Cette étape se déroule à chaque fois qu'une personne se présente devant le système.

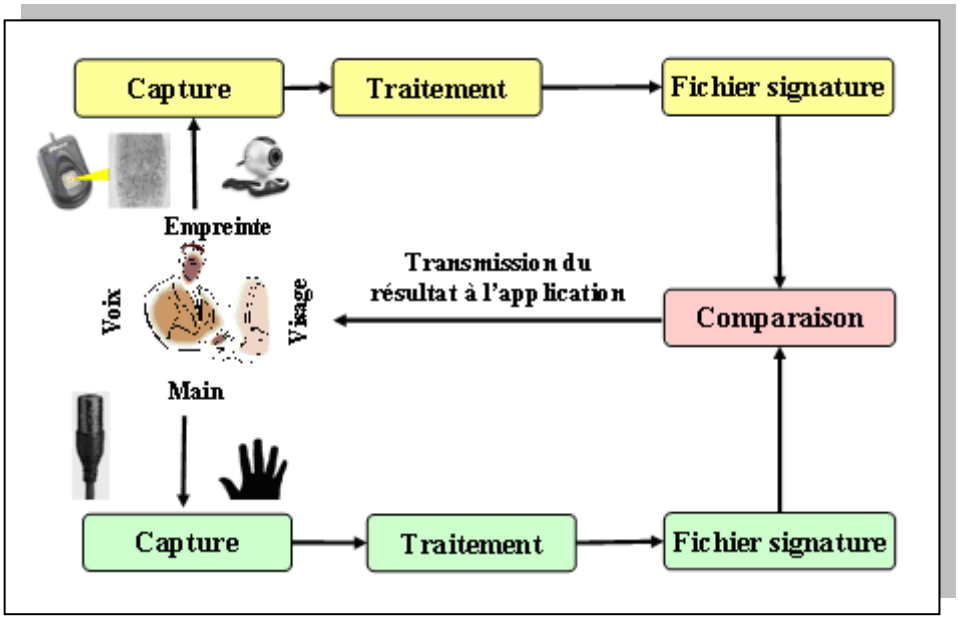

Figure 1.2 Processus d'un système d'identification biométrique.

1.2.5 Identification vs Vérification

Il convient de distinguer deux modes de fonctionnement des systèmes biométriques ; systèmes d'identification et système de vérification [Clusif, 2003].

La vérification (authentification) consiste à confirmer l'identité revendiquée par un utilisateur. C'est une comparaison « *un pour un* » dans laquelle le modèle biométrique saisi est comparé au modèle de référence.

L'identification permet de vérifier que l'identité d'un individu qui se présente existe bien dans la base de référence. C'est une comparaison « *un pour plusieurs* » où le modèle saisi est comparé à tous les modèles stockés dans la base.

1.2.6 Fiabilité des systèmes biométriques

En réalité, il est presque impossible d'obtenir dans un système biométrique une coïncidence absolue entre le fichier "signature" créé lors de l'enrôlement et le fichier "signature" créé lors de l'identification.

Les éléments d'origine (image, son...) utilisés ne peuvent jamais être reproduits à l'identique. C'est là l'inconvénient majeur de la biométrie ; en effet, aucune des mesures utilisées ne se révèle être totalement exacte car il s'agit de caractéristiques concernant un organisme vivant : on s'adapte à l'environnement, on vieillit, on subit des traumatismes plus ou moins importants, on évolue, etc., par conséquent, les mesures changent.

Partant de ce fait, les performances des systèmes biométriques se mesurent via deux taux d'erreur [Clusif, 2003] :

- FRR (*False Rejection Rate*) : se rapporte à la probabilité qu'un système biométrique échoue dans l'authentification ou l'identification d'une personne enregistrée.
- FAR (*False Acceptance Rate*): représente le pourcentage d'acceptations par erreur.

Trouver un compromis entre ces deux taux d'erreur est le souci de tout industriel et chercheur dans le domaine biométrique.

1.2.7 Biométrie Vocale

La biométrie vocale est la technique de reconnaissance la moins intrusive ; elle n'exige généralement aucun contact physique avec le récepteur du système. C'est la seule biométrie qui permet une reconnaissance à distance.

Elle se concentre sur les seules caractéristiques de la voix qui sont uniques à la configuration de la parole d'un individu.

Ces caractéristiques émanent à la fois de facteurs physiologiques (forme de l'appareil vocal, sexe, âge…) et comportementaux (vitesse, rythme, accent…). Ainsi, la voix peut être classée conjointement dans la biométrie physique (morphologique), car la forme de l'appareil vocal influe directement sur la production de la parole, ainsi que dans la biométrie comportementale, car la production de cette dernière est en grande partie apprise et non innée.

Comme toute technologie biométrique, la voix présente des inconvénients qui peuvent influer sur la fiabilité des systèmes.

La voix peut être altérée soit par des facteurs intrinsèques au locuteur (son état physique et émotionnel), soit par des facteurs extrinsèques tels que les conditions d'enregistrement du signal de parole (bruit ambiant, qualité du microphone, etc.).

1.3 CONCLUSION

La biométrie, comme nous avons pu le constater tout au long de ce chapitre, offre de nos jours la solution tant attendue dans le domaine privé ou professionnel de pouvoir sécuriser les systèmes ou bien les individus eux-mêmes.

Elle représente une véritable alternative aux méthodes traditionnelles de sécurité en vérifiant réellement l'identité de la personne et non plus ce que cette dernière possède ou connaît.

Cette technologie est en plein essor et les systèmes développés tant en biométrie morphologique, comportementale ou biologique sont encourageants. Certes, les performances n'ont pas encore atteint les 100 % de fiabilité mais l'avancée que connaît ce domaine est vraiment prometteuse.

La voix, sujet de ce présent travail, présente comme toute autre technique des inconvénients inhérents à l'individu ou à l'environnement. C'est un moyen de reconnaissance très intéressant permettant une authentification à distance.

La voix présente aussi l'avantage majeur d'être ergonomique et non intrusive. Elle est très acceptée par les utilisateurs par rapport aux autres biométries.

Une pléthore de systèmes de reconnaissance vocale a été proposée ces dernières années ; une étude plus approfondie et un état de l'art seront présentés dans le prochain chapitre.

CHAPITRE II

LA RECONNAISSANCE AUTOMATIQUE DU LOCUTEUR

La conscience est la voix de l'âme, les passions sont la voix du corps.
[Jean-Jacques Rousseau]

2.1 INTRODUCTION

La parole est depuis tout temps le moyen de communication privilégié de l'Homme. Elle véhicule, en plus du message linguistique prononcé, plusieurs types d'informations.
Ces informations servent en particulier à déterminer l'*identité* du Locuteur ; elles sont exploitées par les humains pour l'identification des personnes qu'ils connaissent en particulier à distance (au téléphone par exemple).
Les systèmes de Reconnaissance Automatique du Locuteur (RAL) s'intéressent précisément à ces caractéristiques particulières du signal de parole. Cette discipline s'inscrit dans le cadre général de la reconnaissance des formes ; c'est un terme générique qui regroupe les problèmes relatifs à l'identification ou à la vérification du Locuteur sur la base de l'information contenue dans le signal acoustique : il est question de reconnaître une personne à partir de sa voix.
Le champ d'application est très vaste, il va du simple contrôle d'accès, aux applications militaires passant par des applications judiciaires.
Un système de RAL opère en trois étapes : l'analyse acoustique du signal de parole, la modélisation du Locuteur et une dernière étape de décision.

Ce chapitre est une introduction au domaine de RAL ; il présente tout d'abord les différentes tâches liées à ce domaine.
Les principes ainsi qu'un état de l'art sur les méthodes et les techniques afférentes à ces tâches y sont décrits.
Une description anatomique détaillée du Locuteur ainsi que les limites des systèmes d'identification vocale dues à la variabilité intrinsèque et extrinsèque au Locuteur y sont aussi exposées.

2.2 LA VOIX

La voix est un instrument paradoxal. Il est à la fois banal et précieux, fragile et puissant. [Musimem]
La voix de chaque personne dépend des caractéristiques, à la fois anatomiques et comportementales. Avant de parler de la reconnaissance automatique du Locuteur, il est important de le décrire anatomiquement pour comprendre le processus d'émission de la voix et connaître les paramètres qui différencient un Locuteur d'un autre.

A côté de l'aspect anatomique, on présentera aussi dans cette section une description du signal vocal.

2.2.1 Description Anatomique du Locuteur

L'appareil vocal (Figure 2.1) est constitué de structures appartenant à l'appareil respiratoire et à l'appareil digestif. On le décompose classiquement en trois étages [Kob], [Roublot, 2003], [Flanagan, 1972], [Bartkova, 2002] :

1. **La soufflerie** : Elle comprend la musculature respiratoire, les poumons, et les conduits sus-jacents.
   La soufflerie produit le flux d'air qui sera la matière première de la production vocale, expiré par les poumons et acheminé par la trachée vers le larynx.

2. **Le vibrateur** : Il s'agit du larynx qui est un tube situé à l'extrémité supérieure de la trachée, au niveau de la pomme d'Adam. La colonne d'air produite par la soufflerie est mise en vibration sous l'action des cordes vocales.

3. **Les résonateurs :** Ce sont principalement les cavités supra laryngées, à savoir le pharynx, la cavité buccale et les fosses nasales.
   La forme et le volume de ces cavités sont très variables selon les individus ; c'est ce qui explique que chaque personne ait un timbre de voix personnel et identifiable. Par ailleurs, les mouvements des muscles du pharynx et de la bouche (notamment : de la langue) permettent des modifications rapides du volume et de la forme de ces résonateurs qui transforment la voix produite par la vibration laryngée en phonèmes constitutifs de la parole articulée et ce, par l'amplification sélective de certaines fréquences laryngées.

Les cordes vocales sont attachées horizontalement entre le cartilage thyroïde (la "pomme d'Adam" chez l'homme) situé à l'avant et les cartilages aryténoïdes situés à l'arrière. En faisant bouger ces cartilages en parlant, on modifie la longueur et la position des cordes vocales.
Lorsque la personne commence à dire quelques mots, les cartilages aryténoïdes accolent les cordes vocales l'une contre l'autre, fermant ainsi la glotte.
Sous la pression de l'air expiré, les cordes vocales s'écartent, puis se referment aussitôt, entraînant à nouveau une hausse de la pression sous la glotte.

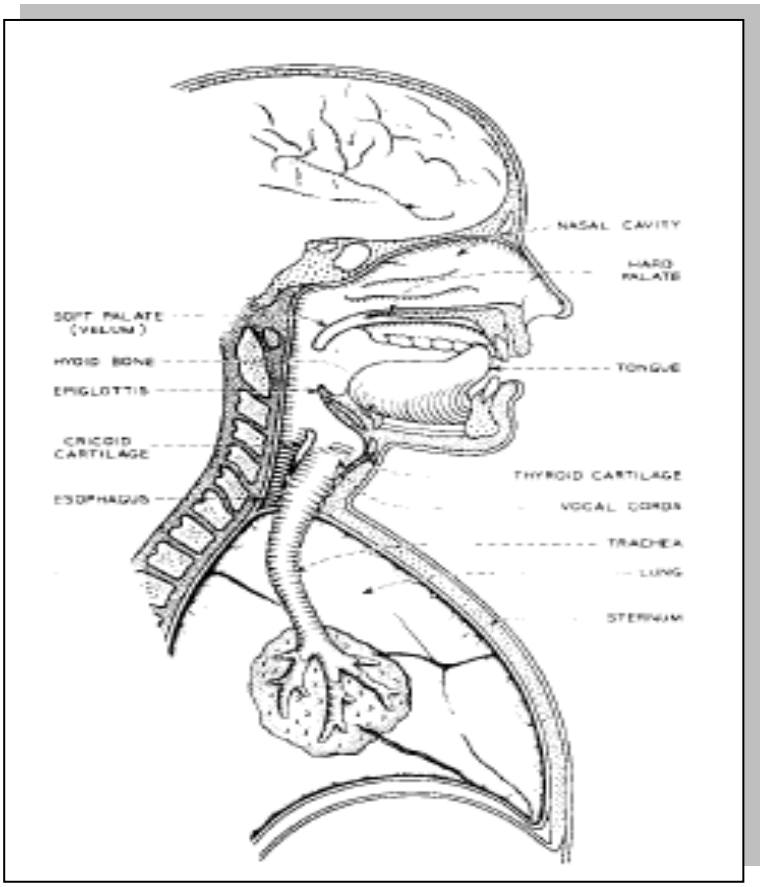

Figure 2.1 Système vocal [Flanagan, 1972].

En ouvrant et fermant la glotte lors de la phonation, les cordes vocales libèrent de façon saccadée l'air emmagasiné dans les poumons. Au cours d'une phrase, le Locuteur modifie ainsi plusieurs fois la fréquence de vibration des cordes vocales pour produire les vibrations acoustiques correspondant à différents sons [Kob], [Roublot, 2003], [Flanagan, 1972].

2.2.2 Description Physique du Signal Vocal

En plus du message linguistique servant à la communication entre individus, le signal de parole véhicule des informations caractéristiques de la personne qui l'a émis comme le timbre de sa voix, sa façon de parler, son état émotionnel ou pathologique, etc.
Ces informations caractéristiques du Locuteur peuvent être classées en deux catégories distinctes :

- Les informations de nature statique telles que les paramètres spectraux caractérisant les conduits vocal et nasal, la moyenne et les variations de la fréquence fondamentale.

- Les informations de nature dynamique reflétant les phénomènes de co-articulation, les trajectoires formantiques ainsi que les informations temporelles (vitesse d'élocution, distribution des pauses).

Nous parlerons ici des caractéristiques statiques du signal vocal.
Ce dernier peut être défini par 4 paramètres principaux [Zwicker et al., 1981], [Reynolds, 1994], [Homayounpour et al., 1994] :

1. **Intensité:** L'intensité d'un son correspond à l'amplitude de la vibration acoustique ; elle caractérise le volume sonore qui nous permet de distinguer un son fort d'un son faible.
   L'intensité vocale varie surtout en fonction de la pression sous glottique.

2. **Timbre:** Le timbre permet de différencier deux sons de même hauteur et de même amplitude. Il est constitué d'un ensemble de fréquences appelé spectre. La richesse du spectre permettra de dire qu'un son est riche, brillant, profond, etc.
   Le timbre est fonction des trois critères suivants : des conditions d'accolement des cordes vocales, de leur épaisseur et enfin des caractéristiques anatomiques des cavités de résonance (pharynx, bouche et cavités nasales).

3. **Hauteur:** La hauteur dépend de la fréquence de la variation de pression acoustique correspondant au son.
   Elle est fonction de la périodicité du mouvement des lèvres glottiques, c'est-à-dire en pratique, du nombre d'ouvertures glottiques par seconde.
   La hauteur dépend aussi de la taille du larynx : plus les cordes vocales sont longues, plus la voix est grave.

4. **Fréquence :** Elle représente le nombre de vibrations de l'air en une seconde.

2.3 DE LA RECONNAISSANCE HUMAINE A LA RECONNAISSANCE AUTOMATIQUE

2.3.1 Reconnaissance Auditive

Utilisée jusqu'à nos jours dans le domaine juridique, l'identification auditive se base essentiellement sur la capacité naturelle de l'être humain à reconnaître une personne en utilisant seulement l'écoute de sa voix.

Cette capacité est cependant variable selon les individus [Ladefoged et al., 1980], [Schmidt et al., 2000] et reste influencée par différents facteurs : familiarité entre l'auditeur et le Locuteur [Van Lancker et al., 1985], [Papcun et al., 1989], [Yarmey et al., 2001], durée des enregistrements, conditions de stress ou de modifications volontaires de la voix, etc.

2.3.2 Reconnaissance par spectrogramme

Une '*empreinte vocale*' est en fait un terme qui fait référence à un spectrogramme du signal vocal [Bolt et al., 1970], [Stevens et al., 1968]. Il s'agit d'un graphique qui représente le signal en trois dimensions : temps, fréquence et intensité.

Le spectrogramme est un outil utile pour le traitement et l'analyse de la voix mais n'a cependant aucun lien avec les empreintes digitales ou génétiques.

L'analyse des empreintes digitales par exemple, bénéficie d'une longue histoire et de bases de données expérimentales de dimension très importante. Dans le domaine vocal, les bases de données disponibles ne comportent pas un nombre suffisant de Locuteurs, de langues et de conditions d'enregistrement pour l'évaluation des méthodes d'authentification criminalistique, à haut niveau de fiabilité.

De plus, la voix présente des différences majeures avec les empreintes digitales et génétiques. Elle évolue dans le temps, elle peut être modifiée volontairement par son porteur, elle est facilement falsifiable, etc. Par conséquent, on ne parle pas d'*empreinte* vocale mais plutôt de *signature* vocale.

La reconnaissance vocale par spectrogramme se fait par comparaison spectrale (spectrographiques) de mots.

2.3.3 Reconnaissance phonétique

Cette méthode utilise une approche linguistique [Nolan, 1983].

L'information recueillie à travers l'étude systématique des sons d'une langue est utilisée par un expert phonéticien pour produire une preuve correspondant à la vraisemblance pour qu'un enregistrement vocal ait été produit par une personne donnée. Plus exactement, il s'agit d'estimer combien de fois il est plus probable d'observer une différence entre les exemples vocaux si ceux-ci proviennent du même Locuteur ou, au contraire, de deux Locuteurs différents (le rapport de vraisemblance Bayésien est utilisé).

2.3.4 Reconnaissance Automatique

Cette approche s'appuie sur un processus automatique consistant à déterminer, parmi une population de Locuteurs connus, la personne ayant prononcé un message donné.

L'approche automatique sous tous ses aspects sera détaillée dans la suite de ce chapitre.

2.4 RECONNAISSANCE AUTOMATIQUE DU LOCUTEUR

2.4.1 Généralité

La Reconnaissance Automatique du Locuteur (RAL) s'inscrit dans le cadre général du traitement automatique de la parole ; son objectif principal est de déterminer l'identité d'une personne à l'aide de sa voix [Campbell, 1997].

Elle tire son essence de la variabilité interlocuteur. Cette variabilité est due principalement aux différences morphologiques de l'appareil vocal d'un individu à l'autre.

On peut classifier les systèmes de RAL suivant leur dépendance au texte. On distingue les systèmes dépendants du texte des systèmes indépendants du texte.

Dans le mode dépendant du texte [Charlet, 1997], la reconnaissance est réalisée à l'aide d'un message connu *a priori* par le système que le Locuteur doit prononcer (mot de passe, code PIN, phrase, ...). Ce message peut être choisi par le Locuteur ou imposé par le système.

Le mode indépendant du texte contrairement au premier n'impose aucune contrainte, sur le message à prononcer, au Locuteur.

Les systèmes de RAL sont sensibles à certains facteurs qui peuvent altérer leur performance ; ces facteurs peuvent être intrinsèques ou extrinsèques au Locuteur. On peut citer :

- L'état pathologique du Locuteur (maladie, émotion, …)
- Vieillissement.
- Facteurs socioculturels.

- Locuteurs non coopératifs.
- Conditions de prise de son.
- Bruit ambiant,…
- Etc.

2.4.2 Différentes tâches en RAL

Les deux tâches pionnières des systèmes de Reconnaissance Automatiques du Locuteur (RAL) sont l'Identification Automatique du Locuteur (IAL) et la Vérification Automatique du Locuteur (VAL) [Atal, 1976], [Doddington, 1985], [Rosenberg et al., 1991].

Récemment, des besoins spécifiques ont stimulé l'apparition de nouvelles tâches comme l'indexation du Locuteur qui consiste à indiquer à quel moment chaque Locuteur intervenant dans une conversation a pris la parole, le suivi de Locuteurs (speaker tracking) ou bien une application connexe à l'indexation qui est la détection d'un Locuteur lors d'une conversation. Dans cette section, nous allons décrire les principales tâches de la RAL qui sont l'IAL et la VAL.

2.4.2.1 Identification Automatique du Locuteur (IAL)

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. Cela est fait en calculant des mesures de similarité entre le signal en entrée et tous les modèles des Locuteurs de la base.

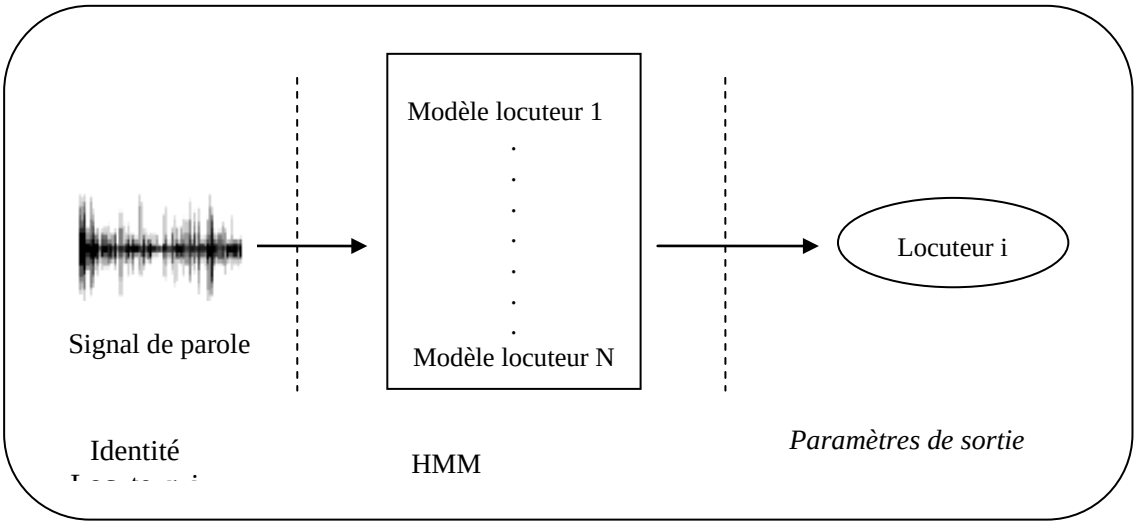

Figure2.2 Schéma typique d'un système d'IAL

L'identité du Locuteur, dont le modèle est le plus proche du signal en entrée, est donnée en sortie du système d'IAL (voir Figure 2.2).
Deux modes d'identification sont possibles :

- Identification en ensemble fermé : c'est le cas où le système doit fournir comme sortie un ensemble d'au moins un Locuteur. En d'autres termes, la séquence fournie en entrée doit être en fait prononcée par un Locuteur connu du système.

- Identification en ensemble ouvert : le système dans ce cas peut être amené à fournir un ensemble vide, car le Locuteur peut ne pas être connu.
  Dans ce mode, le système d'IAL doit décider de la fiabilité de son jugement en acceptant ou rejetant l'identité qu'il a trouvée.

En pratique, la plupart des systèmes d'IAL fournissent un ensemble d'un seul Locuteur qui représente le Locuteur le plus proche.

2.4.2.2 Vérification Automatique du Locuteur (VAL)

La Vérification Automatique du Locuteur est une décision en tout ou en rien. Elle consiste à déterminer à partir d'un message vocal, la véracité de l'identité proclamée par un individu.
Les entrées du système sont donc le signal de parole et l'identité proclamée et la sortie une acceptation ou bien un rejet.

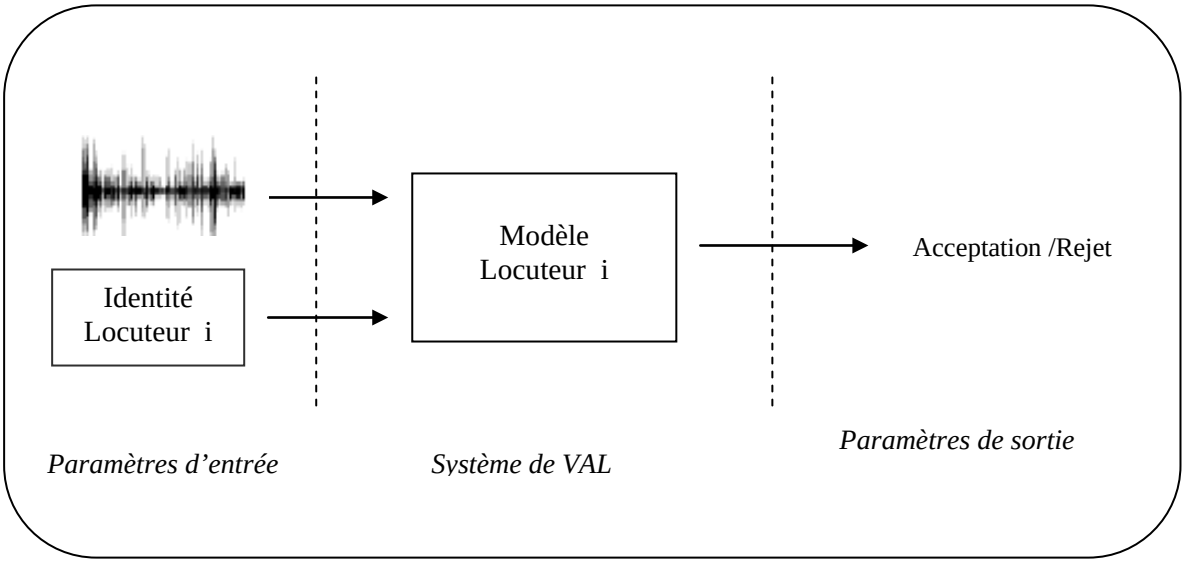

Figure 2.3 Schéma typique d'un système de VAL

Dans la suite de ce chapitre, nous nous intéresserons à l'Identification Automatique du Locuteur (IAL) sujet de ce présent travail.

2.5 STRUCTURE DES SYSTEMES D'IAL

Un système d'IAL se résume à l'enchaînement de trois processus principaux qui sont : la paramétrisation, la reconnaissance et la décision

En premier lieu, le message vocal est analysé acoustiquement. A l'issue de cette analyse, on obtient un ensemble de vecteurs de coefficients pertinents qui vont être utilisés pour la modélisation des Locuteurs.

A la reconnaissance, une mesure de similarité va être calculée entre les paramètres acoustiques du signal prononcé et les modèles contenus dans la base.

La dernière étape du système est un module de décision qui, basé sur une stratégie de décision donnée, fournit la réponse du système.

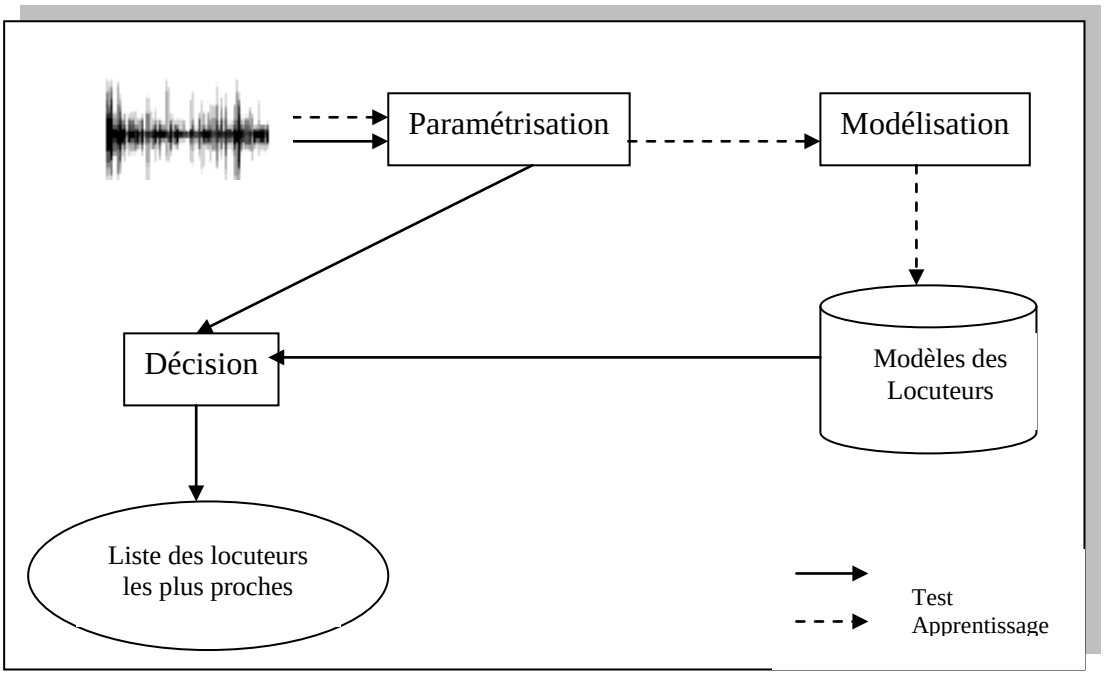

Figure 2.4 Schéma modulaire d'un système d'IAL.

2.5.1 Paramétrisation Acoustique

Le processus de paramétrisation consiste à extraire du signal de parole l'information pertinente et à réduire au maximum la redondance en vue de la reconnaissance.

C'est une représentation plus simple du signal de parole sous forme de vecteurs de paramètres acoustiques.
Le calcul de ces derniers est réalisé en glissant avec une cadence régulière (ex : 10 ms) une fenêtre de pondération d'une longueur variant généralement de 20 à 32 ms.
Le fenêtrage le plus utilisé en traitement du signal de parole est en général le fenêtrage de Hamming. Chaque fenêtre nous permet d'avoir une trame.
Les trames ainsi obtenues sur tout le signal de parole sont traitées par la suite afin de produire les vecteurs de paramètres acoustiques.

Nous retrouvons dans la littérature trois grandes familles de paramètres :

- 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 (ex : Linear Predictive Cepstrum Coefficient « LPCC », Mel Frequency Cepstrum Coefficient « MFCC »). Pour plus d'informations sur le sujet, le lecteur pourra se référer aux travaux suivants [Homayounpour et al., 1994], [Reynolds, 1994], [Charlet, 1997].

- Paramètres prosodiques : Ces paramètres illustrent en général le style d'élocution d'un Locuteur : vitesse d'élocution (débit), durée et fréquence des pauses, pitch ainsi que les caractéristiques de la source glottale (fréquence fondamentale, énergie, ...).

- Paramètres dynamiques : Le vecteur de paramètres issus des paramétrisations précédentes peut être complété par le vecteur correspondant aux dérivées du premier et second ordres de ces paramètres.
  Ces dérivées sont les paramètres dynamiques les plus répandus ; on les appelle aussi coefficients Delta (dérivée première) et Delta-Delta (dérivée seconde) [Furui, 1981], [Soong et al., 1988], [Bernasconi, 1990].
  D'autres paramétrisations pour exploiter les informations dynamiques sont proposées dans la littérature comme l'utilisation des Composantes Principales Temps – Fréquence (TFPC : *Time Frequency Principal Components*) [Magrin Chagnolleau et al., 1999], ou encore la concaténation de trames successives de signal [Hattori, 1992], [Konig et al., 1998],[Fredouille et al., 1998], [Fredouille et al., 2000], etc.

2.5.2 Modélisation des Locuteurs

Le processus d'IAL se base essentiellement sur la phase de modélisation des caractéristiques des Locuteurs. Cette modélisation est réalisée à partir de données d'apprentissage collectées au cours des sessions d'enrôlement.

Les méthodes existantes de modélisation des Locuteurs peuvent être répertoriées en cinq grandes approches :

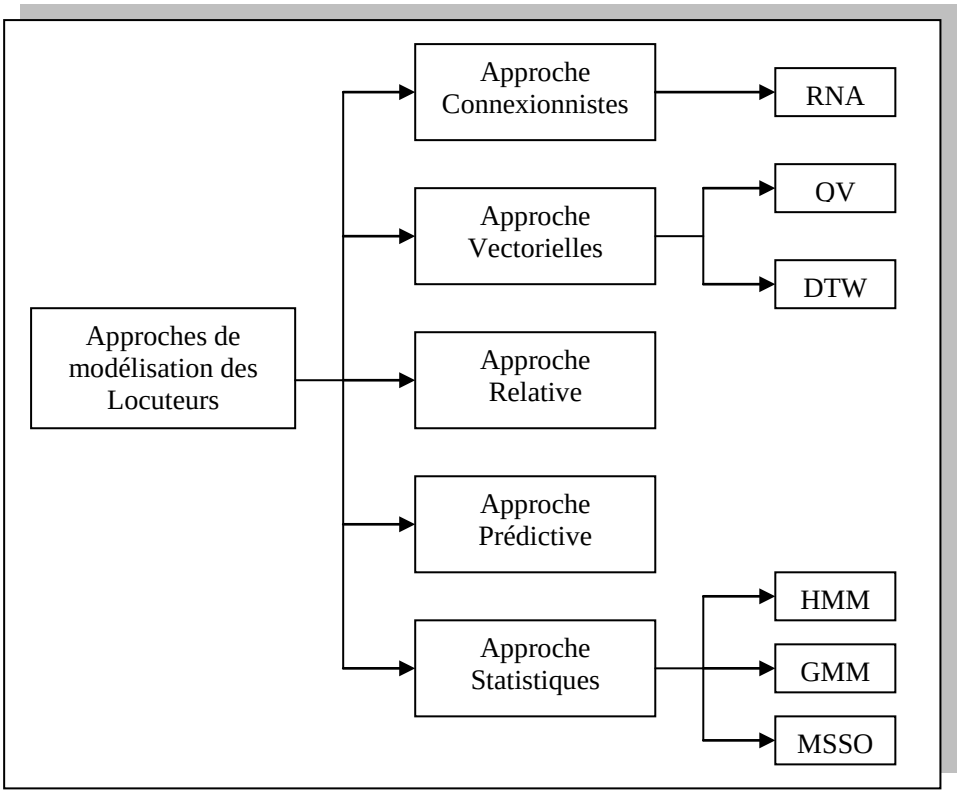

Figure 2.5 Approches de modélisation des Locuteurs

2.5.2.1 Approches Vectorielles

Dans l'approche vectorielle, un modèle de Locuteur est un ensemble de vecteurs de paramètres représentatifs de l'espace acoustique issus de la phase de paramétrisation des signaux d'apprentissage.

Dans cette approche, on retrouve 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.
De par son principe, la programmation dynamique est utilisée exclusivement en mode dépendant du texte [Furui, 1981], [Booth et al., 1993], [Yu et al., 1995].
Bien qu'elle soit facile à mettre en œuvre, 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 des deux formes à comparer.

- ***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, 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 retenue.
La quantification vectorielle s'applique en mode dépendant ou indépendant du texte [Soong et al., 1992], [Mason et al., 1989], [Matsui et al., 1992].
La rapidité et les performances de cette technique dépendent fortement de la taille du dictionnaire : plus la taille du dictionnaire augmente, meilleures sont les performances. Néanmoins, le processus devient d'autant plus lent.

2.5.2.2 Approches Connexionnistes

L'approche connexionniste repose sur la discrimination entre Locuteurs.
Un ensemble de signaux de parole issus d'une population de Locuteurs clients est fourni en entrée à un réseau de neurones pour une étape d'apprentissage.
A l'issue de cette étape, le réseau apprend à discriminer un Locuteur des autres. L'approche connexionniste se résume, par conséquent, à une tâche de classification.

Le Locuteur dans cette approche est représenté par un ou plusieurs réseaux de neurones appris directement des trames obtenues de la phase de paramétrisation [Bennani, 1992] [Bennani et al., 1994].

L'inconvénient majeur de cette approche est la complexité d'apprentissage. De plus, 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.

D'un autre côté, son principal avantage se résume à sa capacité discriminante qui n'exige pas beaucoup d'hypothèses ni beaucoup de connaissances sur l'application.

2.5.2.3 Approches Statistiques

L'inconvénient commun aux méthodes présentées précédemment est qu'elles ne tiennent pas compte de l'ordre dans lequel les vecteurs de paramètres sont présentés.

L'approche statistique résout ce problème en utilisant des techniques qui permettent de construire des modèles qui prennent en considération l'aspect temporel du signal de parole.

Les vecteurs acoustiques issus de la paramétrisation sont donc représentés par des statistiques à long terme.

- ***Modèles de Markov cachés (HMM)***

Les modèles de Markov (HMM : *Hidden Markov Models*) ont été largement utilisés en Reconnaissance Automatique de la Parole. Plus récemment, leur utilisation s'est étendue à la Reconnaissance Automatique du Locuteur.

La modélisation dans ce cas de figure se fait par une succession d'états avec des probabilités de transition d'un état à l'autre.

La reconnaissance se fait par calcul de la vraisemblance qu'une séquence de vecteurs de test soit issue de la chaîne de Markov.

L'utilisation des HMMs en mode dépendant du texte fournit d'excellents résultats [Rosenberg et al., 1991]

- ***Les modèles de mélange de gaussiennes (GMM)***

Les modèles GMM ***(Gaussian Mixture Models)*** sont considérés comme étant la modélisation «état de l'art» des systèmes de Reconnaissance Automatique du Locuteur en mode indépendant du texte.

Cette approche consiste à modéliser un Locuteur par un mélange de gaussiennes qui représente une somme pondérée de M gaussiennes multidimensionnelles [Reynolds, 1992], [Reynolds et al., 2000], [Reynolds, 1995].

Chaque gaussienne *gi* est supposée modéliser un ensemble de classes acoustiques. Elle est caractérisée par son poids *pi*, un vecteur moyen μi de dimension *d* et une matrice de covariance $\sum_i$ de dimension $d \times d$.

L'inconvénient majeur de cette technique est la quantité de signaux d'apprentissage requise pour une bonne estimation des paramètres des modèles (voir Annexe).

- ***Méthodes Statistiques du Second Ordre (MSSO)***

Cette approche est généralement associée à une famille de mesures de similarité entre Locuteurs en vue de la reconnaissance. On peut citer : rapport de vraisemblance, distance de *Kullback-Leiber*, maximum de vraisemblance, etc.

Le modèle d'un Locuteur se résume au triplet $\{x; X_0; M\}$ où x est un vecteur moyen, X_0 est une matrice de covariance ; tous deux estimés à partir de la séquence de M vecteurs acoustiques.

Les mesures de similarité reposent ainsi essentiellement sur une ressemblance entre les matrices de covariance de test et d'apprentissage [Bimbot et al., 1995].

L'avantage majeur des MSSO est leur simplicité de mise en oeuvre. Performantes sur de courtes durées (3 secondes) [Magrin Chagnolleau et al., 1995], elles ne capturent que les caractéristiques stables le long du signal de parole. Les variations locales sont, quant à elles, moyennées et ne sont pas prises en compte par les modèles.

Ces spécificités des MSSO 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.

2.5.2.4 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.

2.5.2.5 Approche Relative

Le principe de la reconnaissance relative des Locuteurs a été initialement appliqué en reconnaissance de la parole dans des techniques d'adaptation rapide [Nguyen et al., 1999], [Kuhn et al., 2000], [Kuhn et al., 1999], [Kuhn et al., 1998a], [Kuhn et al., 1998b], [Kuhn et al., 1998c].

Ces approches ont donné naissance à la notion «d'espace de Locuteurs» où un modèle de Locuteur est représenté par rapport à un ensemble de Locuteurs bien appris.

Le but étant d'hériter, à partir de cet espace représentatif, de quelques connaissances pour la modélisation qu'on ne pouvait pas avoir avec peu de données.

Les principales techniques utilisées dans ce domaine sont :

RMP (*Regression-Based Model Prediction*), Speaker Clustering, RSW (*Reference Speaker Weighting*) et les voix propres (*eigenvoices*).

Plus récemment, la représentation relative a été introduite et appliquée en reconnaissance des Locuteurs.

Les principales techniques utilisées en RAL sont :

1. *Non Directly Acoustic Process*

   Chaque Locuteur dans cette technique est caractérisé par rapport à un ensemble de Locuteurs dont les modèles sont bien appris [Merlin et al., 1999].

Dans [Merlin et al., 1999] à titre d'exemple, les Locuteurs de référence sont tirés d'une façon aléatoire.

Les Locuteurs sont projetés dans le nouvel espace de représentation. Cette projection est faite pour l'évaluation d'un score de vraisemblance entre le signal projeté et l'ensemble des Locuteurs de référence.

L'étape d'apprentissage dans cette technique est réalisée par un algorithme de classification tel que K-means. Par conséquent, chaque Locuteur est représenté par deux vecteurs : vecteur centre de gravité et vecteur des variances.

Dans la phase de test, on mesure l'angle entre le vecteur du signal de test et les centres de gravité des modèles des Locuteurs à identifier.

2. Les Modèles d'Ancrage

Les modèles d'ancrage sont largement inspirés de l'approche précédente.

Ils consistent à représenter et à caractériser un Locuteur par rapport à un ensemble de modèles de Locuteurs bien appris appelés "modèles d'ancrage" (*Anchor Models*).

Pour caractériser un Locuteur, on évalue un score de vraisemblance entre les données du Locuteur et chaque modèle de référence (modèles GMM-UBM).

La phase de test consiste à appliquer la distance euclidienne entre les vecteurs d'un Locuteur cible et d'un Locuteur de test.

3. Les Voix Propres

Dans cette approche, le Locuteur est modélisé par ses coefficients de combinaison des voix propres.

La reconnaissance est faite via le calcul de la similarité entre les Locuteurs et ceci par une distance entre les coefficients [Thyes et al., 2000].

Dans notre travail, nous nous intéressons plus particulièrement à cette dernière technique, à savoir les techniques des «voix propres».

Pour mieux comprendre son principe et d'une façon générale le principe de reconnaissance par placement dans un espace de Locuteurs de référence, nous pouvons formuler le problème par la question suivante :

«Parmi un ensemble de S Locuteurs, peut-on sélectionner ou fabriquer un certain nombre de Locuteurs (de voix) dont chacun reflète une caractéristique de la voix humaine ?»

L'idéal serait d'avoir un espace orthogonal où chacun des axes évalue et détermine le sexe du Locuteur, son âge, l'intensité de sa voix, sa qualité, le débit et le rythme de la parole, etc.

Chaque Locuteur peut être représenté dans cet espace et son modèle λ approximé par la relation :

$$\lambda \approx \sum_{e=1}^{E} w_e \overline{\lambda_e} \qquad \textbf{(2.1)}$$

Où $\overline{\lambda_e}$ représente les vecteurs propres de l'espace représentatif ou les voix propres, E est la dimension de l'espace (le nombre de Locuteurs ou de voix propres).

Ainsi, on associe à chaque Locuteur λ un vecteur caractéristique w *:*

$$w = \{ w_e \}\ e=1,\ldots, E. \qquad \textbf{(2.2)}$$

De par l'équation (2.1), nous constatons que le problème de reconnaissance est divisé en trois phases essentielles comme le montre la figure 2.6 :

La première phase consiste à rechercher les $\overline{\lambda_e}$, ce qui se traduit par la construction de l'espace représentatif.

La deuxième étape consiste en la localisation des Locuteurs dans cet espace, soit en d'autres termes, la détermination des vecteurs w_e.

La dernière phase est une phase de décision où on évalue la proximité spatiale entre les Locuteurs placés dans l'espace.

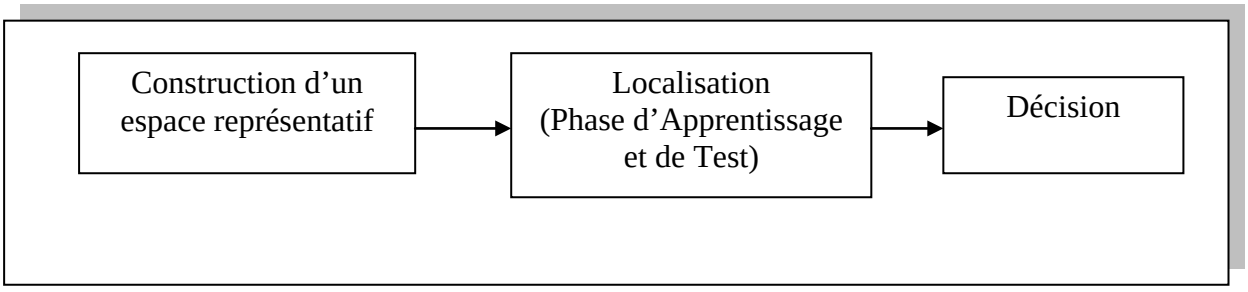

Figure 2.6 Système de reconnaissance par placement dans un espace de référence.

Selon la technique de construction de l'espace de représentation et la technique de localisation des Locuteurs, plusieurs variantes peuvent être explorées.

Dans les travaux de [Nguyen et al., 1999] et [Thyes et al., 2000] par exemple, la construction de l'espace est faite par les algorithmes de réduction de dimensionnalité, tels que l'ACP et l'ALD, ce qui génère un espace orthogonal.

La localisation est par conséquent faite par une projection orthogonale des vecteurs caractéristiques du Locuteur dans l'espace représentatif.

Mami et Charlet quant à eux, ont exploré d'autres voies pour la construction de l'espace.

On retrouve dans [Mami et al., 2002a] un espace obtenu par fusion des voix les plus proches et ceci, en exploitant l'algorithme de regroupement hiérarchique ascendant.

Dans [Mami et al., 2002b], l'espace est construit par un algorithme de sélection : selon un critère donné, on recherche un sous-ensemble de Locuteurs supposé le plus représentatif.

La localisation des Locuteurs dans les deux travaux se fait par la technique des modèles d'ancrage.

2.5.2.6 Tableau Récapitulatif des Approches de Modélisation des Locuteurs

Nous reprenons dans le tableau ci-après les approches de modélisation des Locuteurs en mettant l'accent sur leurs avantages et inconvénients.

Tableau 2.1 Etude comparative des Approches de Modélisation des Locuteurs [Debbeche et al., 2007a], [Debbeche et al., 2007b].

Approches	Avantages	Inconvénients
DTW	• Très rapide. • Présente des performances relativement bonnes.	• Utilisée exclusivement en mode dépendant du texte. • Très sensible à la qualité d'alignement des vecteurs et au choix du point de départ.
QV	• S'applique en mode dépendant ou indépendant du texte.	• Sa rapidité et ses performances dépendent fortement de la taille du dictionnaire.
Approches Connexionnistes	• Bonne performance.	• Complexité d'apprentissage. • L'ajout d'un nouveau client nécessite le réapprentissage de tous les modèles.
Approche Prédictive	• L'information dynamique transportée par le signal de parole est prise en considération.	• Les performances obtenues ne sont pas assez suffisantes pour un usage pratique.
Approche Relative	• La modélisation d'un nouveau Locuteu ne se fait plus de façon absolue mais relativement à un ensemble de Locuteurs bien appris.	• Le taux d'identification dépend de la quantité de données d'apprentissage pour la construction des Locuteurs de référence.
HMM	• Prend en considération l'aspect temporel du signal de parole • Excellents résultats en mode dépendant du texte.	• Utilisée uniquement en mode dépendant du texte.
GMM	• Très bonnes performances en mode indépendant du texte.	• Quantité importante de signaux d'apprentissage requise pour une bonne évaluation des paramètres du modèle.
MSSO	• Simplicité de mise en oeuvre • Performante sur de courtes durées.	• Ne capture que les caractéristiques stables le long du signal de parole. • les variations locales ne sont pas prises en compte.

2.5.3 Décision

Après avoir comparé le signal de test à tous les modèles de Locuteurs connus du système, on obtient un ensemble de mesures de similarité qui va servir d'entrée au module de décision.
Ce dernier a pour tâche de rechercher la mesure de similarité maximale ou bien minimale en terme de distance et d'indiquer l'identité du Locuteur.
Pour mesurer les performances d'un système d'IAL, on utilise généralement le taux d'identification correcte I_c ou incorrecte I_i qu'on obtient par les formulations suivantes :

$$I_c = \frac{\text{Nombre de tests correctement identifiés}}{\text{Nombre total de tentatives}} \qquad \textbf{(2.3)}$$

Et

$$I_i = \frac{\text{Nombre de tests mal identifiés}}{\text{Nombre total de tentatives}} \qquad \textbf{(2.4)}$$

Avec

$$I_c + I_i = 100\% \qquad \textbf{(2.5)}$$

2.6 CONCLUSION

Dans ce chapitre, nous avons introduit le principe de la Reconnaissance Automatique du Locuteur.
Nous avons décrit en premier lieu le Locuteur sous ses deux facettes anatomiques et acoustiques afin de justifier l'utilisation de la voix dans le domaine de reconnaissance.
Les différentes étapes d'un système de RAL ont été présentées ainsi qu'un état de l'art sur les approches de modélisation du Locuteur.
Nous avons mis l'accent sur l'approche relative qu'on utilisera par la suite pour la conception de notre système d'identification.

CHAPITRE III

LA FUSION DE DONNEES

On ne connaît pas complètement une science tant qu'on n'en sait pas l'histoire.
[Auguste Comte]

Cours de philosophie positive, 1830-1842

3.1 INTRODUCTION

La fusion de l'information est un sujet relativement ancien qui trouve ses origines à partir du moment où les chercheurs ont fait leurs premières tentatives d'imitation de l'intelligence humaine.

De nos jours, cet axe de recherche est en pleine effervescence. Cela est dû au fait que de très nombreuses applications nécessitent la résolution de problèmes abordant des données imparfaites ou imprécises, ou bien en général, l'estimation de grandeurs à partir d'informations, parfois contradictoires et hétérogènes, issues de plusieurs sources.

La fusion d'informations a beaucoup évolué ces dernières années dans différents domaines, et en particulier en vision et en robotique, les sources d'informations se sont multipliées, qu'il s'agisse de capteurs, d'informations *a priori*, de connaissances génériques, etc.

Chaque source d'information étant en général imparfaite, il est important d'en combiner plusieurs afin d'avoir une meilleure connaissance du « monde ».

La fusion d'informations peut alors se définir comme la combinaison d'informations afin d'obtenir une information globale plus complète, de meilleure qualité, et permettant de mieux décider et agir. Parallèlement, les méthodes pour modéliser les connaissances et les informations imparfaites et pour les combiner ont connu des développements théoriques importants et leurs champs d'applications se sont étendus. Ces méthodes sont souvent issues des théories de la décision, de l'incertain et de l'intelligence artificielle.

L'ampleur que prend la fusion d'informations suit celle que prennent les technologies et le traitement de l'information en général.

3.2 POURQUOI LA FUSION DE DONNEES ?

Les progrès permanents de l'informatique aussi bien du point de vue matériel que logiciel permettent de disposer d'informations de plus en plus riches et complexes, de nature et de fiabilité différentes. D'un autre côté, les problèmes abordant des données imparfaites ou entachées d'imprécision apparaissent de plus en plus fréquemment.

Parallèlement à ces états de fait, il est de plus en plus demandé aux systèmes d'information, de communication ou de commandement d'aider ou de coopérer avec les utilisateurs du domaine dans le but de prendre une décision.

Inspirée des mécanismes de raisonnement de l'esprit humain qui, lorsqu'il s'efforce de résoudre un problème donné, commence souvent par rassembler un maximum d'information provenant de diverses sources, la fusion de données est apparue.
Cette dernière permet de mettre à profit un nombre maximum de données, généralement de capteurs différents, en tenant compte de la diversité de leurs imperfections et en tentant de pallier les faiblesses de certaines avec les point forts des autres.
Ceci, dans le but de fournir une information élaborée, dédiée et pertinente vis-à-vis du contexte, qui ne peut, dans la majorité des cas, être obtenue en employant une seule source de données [Waltz, 1986], [Llinas et al., 1990], [Hall, 1992], [Klein, 1993].

3.3 DEFINITION DE LA FUSION DE DONNEES

3.3.1 Définitions diverses non satisfaisantes de la fusion de données

Diverses définitions de la fusion de données ont été proposées dans la littérature. Nous passerons en revue certaines d'entre elles en soulignant leurs avantages, inconvénients et points faibles.
Dans le domaine de la géographie par exemple comportant des images d'instruments aéroportés et des analyses d'intelligence rassemblée, les documents du consortium ouvert de GIS [Gis, 2000] définissent la fusion comme :

> « *Processus qui permet d'organiser, de fusionner et de lier des éléments de l'information disparates (exp. carte, images, textes, vidéo, etc.) pour produire une représentation cohérente et compréhensible d'un ensemble réel ou hypothétique d'objets et / ou d'événements dans l'espace et le temps* ».

Dans ces documents, la fusion est clairement définie comme étant un ensemble d'algorithmes, de techniques et d'opérateurs.

Dans le domaine de l'observation de la terre de l'espace, [Pohl et al., 1998] ont proposé la définition suivante :

> « *La fusion d'image est la combinaison de deux ou plusieurs images différentes pour former une nouvelle image en employant un certain algorithme* ».

Cette définition est limitée aux images.

[Mangolini, 1994] a prolongé la fusion de données à l'information en général et a ajouté une référence à la qualité. Il a défini la fusion de données comme :

> « *Ensemble de méthodes, d'outils et de moyens utilisant des données venant de diverses sources de nature différente, afin d'augmenter la qualité (au sens large) d'information sollicitée* ».

Cette définition met l'accent sur les méthodes, mais elle est limitée à ces dernières.

Dans les mathématiques appliquées et le traitement d'images, la définition proposée par [Hall et al., 1997] se rapporte également à la qualité de l'information et détaille les buts de la fusion de données. Cependant, elle met l'accent toujours sur les méthodes.

> « *Les techniques de fusion de données combinent des données provenant de capteurs multiples, et l'information connexe aux bases de données associées, pour obtenir une précision améliorée et des inférences plus spécifiques qui ne pourraient être réalisées en employant un seul capteur.* »

[Li et al., 1993] de leur côté ont écrit :

> « *La fusion se rapporte à la combinaison d'un groupe de capteurs en vue de produire un signal simple d'une plus grande qualité et fiabilité* ».

La qualité et la fiabilité sont mentionnées, mais il n'y a aucune référence aux concepts. En outre, cette définition est limitée aux capteurs et au signal.

3.3.2 Nouvelles définitions de la fusion de données

La définition de la fusion de données ne devrait pas être limitée aux méthodes et aux techniques ou se rapporter aux modèles ou aux architectures fonctionnels des systèmes.
Vu le manque constaté de définitions appropriées, un groupe de travail européen a été constitué en 1996 [Wald, 2000] sous l'auspice de SEE, la filiale française de l'IEEE, l'EARSeL et l'ISPRS (*the European affiliate of the International Society for Photogrammetry and Remote Sensing*).

Au cours de plusieurs réunions, la discussion s'est concentrée sur la formalisation de la fusion de données. Les résultats principaux étaient sur des définitions et des termes de références.
La définition suivante a été finalement convenue en janvier 1998 [Wald, 1998], [Wald, 1999] :

> « *La fusion de données constitue un cadre formel dans lequel s'expriment les moyens et techniques permettant l'alliance des données provenant de sources diverses. Cette fusion vise à obtenir une information de meilleure qualité* »

Les sens exacts des mots « meilleure » et « qualité » dépendront de l'application considérée.

[Bloch et al., 2001] ont à leur tour proposé une définition de la fusion lors des travaux du Groupe Européen de Travail sur la Fusion (FUSION) :

> « *La fusion consiste à réunir ou agréger des informations provenant de différentes sources, et à exploiter cette information réunie ou agrégée, dans diverses applications comme la réponse à une question, la prise de décision, une estimation numérique, etc*. »

Cette définition met l'accent sur deux éléments principaux. D'abord, elle met l'emphase sur la combinaison de l'information, puis l'accent est mis sur l'objectif de la fusion.

La définition de la fusion de données que nous avons choisie et que nous utiliserons tout au long de notre travail est celle donnée par [Bloch] :

> *« La fusion d'informations consiste à combiner des informations hétérogènes issues de plusieurs sources afin d'améliorer la prise de décision. »*

Cette définition est suffisamment générale pour englober la diversité des problèmes de fusion que l'on rencontre. Son intérêt est qu'elle est focalisée sur les étapes de combinaison et de décision ; ces deux opérations pouvant prendre des formes différentes suivant les problèmes et les applications.
Pour chaque type de problème et d'application, cette définition pourra être plus spécifique en répondant à un certain nombre de questions : quel est le but de la fusion ? Comment s'exprime la décision ? Quelles sont les informations à fusionner ? Quelles sont leurs origines ? etc.

3.3.3 Définition JDL de la fusion de données

En plus des définitions exposées dans les sections précédentes, on retrouve dans la littérature une autre définition de la fusion appelée JDL (proposée par le Département de la Défense des Etats-Unis d'Amérique).
C'est le fruit d'un effort fourni par le groupe de travail *Joint Directors of Laboratories* (JDL) *Data Fusion Working Group* établi en 1986 [Kessler et al., 1992], [JDL, 1991].
Ce groupe a aussi donné une codification de la terminologie et a créé un modèle de processus et un lexique pour la fusion de données.
Ceci dans le but d'améliorer la communication parmi les chercheurs et développeurs de systèmes en particulier dans le domaine militaire.
La définition de JDL de la fusion de données [JDL, 1991] et qui a été encore raffinée [DSTO, 1994] est formulée comme suit :

> « *Multilevel, multifaceted process dealing with the automatic detection, association, correlation, estimation, and combination of data and information from single and multiple sources.* »[1]

Cette définition est plus générale que les précédentes en ce qui concerne les types d'information qui peuvent être combinés. Elle est très appréciée de par la communauté militaire.

Le JDL a aussi proposé un modèle fonctionnel qui illustre les fonctions primaires, les informations pertinentes et les bases de données pour effectuer la fusion de données.
Le mot « *multilevel* » cité dans la définition se rapporte aux quatre niveaux du modèle fonctionnel. Par conséquent, on devrait joindre la description du modèle fonctionnel à la définition susmentionnée.

[1] La définition JDL ainsi que le tableau 3.1 n'ont pas été traduits en français pour préserver le sens.

Dans la littérature, cette définition est nécessairement associée à une description des quatre niveaux hiérarchiques (voir Figure 3.1).

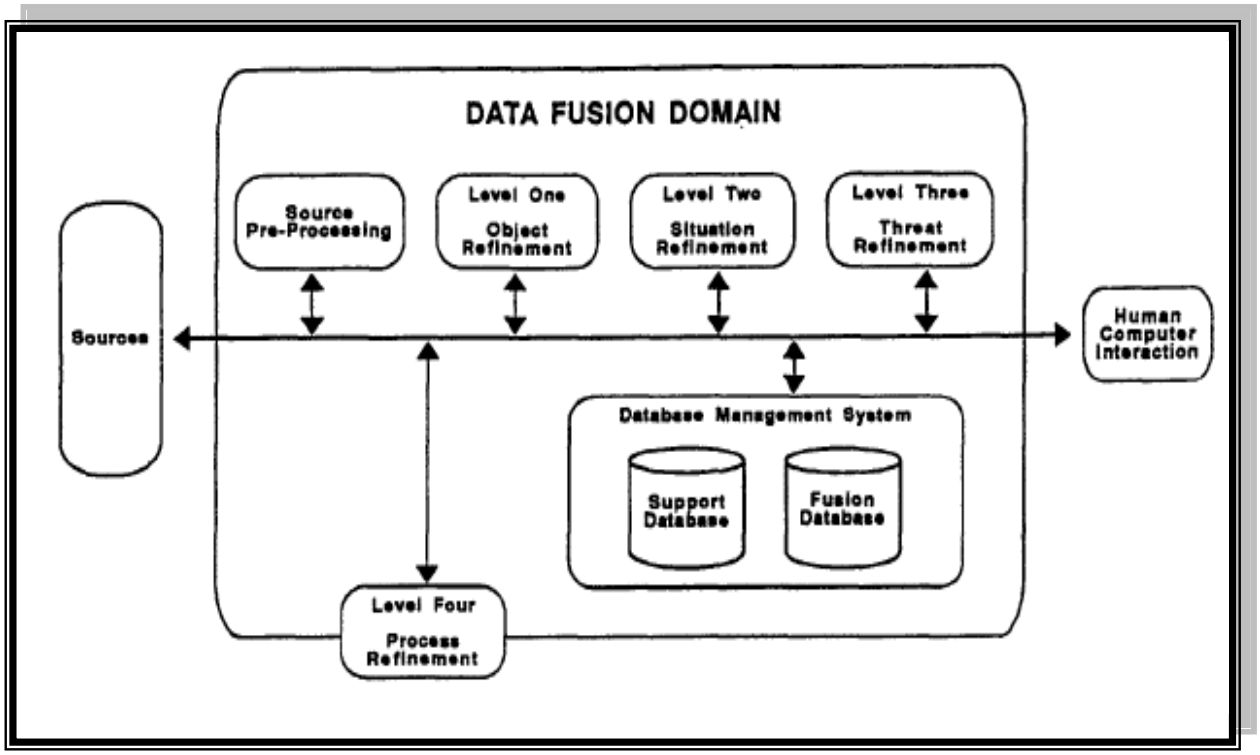

Figure 3.1 Modèle fonctionnel JDL de fusion de données [Hall et al., 1997]

Un résumé de ces niveaux est présenté dans le tableau 3.1.

Tableau 3.1 Composants du modèle JDL [Hall et al., 1997].

Sources	The sources provide information at a variety of levels ranging from sensor data to a priori information from databases to human input.
Process Assignment	Source preprocessing enables the data fusion process to concentrate on the data most pertinent to the current situation as well as reducing the data fusion processing load. This is accomplished via data prescreening and allocating data to appropriate processes.
Object Refinement (Level 1)	Level 1 processing combines locational, parametric, and identity information to achieve representatives of individual objects. Four key functions are: - transform data to a consistent reference frame and units; - estimate or predict object position, kinematics, or attributes; - assign data to objects to permit statistical estimation; and - refine estimates of the objects identity or classification.
Situation Refinement (Level 2)	Level 2 processing attempts to develop a contextual description of the relationship between objects and observed events. This processing determines the meaning of a collection of entities and incorporates environmental information, a priori knowledge, and observations.
Threat Refinement (Level 3)	Level 3 processing projects the current situation into the future to draw inferences about enemy threats, friendly and enemy vulnerabilities, and opportunities for operations. Threat assessment is especially difficult because it deals not only with computing possible engagement outcomes, but also assessing an enemy's intent based on knowledge about enemy doctrine, level of training, political environment, and the current situation.
Process Refinement (Level 4)	Level 4 processing is a mela-process, i.e., a process concerned about other processes. The three key Level 4 functions are: - monitor the real-time and long-term data fusion performance; - identif y information required to improve the multi-level data fusion product; and - allocate and direct sensor and sources to achieve mission goals.
Database Management System	Database management is the most extensive ancillary function required to support data fusion due to the variety and amount of managed data, as well as the need for data retrieval, storage, archiving, compression, relational queries, and data protection.
Human-Computer Interaction	In addition to providing a mechanism for human input and communication of data fusion results to operators and users, the human-computer interaction (HCI) includes methods of directing human attention as well as augmenting cognition, e.g., overcoming the human difficulty in processing negative information.

3.4 CONCEPTS DE LA FUSION DE DONNEES

3.4.1 Caractéristiques générales des données

Les caractéristiques générales des informations à fusionner doivent être prises en compte dans un processus de fusion [Bloch].
Une première caractéristique concerne le type de l'information à fusionner. Il peut s'agir d'observations directes, de résultats de traitements sur ces observations, de connaissances plus génériques exprimées sous forme de règles par exemple, ou d'avis d'experts. Ces informations peuvent être exprimées sous forme numérique ou sous forme symbolique.
D'autres distinctions sur les types de données sont également intéressantes à souligner, car elles donnent lieu à des modélisations et à des types de traitements différents : données fréquentes ou rares, informations factuelles ou génériques, génériques ou spécifiques, etc.

Une des caractéristiques importantes de l'information en fusion est son imperfection : c'est l'essence même du domaine de fusion.
Elle peut prendre diverses formes :

- L'**incertitude :** elle est relative à la vérité d'une information et caractérise son degré de conformité à la réalité [Dubois et al., 1988]. Elle fait référence à la nature de l'objet ou du fait concerné, à sa qualité, à son essence ou à son occurrence.
- L'**imprécision :** concerne le contenu de l'information et mesure son défaut quantitatif de connaissance sur une mesure [Dubois et al., 1988].
- L'**incomplétude :** caractérise l'absence d'informations apportées par la source sur certains aspects du problème.
- L'**ambiguïté :** exprime la capacité d'une information de conduire à deux interprétations. Elle peut provenir des imperfections précédentes.
- Le **conflit :** caractérise deux ou plusieurs informations conduisant à des interprétations contradictoires et donc incompatibles. Les situations conflictuelles sont fréquentes dans les problèmes de fusion, et posent toujours des problèmes difficiles à résoudre.

  La détection des conflits n'est pas chose facile ; ils peuvent facilement être confondus avec d'autres types d'imperfection, ou même avec la complémentarité des sources.

D'autres caractéristiques de l'information sont plus positives, et sont exploitées pour limiter les imperfections :

- **La redondance :** représente la qualité de sources qui apportent plusieurs fois la même information. La redondance entre les sources est souvent observée dans la mesure où les sources donnent des informations sur le même phénomène.
  Idéalement, la redondance est exploitée pour réduire les incertitudes et les imprécisions.

- **La complémentarité :** est la propriété des sources qui apportent des informations sur des grandeurs différentes. Elle vient du fait qu'elles ne donnent en général pas d'informations sur les mêmes caractéristiques du phénomène observé.
  Elle est directement exploitée dans le processus de fusion pour avoir une information globale plus complète et pour lever les ambiguïtés.

3.4.2 Types de Fusion

Après avoir connu les différents types et caractéristiques des données à fusionner, il convient de noter qu'une formalisation possible de la fusion de ces dernières introduit trois niveaux conceptuels [Bloch et al., 2001], [Dubois et al., 2004], [Dasarathy, 1997] qui sont :

- **La fusion de données :** c'est le niveau conceptuel le plus bas. Elle consiste essentiellement à marier des informations de bas niveau comme par exemple des primitives, dans le but de rendre l'information la moins bruitée que celle obtenue avec une seule source d'information.

- **La fusion de décisions :** ce type de fusion agit au niveau de l'espace de décision. Il effectue l'association d'informations élaborées (numériques ou symboliques) qui peuvent être considérées comme des propositions de décision.

- **La fusion de modèles :** ce cas est celui dans lequel les aspects complémentaires de différentes approches sont mis à partie pour combler les imperfections dont souffrent chacune d'entre elles indépendamment.

3.4.3 Etapes du Processus de Fusion de Données

L'opération de fusion s'effectue en plusieurs étapes [Hall et al., 1997], [El Faouzi, 2000], [El Faouzi, 2004] ; chacune correspondant à un ou plusieurs traitements des données à fusionner :

1. Représentation homogène et recalage des informations pertinentes : les données à fusionner sont souvent hétérogènes et il est impossible de les combiner sous leur forme initiale. On est alors amenés à rechercher un espace de représentation commun dans lequel les différentes informations pertinentes disponibles renseignent sur une même entité. Un premier traitement consiste donc à transformer certaines de ces informations initiales en informations équivalentes dans un espace commun dans lequel s'effectuera la fusion.

2. Modélisation des connaissances : chaque jeu de données propre à chaque source n'est pas forcément exploitable en tant que tel, notamment si l'information fournie est très imparfaite et ne donne qu'un aspect de la réalité. Cependant, même imparfaite, toute information peut apporter de la connaissance sur l'état du système. Donc, une étape essentielle du processus de fusion consiste à modéliser et à évaluer la connaissance apportée par chaque source. Elle est couplée au choix d'un cadre théorique adapté.

3. Fusion : c'est à ce niveau du processus que l'opération de fusion proprement dite est réalisée. Les informations recalées et modélisées sont combinées selon une règle de combinaison propre au cadre théorique choisi.

4. Décision par choix d'une stratégie : la fusion doit permettre de choisir l'information la plus vraisemblable, au sens d'un certain critère, parmi toutes les hypothèses possibles. En ce sens, la fusion de données aboutit bien souvent à une classification (affectation d'un ensemble de mesures aux hypothèses possibles). Le critère de décision dépend du cadre théorique dans lequel le processus de fusion a été développé, et de l'objectif à atteindre.

3.4.4 Architectures des Systèmes de Fusion de Données

Deux types d'architectures en système de fusion de données sont à distinguer.
Les premiers systèmes de fusion à avoir été mis en place ont été de types multi-capteurs. C'est le cas intuitif dans lequel l'objet est observé par plusieurs capteurs physiques (ou bien le

même mais selon différents angles d'observation). Dans ce cas de figure, les capteurs physiques sont employés pour accéder à différentes informations issues d'un ou de plusieurs objets d'une scène réelle.
Cependant, d'autres types d'architecture peuvent être envisagés comme par exemple les systèmes de fusion monocapteurs. De tels systèmes sont basés sur l'utilisation d'un seul capteur physique. Des informations nouvelles sont extraites de l'information brute initiale via des capteurs fictifs définis à partir de connaissances *a priori*.

3.4.5 Domaines d'application

La fusion est fédératrice d'actions inter-communautés [Reynaud, 1994]. On peut citer la communauté CHM (Communication Homme Machine), la communauté IA (Intelligence Artificielle) avec tous les travaux qui portent sur l'introduction de formes variées de raisonnement dans les algorithmes de traitement du signal et de l'image, et bien entendu la communauté Robotique.
Par conséquent, les domaines d'applications de la fusion peuvent être répertoriés comme suit :

1. Applications militaires comme [Hall et al., 1991] :
    - Détection, identification et suivi de cibles.
    - Surveillance des champs de bataille.
    - Détection de mines enfouies ou sous-marines.
    - Etc.

2. Applications aéronautiques et spatiales :
    - Imagerie satellitaire.
    - Commande d'engins spatiaux (fusées et robots).

3. Applications médicales :
    - Observation du corps et des pathologies.
    - Aide au geste et au diagnostic médical.

4. Robotique et véhicules intelligents [Abidi et al., 1992]:
    - Robots d'assistance humaine (fauteuils roulants, véhicules automobiles, machines agricoles,…)

- Robots autonomes en environnement difficile (robots sous-marins, robots d'intervention, micro-robots,…)

5. Assistance à l'opérateur humain :
   - Salle de contrôle (aiguilleurs du ciel).
   - Aide au diagnostic
6. Etc.

3.5 Avantages de la fusion de données

La fusion de données offre de multiples avantages (Figure 3.2) [Llinas et al., 1990] ; nous en citons :

- Robustesse et fiabilité : Le système est opérationnel même si une des sources d'information ou plusieurs d'entre elles fonctionnent mal ou sont défectueuses.
- Accroissement de la qualité d'information déduite et réduction de la vulnérabilité du système.
- Ambiguïté réduite : L'information plus complète fournit une meilleure discrimination entre les hypothèses disponibles.
- Ce domaine fournit une solution à l'explosion d'information qui est disponible de nos jours.

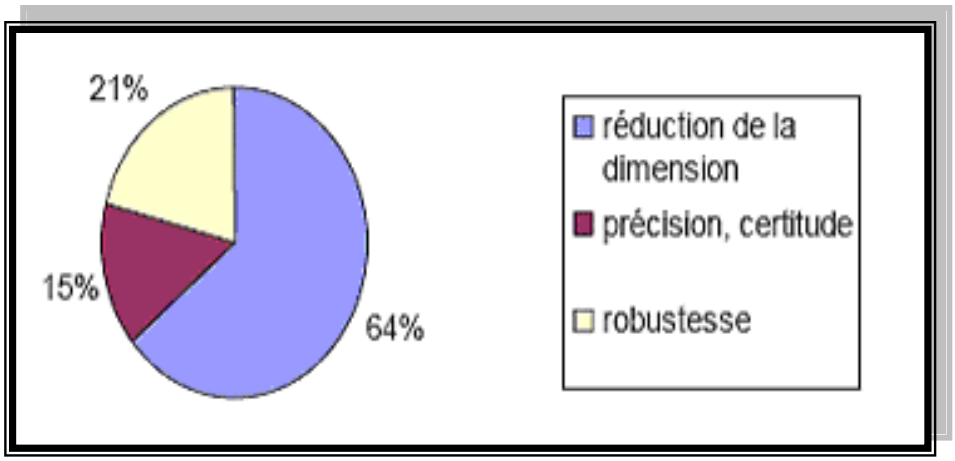

Figure 3.2 Objectifs de la fusion de données [Valet et al., 2000].

3.6 APPROCHES DE FUSION DE DONNEES

On présente dans cette section les principes de base des théories les plus classiquement utilisées en fusion, à savoir, théorie des probabilités, théorie des possibilités, et théorie de l'évidence.

Toutes les trois possèdent des fondements mathématiques solides, bien que les deux dernières soient de loin les plus récentes. On donnera aussi un aperçu sur les réseaux de neurones, alternatives possibles pour faire de la fusion.

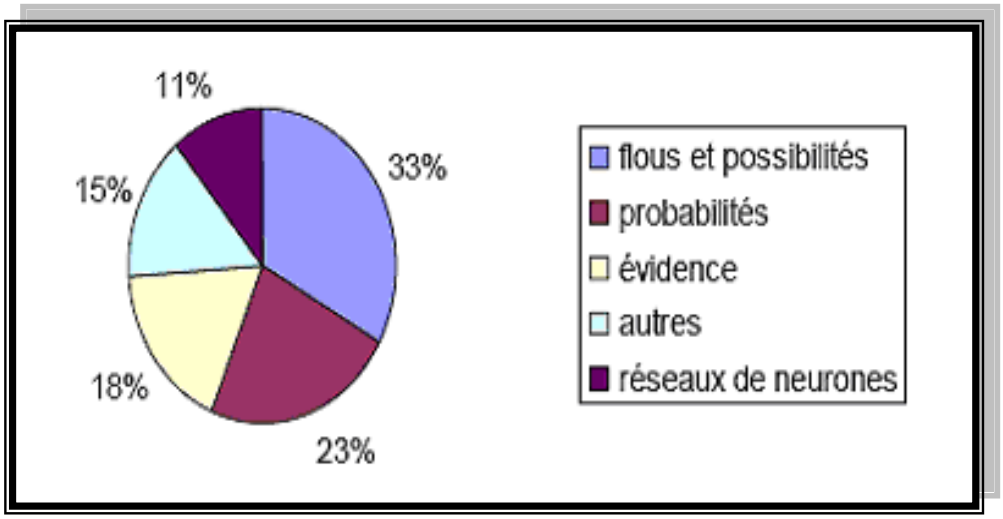

Figure 3.3 Méthodes de fusion de données [Valet et al., 2000].

3.6.1 Théorie des probabilités

La théorie des probabilités n'a plus à être présentée [Saporta, 1990]. C'est l'une des approches les plus utilisées sur des applications pratiques. Par contre, elle n'apparaît ces dernières années que dans 23 % des publications de recherche (voir Figure 3.3) [Valet et al., 2000], ce qui signifie que l'effort de recherche n'est pas prépondérant.

Cette théorie met à la disposition de l'utilisateur un certain nombre d'outils mathématiques qui lui permettent de régler la majorité des cas qui peuvent se rencontrer. Cependant, si en tant que théorie mathématique, la théorie des probabilités n'a pas à être justifiée, il en va autrement lorsqu'on cherche à appliquer le calcul des probabilités : on ne peut alors éluder la question de la nature de la probabilité et de la validité du modèle probabiliste.

Cette théorie est très performante lorsque l'on a une approche statistique du problème à traiter, ce qui n'est pas toujours le cas.

3.6.1.1 La distribution Gaussienne ou normale

La loi normale est une bonne représentation de l'état des connaissances sur les erreurs qui affectent les mesures.

En fait, la normalité n'est pas une hypothèse de nature physique, mais la description d'un état de connaissance de l'expérimentateur.

Supposons que l'on dispose de *n* observations s_i de la valeur de *x* avec $i \in [1, n]$.

On cherche à connaître la valeur vraie x_0.

Alors on a $s_i = x_0 + e_i$ où e_i est l'erreur inconnue faite lors de la mesure s_i.

Si l'on attribue une distribution gaussienne $p\left(s_i / x_0\right)$ à ces erreurs, on aura alors :

$$p\left(s_i / x_0\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp - \frac{1}{2}\left[\frac{\sum_n (s_i - \hat{x})^2}{\sigma^2}\right] \qquad \textbf{(3.1)}$$

Avec

$$\hat{x} = 1/n \sum_n s_i \qquad \textbf{(3.2)}$$

Et

$$\sigma^2 = 1/n \sum_n (s_i - \hat{x})^2 \qquad \textbf{(3.3)}$$

Seuls les deux premiers moments sur les données sont utilisés pour estimer la valeur de x à $\hat{x}$, et l'imprécision à σ.

La distribution gaussienne possède un certain nombre de propriétés mathématiques intéressantes qui explique en partie son utilisation fréquente :

- Le produit de deux fonctions gaussiennes est une fonction gaussienne,
- La convolution de deux fonctions gaussiennes est une fonction gaussienne,
- La transformée de Fourier d'une fonction gaussienne est une fonction gaussienne.

La distribution gaussienne est largement utilisée pour modéliser l'imprécision d'une source d'information.

Soit la mesure *s* de la grandeur *x*. Connaissant la distribution gaussienne représentant la distribution de probabilités des erreurs de mesures, la valeur estimée $\hat{x}$ de x est égale à s, et l'imprécision est estimée à σ.

$$p\left(x/s\right)=\left(\frac{1}{2\pi\sigma^2}\right)^{1/2}\exp-\frac{1}{2}\left[\frac{x-s}{\sigma}\right]^2 \qquad \textbf{(3.4)}$$

Les distributions de probabilités sont a
ussi utilisées pour modéliser l'incertitude sur des hypothèses.
Soit $\Omega=\{H_i, i=1,N\}$ l'ensemble des hypothèses possibles, et soit *s* une mesure. Compte tenu des capacités de la source d'information, on pourra définir une distribution de probabilités $P\left(\bullet/s\right)$ de la façon suivante :

$$P\left(\bullet/s\right): \qquad \Omega \to [0,1]$$
$$H_i \mapsto P\left(H_i/s\right)$$

3.6.1.2 Fusion d'informations

L'inférence Baysienne représente la méthode privilégiée de combinaison dans la théorie des probabilités. Elle se décline dans le cas continu comme dans le cas discret. Il s'agit d'évaluer la plausibilité de toute proposition en calculant la probabilité qu'elle soit vraie ou conditionnée par toute l'information disponible.

- **Cas discret**

  Supposons que l'on dispose de *n* observations s_i avec $i \in [1,n]$ pour estimer dans quelle mesure les hypothèses H_j de Ω sont vraies. Pour chaque source d'information, on dispose des probabilités conditionnelles $P\left(s_i/H_j\right)$ modélisant l'incertitude sur les mesures, c'est-à-dire la probabilité d'avoir la mesure s_i sachant que l'hypothèse H_j est vraie. On suppose aussi que l'on dispose de la probabilité *a priori* $P(H_j)$sur les hypothèses. Le théorème de Bayes s'exprime alors de la façon suivante (si les sources d'information sont indépendantes) :

$$P\left(H_j/s_i, i=1,n\right)=\frac{P(H_j)\bullet\prod_i P\left(s_i/H_j\right)}{\sum_{Hk}\left[P(H_k)\bullet\prod_i P\left(s_i/H_k\right)\right]} \qquad \textbf{(3.5)}$$

- **Cas continu**

  Supposons que l'on dispose de n observations s_i avec $i \in [1,n]$, mais cette fois pour estimer la valeur de la variable x. On suppose aussi que les erreurs sur les mesures sont modélisées par une distribution gaussienne :

$$P\left(x/s_i\right) = \left(\frac{1}{2\pi\sigma_i^2}\right)^{1/2} \exp - \frac{1}{2}\left[\frac{(x - s_i)}{\sigma_i}\right]^2 \qquad \textbf{(3.6)}$$

  Alors, le calcul de la valeur estimée $\hat{x}$ de x est basé sur l'application du théorème de Bayes :

$$P\left(x/s_i = 1,n\right) = \frac{P(x) \bullet \prod_i P\left(x/s_i\right)}{P(x) \bullet \int_{y \in Ix} \prod_i P\left(y/s_i\right) \bullet dy} = \frac{\prod_i P\left(x/s_i\right)}{\int_{y \in Ix} \prod_i P\left(y/s_i\right) \bullet dy} \qquad \textbf{(3.7)}$$

  On obtient une distribution gaussienne :

$$P\left(x/s_i = 1,n\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp - \frac{1}{2}\left[\frac{(x - \hat{x})}{\sigma}\right]^2 \qquad \textbf{(3.8)}$$

Où

$$\hat{x} = \frac{\sum_i s_i/\sigma_i}{\sum_i 1/\sigma_i} \qquad \textbf{(3.9)}$$

Tel que

$$\frac{1}{\sigma} = \sum_i \frac{1}{\sigma_i} \qquad \textbf{(3.10)}$$

3.6.1.3 Décision

Dans le cadre de la théorie des probabilités, la théorie de la décision est bien connue. Elle s'appuie sur l'utilisation du théorème de Bayes. Il s'agit essentiellement de maximiser la

probabilité *a posteriori*, ou plus généralement, la vraisemblance. On pourra aussi utiliser une représentation graphique plus connue sous le nom "d'arbre de décision".

3.6.2 Théorie de l'évidence

La théorie de l'évidence, appelée aussi théorie de la croyance ou théorie de *Dempster Shafer*, est relativement récente, surtout par rapport à la théorie des probabilités. Elle suscite beaucoup d'intérêts par sa nouveauté et sa puissance.
Elle est fondée sur les travaux de Dempster [Dempster, 1967] [Dempster, 1968], puis a été formalisée par Shafer [Shafer, 1976].
Cette théorie est à l'heure actuelle employée dans des domaines divers : fusion multi-capteur, reconnaissance des formes, etc. De plus, elle est considérée comme un cadre fédérateur des mesures de confiance.

3.6.2.1 Distribution de masses

Dans le cadre de la théorie de l'évidence, l'ensemble est appelé *cadre de discernement*. On suppose que les hypothèses dans Ω sont exclusives et que le cadre de discernement est exhaustif. On définit une masse de probabilités élémentaire, appelée *masse de croyance*, qui caractérise la véracité d'une proposition A pour une source d'information S.
La masse *m* associée à cette source est alors définie par :

$$m : 2^{\Omega} \rightarrow [0,1]$$

Et vérifie les propriétés suivantes :

$$m(\phi) = 0 \qquad \textbf{(3.11)}$$

$$\sum_{A \subseteq \Omega} m(A) = 1 \qquad \textbf{(3.12)}$$

Cette fonction se différencie d'une probabilité par le fait que la totalité de la masse de croyance est répartie non seulement sur les hypothèses singletons wq, mais aussi sur les hypothèses combinées.
La modélisation issue de la fonction *m* est appelée jeu de masses.

A partir de la fonction *m*, on définit respectivement les fonctions de *crédibilité Cr* et de *plausibilité Pl* par :

$$Cr(A) = \sum_{B \subseteq A} m(B) \qquad \textbf{(3.13)}$$

$$Pl(A) = \sum_{(A \cap B) \neq \phi} m(B) = 1 - Cr(\overline{A}) \qquad \textbf{(3.14)}$$

Avec

$Cr(\phi) = 0$ Et $Cr(\Omega) = 1$

Où $\overline{A}$ représente l'événement contraire de la proposition *A*.

La crédibilité Cr (A) mesure la force avec laquelle on croit en la véracité de la proposition.

La plausibilité *Pl* (A), fonction duale de la crédibilité, mesure l'intensité avec laquelle on ne doute pas de A.

Les deux grandeurs présentées ci-dessus englobent la probabilité inconnue *P* (*A*) d'un événement *A* si celle-ci existe :

$$Cr(A) \leq P(A) \leq Pl(A) \qquad \textbf{(3.15)}$$

3.6.2.2 Fusion d'informations

Si l'on dispose de plusieurs sources à partir desquelles sont définies des distributions de masse, il est possible d'en déduire une distribution de masse qui tienne compte de toutes les informations disponibles. Elle est obtenue en utilisant la règle proposée par Dempster sous sa forme non normalisée appelée aussi *somme orthogonale* (commutative et associative) que l'on note pour deux sources *S*1 et *S*2 :

$$m = m^{S1} \oplus m^{S2} \qquad \textbf{(3.16)}$$

$$m(A) = \sum_{B \cap C = A} m^{S1}(B) \bullet m^{S2}(C) \qquad \textbf{(3.17)}$$

Où $\oplus$ représente l'opérateur de combinaison.

Cette combinaison a pour effet d'affecter la masse à des propositions dont le nombre d'éléments est plus faible que celui des propositions initiales. En effet, *A* est un sous-ensemble de *B* et de *C* car $A = B \cap C$.

La distribution de masse finale est donc plus précise que chacune des deux distributions initiales.

3.6.2.3 Conflit

La particularité de la théorie de l'évidence, est qu'elle propose une modélisation qui permet la quantification du conflit entre des sources. Par exemple, pour deux sources *S*1 et *S*2 :

$$m(A) = \frac{1}{1-k} \sum_{B \cap C = A} m^{S1}(B) \bullet m^{S2}(C) \qquad \textbf{(3.18)}$$

Où *k* (coefficient reflétant le conflit) est défini par :

$$k = m(\phi) = \sum_{B \cap C = \phi} m^{S1}(B) \bullet m^{S2}(C) \qquad \textbf{(3.19)}$$

$\frac{1}{1-k}$ est un terme de normalisation. Lorsque *k* est égal à 1, les sources sont en conflit total et les informations ne peuvent être fusionnées. Dans le cas contraire, lorsque *k* est nul, les sources sont en parfait accord.

Cette règle de fusion, déduite de la règle de conditionnement [Smets, 1990b], a été critiquée dans plusieurs travaux (parmi lesquels [Yager, 1987]), en particulier dans le cas de sources en conflit total.

Les causes du conflit peuvent être multiples : mauvais fonctionnement du capteur, mauvaise définition du cadre de discernement, mauvaise définition des fonctions de croyance, etc.

3.6.2.4 Décision

Dans la théorie des fonctions de croyances, plusieurs règles de décision sont possibles telles que le maximum de plausibilité, de croyance et de probabilité pignistique [Smets, 1990a], auxquelles on peut associer des critères de rejet, de coût, de préservation éventuelle des ambiguïtés, etc.

Pour plus de détails sur cette théorie, le lecteur peut se référer à [Smets, 1991] [Yager et al., 1994]

3.6.3 Théorie des possibilités

La théorie des possibilités suscite actuellement un intérêt général de la part des chercheurs qui éprouvent les besoins de généraliser des modes de raisonnement naturels, d'automatiser la prise de décision dans leur domaine et de construire des systèmes artificiels effectuant les tâches habituellement prises en charge par les humains.

Cette théorie a été définie par Zadeh [Zadeh, 1978] et développée par Dubois et Prade [Dubois et al., 1988] à partir de la théorie des sous-ensembles flous [Zadeh, 1965]. C'est un outil mathématique qui permet une gestion efficace de l'imprécision et de l'incertitude qui peuvent être inhérentes à certaines données.

Il est important de rappeler la différence entre imprécision et incertitude :

L'imprécision concerne le contenu de l'information ; elle fait référence à la description incomplète d'un état de la réalité par une proposition.

L'incertitude quant à elle concerne la validité de cette information par rapport à une référence extérieure ; c'est aussi le fait de ne pas connaître ou prévoir l'état de la réalité pour déterminer la valeur de vérité d'une proposition. Elle est liée à la théorie de la probabilité.

On dira par exemple que « la distance est d'environ 20 cm » quand on parle en terme d'imprécision et « il est possible que la distance soit de 20 cm » lorsqu'il s'agit d'incertitude.

Pour mieux illustrer la différence entre ces deux concepts, on peut aussi citer l'exemple classique de Jim Bezdek [Bezdek, 1993]:

"Supposons qu'on se trouve dans un désert, après des jours d'errance... Presque mort de soif, on trouve alors 2 bouteilles remplies d'un liquide.

Sur la bouteille A, une étiquette annonce "potable avec un degré 0.91", et sur la bouteille B, l'étiquette dit "potable avec une probabilité 0,91".

Laquelle de ces 2 bouteilles peut-on boire ?".

Si l'on traduit les indications des étiquettes, on en retire qu'en buvant la bouteille A, on pourra s'en tirer avec comme seuls risques, quelques problèmes intestinaux non mortels... Par contre, en buvant la bouteille B, il y a une probabilité non négligeable (9 % de chance) que le liquide puisse être nocif (acide, ...) et carrément pas buvable.

3.6.3.1 Distribution de possibilités

Il est possible d'introduire les possibilités avec des considérations de logique.

En logique classique, toute proposition $A \subset X$ peut appartenir à trois ensembles *V*, *F* et *U* correspondant respectivement aux ensembles des propositions certainement vraies ou fausses, ou au contraire, incertaines. Ce dernier ensemble représente l'ensemble des propositions indécidables en l'état actuel des informations.

Les possibilités peuvent alors être présentées facilement car elles généralisent ces notions en introduisant des nuances.

Une première mesure $N(A) : X \rightarrow [0, 1]$ quantifie dans quelle mesure la proposition A appartient à l'ensemble *V* des propositions certainement vraies.

Si $N(A) = 1$, alors *A* est certainement vraie ;

Si $N(A) = 0$, alors *A* ne peut pas être classée pour l'instant dans cet ensemble.

Il faut alors introduire une deuxième mesure $\Pi(A) : X \rightarrow [0, 1]$ pour savoir si éventuellement la proposition pourra être classée dans l'ensemble des propositions vraies ou non.

Si $\Pi(A) = 0$, alors *A* est définitivement fausse ;

Si $\Pi(A) = 1$, *A* pourra être éventuellement classée vraie.

Il est possible de calculer ces deux fonctions à l'aide d'une fonction appelée distribution de possibilités π *des* propositions élémentaires dans [0,1].

On a alors les relations suivantes :

$$N(A) = 1 - \max_{x \notin A} \pi(x) \qquad \textbf{(3.20)}$$

$$\Pi(A) = \max_{x \in A} \pi(x) \qquad \textbf{(3.21)}$$

Avec la contrainte :

$$\max_{x \in X} \pi(x) = 1 \qquad \textbf{(3.22)}$$

3.6.3.2 Fusion d'informations

Afin de diminuer l'incertitude et l'imprécision sur les données, la solution est de fusionner le maximum d'informations, soit en faisant tendre N (A) vers 1, soit $\Pi(A)$ vers 0.

Pour fusionner en théorie des possibilités, l'information reflétée par chaque source doit être supportée par une distribution de possibilités.

Ces distributions peuvent être combinées conjonctivement ou disjonctivement ce qui correspond à une intersection ou une union.

Les opérateurs d'union et d'intersection couramment utilisés sont respectivement le max et le min, mais si on est moins exigeant sur la préservation des propriétés classiques des opérateurs ensemblistes, l'utilisation d'autres opérateurs de type t-conorme et t-norme [Weber, 1983], qui généralisent l'union et l'intersection, est envisageable.

Tableau 3.2 Exemples de t-normes et t-conormes.

	t-norme	t-conorme
Zadeh	min(x,y)	max(x,y)
Probabiliste	xy	x+y-xy
Lukasiewicz	max(x+y-1, 0)	min(x+y, 1)
Weber	x si y=1, y si x=1 0 sinon	x si y=0, y si x=0 1 sinon

La combinaison conjonctive suppose que les sources sont fiables, ou du moins certaines d'entre elles. La combinaison disjonctive, quant à elle, est moins contraignante en terme de fiabilité, dans le sens où elle nécessite seulement qu'une seule source soit fiable.

Evidemment, pour le cas particulier où on sait à l'avance qu'il n'y a aucune source fiable dans le processus de fusion, une autre attitude raisonnable consiste à choisir l'ignorance totale comme résultat de combinaison.

Un autre type de combinaison, intermédiaire entre les précédents, consiste à considérer le cas où on sait qu'il y a K sources fiables parmi les N disponibles, sans toutefois pouvoir les identifier.

Cette combinaison se résume à considérer tous les sous-ensembles à k éléments (qui sont susceptibles d'être les K sources fiables), à combiner conjonctivement les éléments de ces sous-ensembles, puis à combiner disjonctivement le résultat de chacune de ces combinaisons intermédiaires (attitude de prudence pour ne pas perdre un de ces sous-ensembles censé contenir l'information vraie). On aboutit alors à la formulation suivante, appelée combinaison quantifiée :

$$\forall x, \pi_k(x) = \max_{K \subseteq N, i \in K} \min \pi_i(x) \qquad \textbf{(3.23)}$$

L'utilisation d'un seul mode de combinaison figé n'est pas toujours souhaitable. Il est en effet intéressant d'évoluer du mode conjonctif, quand les sources s'avèrent fiables et non conflictuelles, vers un mode disjonctif lorsqu'elles s'avèrent par contre conflictuelles (ou non fiables). Dans ce but, Dubois et Prade [Dubois et al., 1994] ont développé une règle de combinaison adaptative.

Son principe est de déterminer deux valeurs n (optimiste) et m (pessimiste) encadrant l'ensemble des sources fiables. Ces valeurs correspondent respectivement au nombre maximum de sources (distributions) ayant une intersection non nulle de leur support (c'est-à-dire susceptibles d'être en concordance) et au nombre maximum de sources (distributions) ayant une intersection non nulle de leur noyau (c'est-à-dire en concordance totale).

De plus, on suppose qu'une fois ces deux valeurs déterminées, on ne sait pas distinguer les sources qui sont fiables de celles qui ne le sont pas. On utilise alors la combinaison quantifiée. La règle de Dubois et Prade se formule comme suit :

$$\pi(x) = \max\left(\frac{\pi_n(x)}{h(n)}, \min(\pi_m(x), 1 - h(n))\right) \qquad \textbf{(3.24)}$$

Avec

$$h(n) = \sup\left[h(K), [K] = n\right] \qquad \textbf{(3.25)}$$

$$h(K) = \sup\left[\min_{i \in K} \pi_i(x)\right] \qquad \textbf{(3.26)}$$

$$m = \sup\left[|K|, h(K) = 1\right] \qquad \textbf{(3.27)}$$

$$n = \sup\left[|K|, h(K) \succ 0\right] \qquad \textbf{(3.28)}$$

h (K) désigne la plus grande intersection entre les sources appartenant au sous-ensemble K de l'ensemble N des sources .

h (n) exprime la plus grande intersection non nulle du plus grand nombre de sources possibles .

Le choix d'un opérateur en théorie des possibilités peut se faire selon plusieurs critères [Bloch, 1996]. Un premier critère est le comportement de l'opérateur.
Des comportements sévères, indulgents ou prudents se traduisent sous forme mathématique de conjonction, disjonction ou compromis.
La combinaison est conjonctive si on opte pour un comportement sévère et disjonctive si le comportement est indulgent. Elle est dite de compromis si le comportement est prudent.
Cette distinction ne suffit pas à classer les opérateurs dont le comportement n'est pas toujours le même. Ainsi, la classification définie dans [Bloch, 1996] ne décrit pas les opérateurs seulement comme conjonctifs ou disjonctifs, mais aussi en fonction de leur comportement selon les valeurs des informations à combiner. Ainsi, les trois classes proposées correspondent aux :

1. Opérateurs Autonomes à Comportement Constant (ACC) : le résultat ne dépend que des valeurs à combiner (le calcul ne fait intervenir aucune autre information) et le comportement est le même quelles que soient ces valeurs.
2. Opérateurs Autonomes à Comportement Variable (ACV) : le comportement dépend des valeurs numériques des informations à fusionner.
3. Opérateurs Dépendant du Contexte (DC), par exemple d'une connaissance plus globale telle que la fiabilité des capteurs, ou encore le conflit entre les sources.

Cette classification, qui regroupe tous les opérateurs classiquement utilisés, constitue un premier critère de choix d'un opérateur pour une application spécifique.
Un deuxième critère est donné par les propriétés des opérateurs et leur interprétation en termes de fusion de données incertaines, imprécises, incomplètes ou encore ambiguës.
Enfin, l'étude du comportement des opérateurs en termes de qualité de la décision à laquelle ils conduisent et de réaction face aux situations conflictuelles, conduit à un autre critère de choix.
La capacité des opérateurs à combiner des informations quantitatives (numériques) ou qualitatives peut être également un critère de choix.
Par exemple, le min, le max et tout filtre de rang sont intéressants à ce titre puisqu'ils peuvent combiner les deux types d'informations.

3.6.3.3 Décision

Dans la théorie des possibilités, un certain nombre de critères peuvent être utilisés pour la prise de décision. Les deux principaux sont :
Le maximum de possibilité : c'est le critère "optimiste" qui privilégie la solution qui semble être la plus probable.
Le maximum de nécessité : c'est le critère "prudent" qui privilégie la solution pour laquelle une erreur serait risquée.

6.4 Les réseaux de neurones

Les réseaux de neurones sont souvent utilisés en classification, en identification et en reconnaissance des formes. En fusion de données, ils permettent de traiter la variabilité de procédés complexes et la disparité des informations provenant des capteurs. De plus, ils nécessitent en général des temps de traitement très courts. En fonction des valeurs d'entrée, ils évaluent des grandeurs de sortie.
L'intérêt des réseaux de neurones, c'est qu'ils ne nécessitent pas de modèle formel du phénomène observé [Dai et al., 1998] puisque leur fonctionnement est basé sur l'utilisation d'une connaissance obtenue par apprentissage.

6.5 Discussion

Nous avons passé en revue les différentes approches possibles pour traiter un problème de fusion de données.
La question qui se pose, et qui laisse souvent l'utilisateur perplexe, est comment choisir une approche pour traiter une application particulière ?
Chacune des approches présente des avantages ainsi que des inconvénients. L'approche probabiliste, par exemple, ne prend pas en compte le cas de données en conflit.
De plus, dans [Dubois et al., 1994] Dubois et Prade ont largement mis en évidence les limites de cette approche en particulier dans des cas d'informations pauvres.
Ils proposent en contrepartie l'utilisation de la théorie des possibilités pour remédier à ces limitations.
La théorie des possibilités est aussi bien adaptée à la modélisation de l'imprécision. Elle présente cependant, le problème de choix de l'opérateur de combinaison.
La théorie de l'évidence, quant à elle, est bien appropriée pour la modélisation de la méconnaissance (i. e. Incomplétude d'informations).

Pour de plus amples informations, le lecteur pourra se référer à [Rombaut, 2001],[Rombaut, 2008].

Le choix d'une approche de modélisation n'est pas trivial. Il nécessite un effort de compréhension de toutes les théories en concurrence et un examen minutieux de la situation à modéliser.

Nous trouvons dans [Ben Amor et al., 2004] un guide pratique pour faciliter ce choix (résumé dans la figure 3.4) ; il procède selon les étapes suivantes:

Étape 1 :

Identification de la nature d'imperfection de l'information pour les évaluations des actions selon le critère à construire. À cette étape, il faut cocher l'une des deux cases suivantes selon la nature prédominante de l'imperfection de l'information présente :

- Incertitude
- Imprécision

Il faut en effet préciser si les imperfections de l'information reliées à ce critère sont de l'ordre des :

1. Incertitudes au sens d'un doute sur la validité d'une connaissance :
   - ✓ Données recueillies par un intermédiaire peu fiable (pas sûr de lui, susceptible de se tromper ou de donner intentionnellement des informations erronées),
   - ✓ Données difficiles à obtenir ou à vérifier,
   - ✓ Données prévisionnelles,
   - ✓ Données de nature aléatoire,
   - ✓ Incertitudes dues à des imprécisions ou à des incomplétudes.

2. Imprécisions au sens d'une difficulté dans l'énoncé d'une connaissance :
   - ✓ Des catégories aux limites mal définies ("jeune", "centre ville", …),
   - ✓ Des situations intermédiaires entre le tout et le rien ("presque noir"),
   - ✓ Le passage progressif d'une propriété à une autre (notion de distance : proche, éloigné, …),
   - ✓ Des valeurs approximatives ("environ 2 km").

Etape 2 :

Cette étape commence à partir de l'une des deux cases : incertitude ou imprécision.

Si on a identifié un contexte d'incertitude pour le critère à construire, il faut répondre par oui (O) ou par non (N) à la question 1:

(1) Peut-on énumérer les différents états possibles influençant ou représentant les évaluations selon ce critère ?

L'identification d'un contexte d'imprécision par contre est suivie par la question 2 à laquelle il faut répondre également par oui (O) ou par non (N) :

(2) Les imprécisions portent-elles sur des données numériques approximatives que l'on peut exprimer par des intervalles ?

Étape 3 :

L'étape 3 procède à partir des réponses données aux questions 1 et 2.

Si, suite à la question 1, on constate qu'on est dans l'impossibilité d'énumérer les différents états possibles influençant ou représentant les évaluations selon le critère à construire (N), on devrait avoir recours à la théorie des possibilités pour modéliser l'incertitude en présence. Dans le cas contraire (O), on continue l'investigation par le biais du test de l'aléatoire (A) :

Test A :

• *Les évaluations selon ce critère sont des données numériques ou du moins mesurables sur des échelles standard (ratio, intervalle, …).*

• *Il existe peu d'intervenants humains non experts dans la situation d'incertitude à modéliser ; ces derniers introduisent des éléments d'imprécision par des descriptions subjectives ou des connaissances formulées en langage naturel.*

• *Il n'existe pas d'importantes connaissances graduelles ou de classes aux limites mal définies caractérisant la situation à modéliser.*

Si toutes les propositions énoncées dans le test A sont vérifiées, on y répondra par oui (O) et dans le cas contraire, on y répondra par non (N).

Si la question 2 montre que les imprécisions sont dues à des données numériques approximatives que l'on peut exprimer par des intervalles (O), il sera naturel de recourir à une modélisation par les intervalles. Sinon (N), on utilisera le langage du flou.

Etape 4 :

Si l'issue du test A est négative pour l'une ou l'autre des propositions énoncées (N), la théorie des possibilités est encore la plus en mesure de prendre en compte l'incertitude de la situation.
Si l'issue est positive (O) à toutes ces propositions, on continue avec la question 3 :

> *(3) Peut-on recueillir l'information sur le critère considéré (les évaluations des actions) selon des données stochastiques (modélisables par des distributions de probabilités) ?*

On répondra à cette question par oui (O) ou par (N).

Etape 5 :

Si la réponse à la question 3 est positive (O), on pose la question 4 à laquelle on répondra par oui (O) ou par non (N). Sinon, on pose la question 5.

> *(4) Connaît-on la distribution de probabilité avec précision ?*
>
> *(5) Les informations que l'on peut recueillir sur l'ensemble des états possibles identifiés à l'étape 2 portent-elles sur des sous-ensembles quelconques (Q) ou emboîtés (E) de cet ensemble ?*

Etape 6 :

Si la réponse à la question 4 est positive (O), on est amené à une modélisation par le langage des probabilités. Sinon (N), c'est le langage de l'ambiguïté qui est le plus approprié.
Si la réponse à la question 5 conduit à des sous-ensembles quelconques (Q), on optera pour la théorie de l'évidence comme langage de modélisation de l'imperfection de l'information.
Dans le cas d'ensembles emboîtés (E), on s'orientera vers la théorie des possibilités.

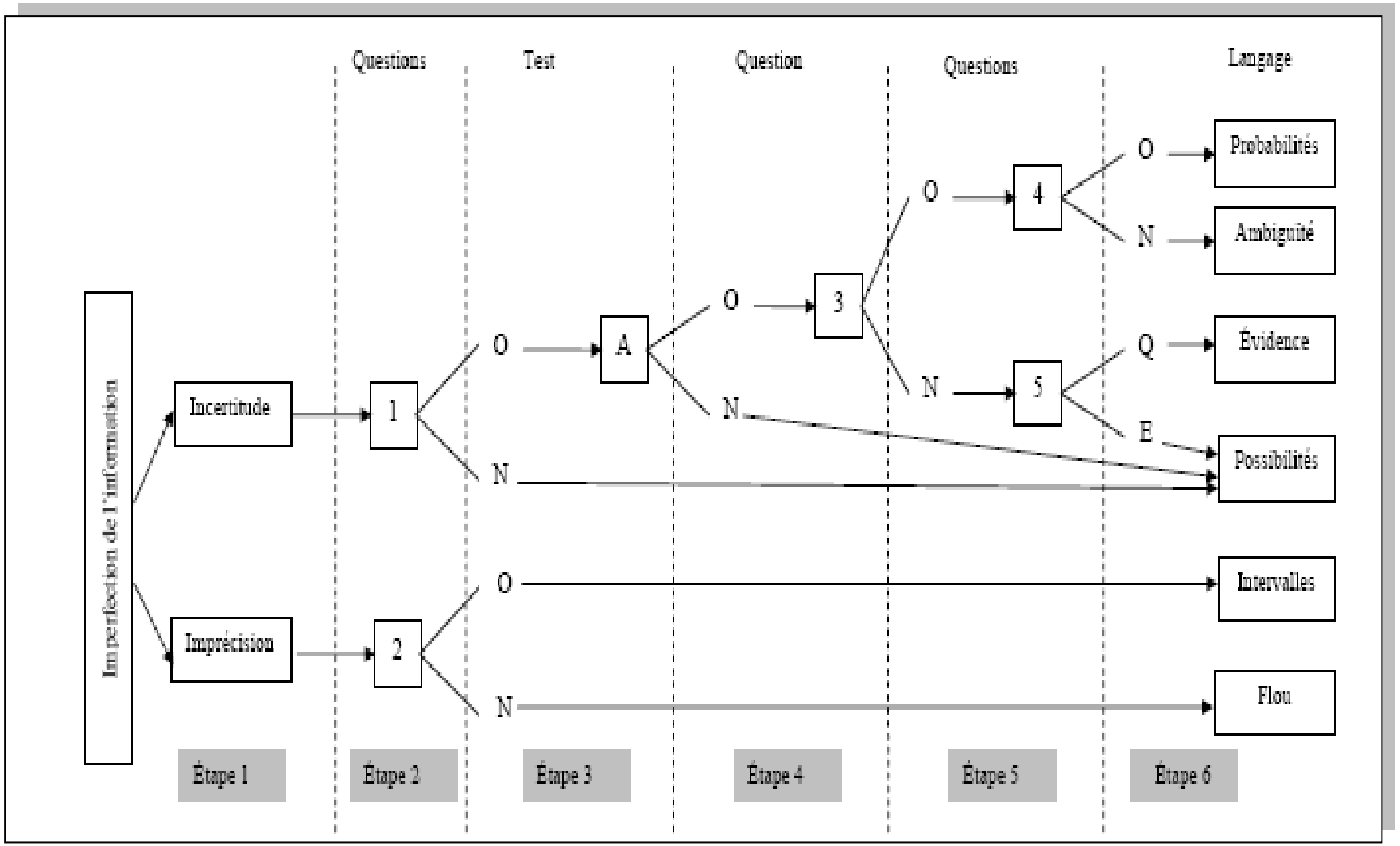

Figure 3.4 Choix d'une approche de modélisation des imperfections de l'information.

3.7 CONCLUSION

La fusion de données est un domaine très prometteur qui tire son essence de la multitude de sources et de capteurs de données existants de nos jours.

Accroître les performances des systèmes en renforçant leur robustesse et leur fiabilité et en réduisant l'erreur due essentiellement à l'imprécision, est le défit relevé par cet axe de recherche.

En effet, le résultat obtenu par la fusion est dans la majorité des cas plus élaboré et plus pertinent qu'un résultat obtenu par une seule source de données.

Nous avons passé en revue dans ce chapitre, les différentes définitions attribuées à la fusion de données en plus de ses différents aspects et avantages.

Les approches les plus connues en fusion ainsi que leurs principes ont été exposés.

Le chapitre est clôturé par une discussion, qui met en relief les avantages et les inconvénients de chacune des approches, suivie de la démarche proposée par [Ben Amor et al., 2004] pour en choisir une selon le problème abordé.

Nous visons par cela à bien comprendre le domaine afin de choisir convenablement l'approche qui va être utilisée pour la conception de notre système qui porte sur la fusion de paramètres acoustiques et anatomiques du Locuteur.

CHAPITRE IV

SYSTEME ACOUSTICO-ANATOMIQUE POUR L'IDENTIFICATION DES LOCUTEURS

---

*L'*homme *et sa* sécurité doivent constituer *la* première préoccupation *de* toute aventure technologique.

[Albert Einstein]

4.1 INTRODUCTION

Les chapitres précédents contenant un état de l'art sur les domaines touchant notre travail, représentent une introduction à ce chapitre.

C'est à partir de ce dernier, le plus conséquent, que commence notre contribution.

Notre objectif étant la conception d'un système d'IAL basé sur une fusion de paramètres acoustiques et de paramètres anatomiques du Locuteur et ce, dans le but d'améliorer le taux d'identification.

Pour ce faire, nous proposons une architecture d'un système acoustico-anatomique qui repose sur une nouvelle paramétrisation du Locuteur.

Cette paramétrisation est le fruit d'un algorithme de fusion qui représente le troisième volet de nos contributions.

Nous parlerons aussi dans ce chapitre, plus exactement au niveau de la paramétrisation du Locuteur, de la physiologie des cordes vocales afin de mettre l'accent sur leur importance dans le processus de phonation et de justifier leur utilisation dans notre système.

Nous détaillerons dans un second plan, l'approche de modélisation utilisée et qui est une modélisation relative.

Le corpus proposé pour une éventuelle implémentation du système est présenté dans la dernière section du chapitre.

4.2 PRESENTATION DU SYSTEME

4.2.1 Architecture du Système Acoustico-Anatomique

Après une étude de l'état de l'art du domaine d'IAL dans [Debbeche et al., 2007a], nous proposons une architecture acoustico-anatomique pour un système d'Identification Automatique du Locuteur [Debbeche et al., 2007b].

Notre système passe tout d'abord par une étape « off-line » consistant à exploiter une base de données complètement différente de celle qu'on utilise pour l'identification, et ceci dans le but d'obtenir des modèles GMM qui serviront à la construction de l'espace de référence (espace propre).

La deuxième partie du système englobe l'étape d'apprentissage et de test et ce, sur le corpus comprenant nos clients.

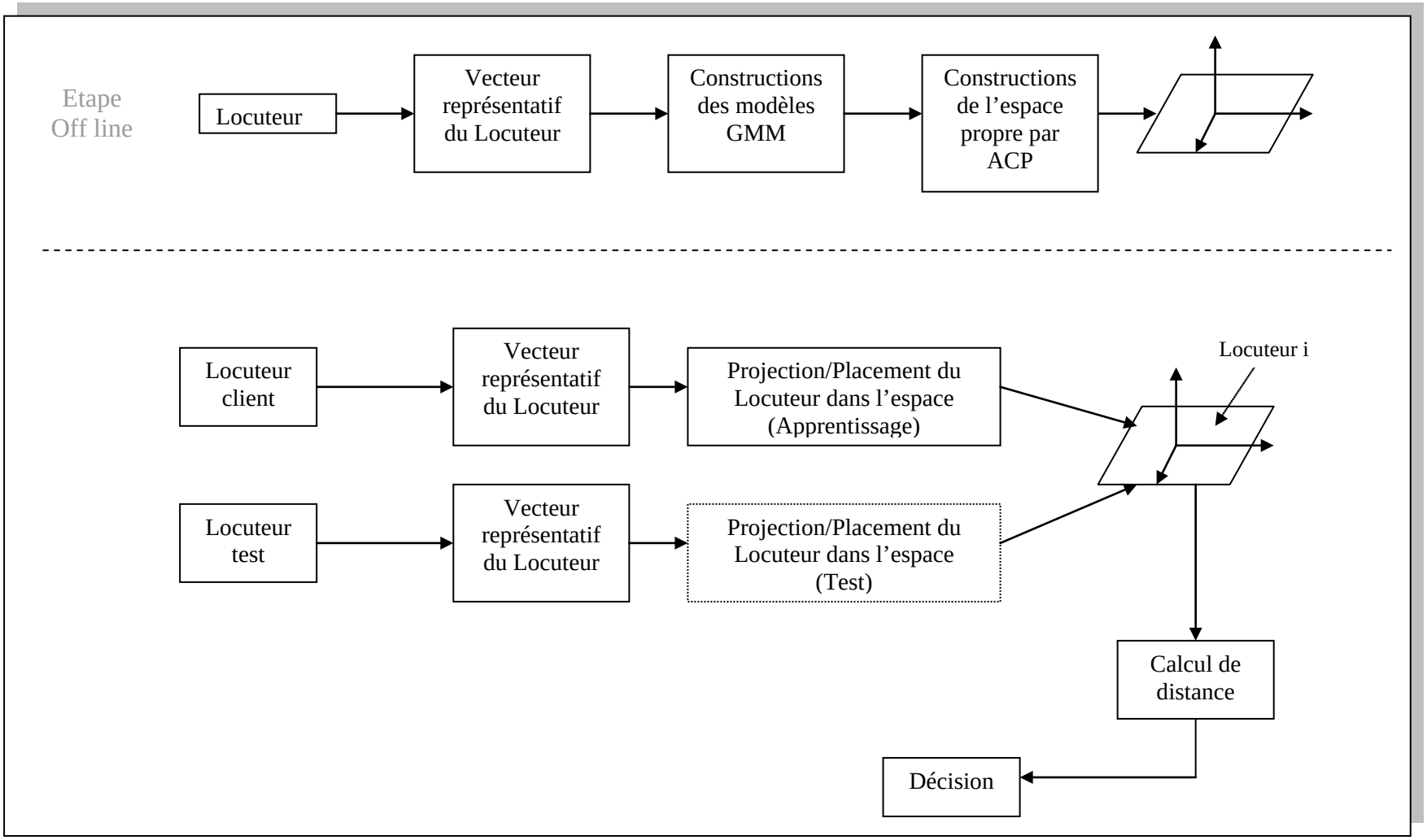

Figure 4.1 Architecture générale du système Acoustico-Anatomique.

Dans cette deuxième partie, nous retrouvons aussi la phase de décision ou d'identification où nous calculons la distance entre tous les vecteurs des Locuteurs clients et le vecteur du Locuteur de test.
L'architecture générale du système est schématisée dans la figure 4.1.

4.2.2 Paramétrisation du Locuteur

La paramétrisation du Locuteur que nous proposons [Debbeche et al., 2008a], [Debbeche et al., 2008b], schématisée dans l'architecture par le « vecteur représentatif du Locuteur », est obtenue par fusion de paramètres acoustiques et de paramètres anatomiques du Locuteur (voir figure 4.2).

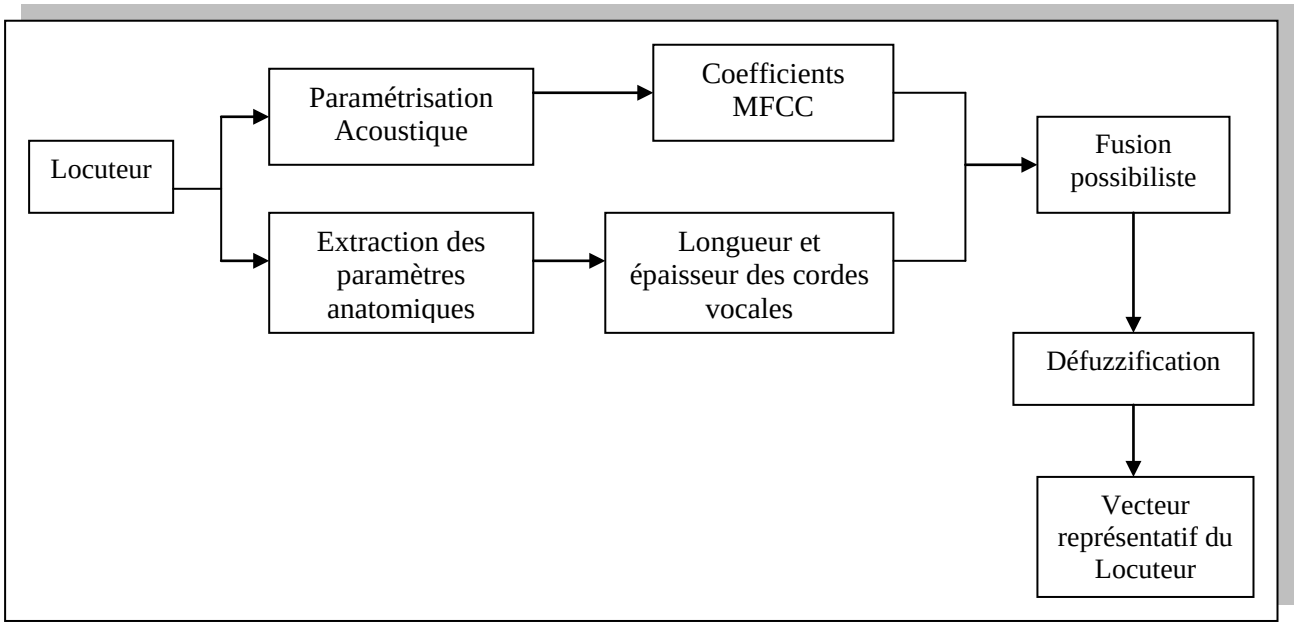

Figure 4.2 Paramétrisation du Locuteur.

4.2.2.1 Paramétrisation Acoustique

L'espace des paramètres acoustiques que nous utilisons dans notre étude est composé de 39 coefficients MFCC.
A chaque trame du signal de parole, nous associons un vecteur de représentation acoustique composé de l'énergie temporelle de la trame, des 12 premiers MFCC en plus de leurs dérivées premières et secondes.

4.2.2.2 Paramétrisation Anatomique

Les deux paramètres anatomiques choisis pour la paramétrisation du Locuteur sont la longueur *L* et l'épaisseur *d* de ses cordes vocales.

Ce choix repose sur le fait que cette partie du système vocal joue un rôle très important dans le processus de phonation (section 2.2.1 du chapitre II).

Rappelons que le son se produit lorsque le souffle (air provenant des poumons et s'écoulant dans le conduit vocal) passant au travers des cordes vocales les fait vibrer et est ainsi modulé par leur vibration.

Avant d'exploiter ces paramètres, nous nous devons d'abord de connaître la physiologie des cordes vocales. Ces dernières sont constituées d'une structure hétérogène de tissus en plusieurs couches appelées vocalis, ligaments et muqueuse.

Les figures 4.3 et 4.4 illustrent cette structure ainsi que les muscles qui contrôlent leur position et leur tension.

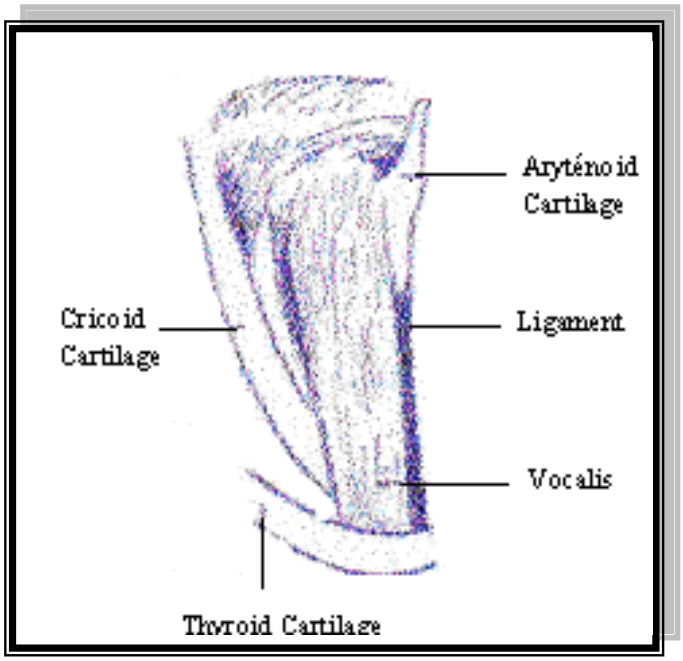

Figure 4.3 Position des cordes vocales. [Hirano, 1974]

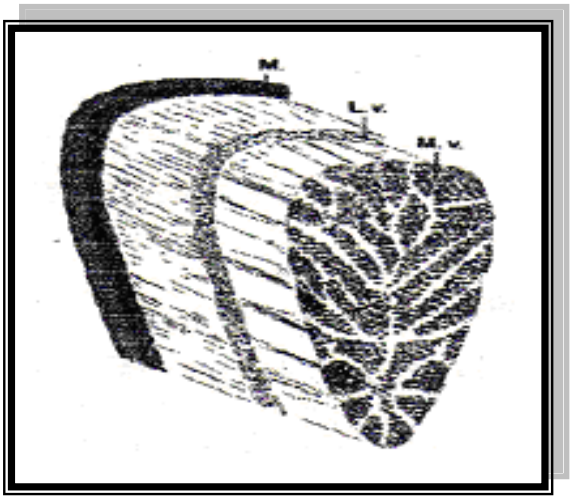

Figure 4.4 Couches des cordes vocales. [Hirano, 1974]

La motivation majeure de l'exploitation des paramètres anatomiques dans l'Identification Automatique du Locuteur, est la différence des dimensions de l'appareil vocal notée entre les individus. Aussi, la possibilité de mesurer ces dimensions à travers la technologie MRI (Magnetic Resonance Imaging) [Demolin et al., 1996].

Il faut savoir que chaque personne possède des dimensions de cordes vocales différentes des autres, ce qui donne des timbres de voix différents.

Par exemple, dans un cas normal, les cordes d'une femme sont plus longues que celle d'un homme. De plus, plus les cordes vocales sont petites, plus la voix est grave et vice-versa.

Pour mieux illustrer cette différence, nous trouvons dans [Kob] les valeurs typiques de la géométrie glottale (voir tableau 4.1, figure 4.5).

Tableau 4.1 Valeurs typiques de la géométrie glottale.

Longueur des cordes vocales	L_G	14 - 18 *mm*
Epaisseur des cordes vocales	d	5 – 10 *mm*
Distance glottale	h	0 - 3 *mm*

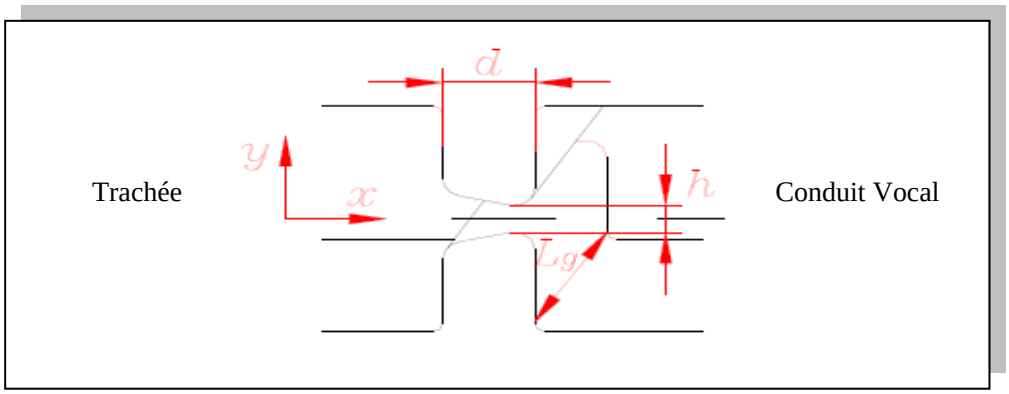

Figure 4.5 Géométrie de la glotte.

4.2.2.3 Fusion Possibiliste

La théorie des possibilités, comme nous l'avons exposée dans le troisième chapitre, permet une gestion efficace de l'imprécision et de l'incertitude qui peuvent être inhérentes à certaines données.

Elle est donc parfaitement adaptée à notre cas puisque les informations extraites, qu'elles soient d'ordre acoustique ou anatomique, sont toujours entachées de bruit et d'erreur.

Ainsi, pour fusionner nos données, nous suivons les étapes de l'algorithme de fusion schématisé dans la figure 4.6.

4.2.3 L'Algorithme Proposé pour la Fusion

La figure 4.6 représente l'algorithme général de la méthode de fusion. Il est composé de quatre parties principales :

1. L'information acoustique est extraite et modélisée sous forme de distributions de possibilités.
2. Les distances anatomiques exploitées sont à leur tour représentées par des distributions de possibilités.
3. La fusion possibiliste nous permet de fusionner les distributions de possibilités obtenues des étapes précédentes et produit une distribution fusionnée.
4. La défuzzification nous permet d'obtenir le vecteur acoustico-anatomique représentatif du Locuteur.

4.2.3.1 Modélisation possibiliste de l'information acoustique

Etant donné que le signal vocal obtenu lors de la phase d'enrôlement est toujours entaché de bruit (bruit ambiant, silence, etc.), l'application de la théorie des possibilités s'avère pertinente.

De ce fait, pour chaque vecteur acoustique $V = (x_1, ..., x_d)$ (où d est la dimension acoustique ; 39 dans notre cas) nous attribuons une distribution de possibilités :

$$\Pi(V) = (\pi_1, ..., \pi_d) \qquad \textbf{(4.1)}$$

Cette distribution est obtenue par une série d'expérimentations où nous travaillons sur les vecteurs acoustiques d'une même phrase.

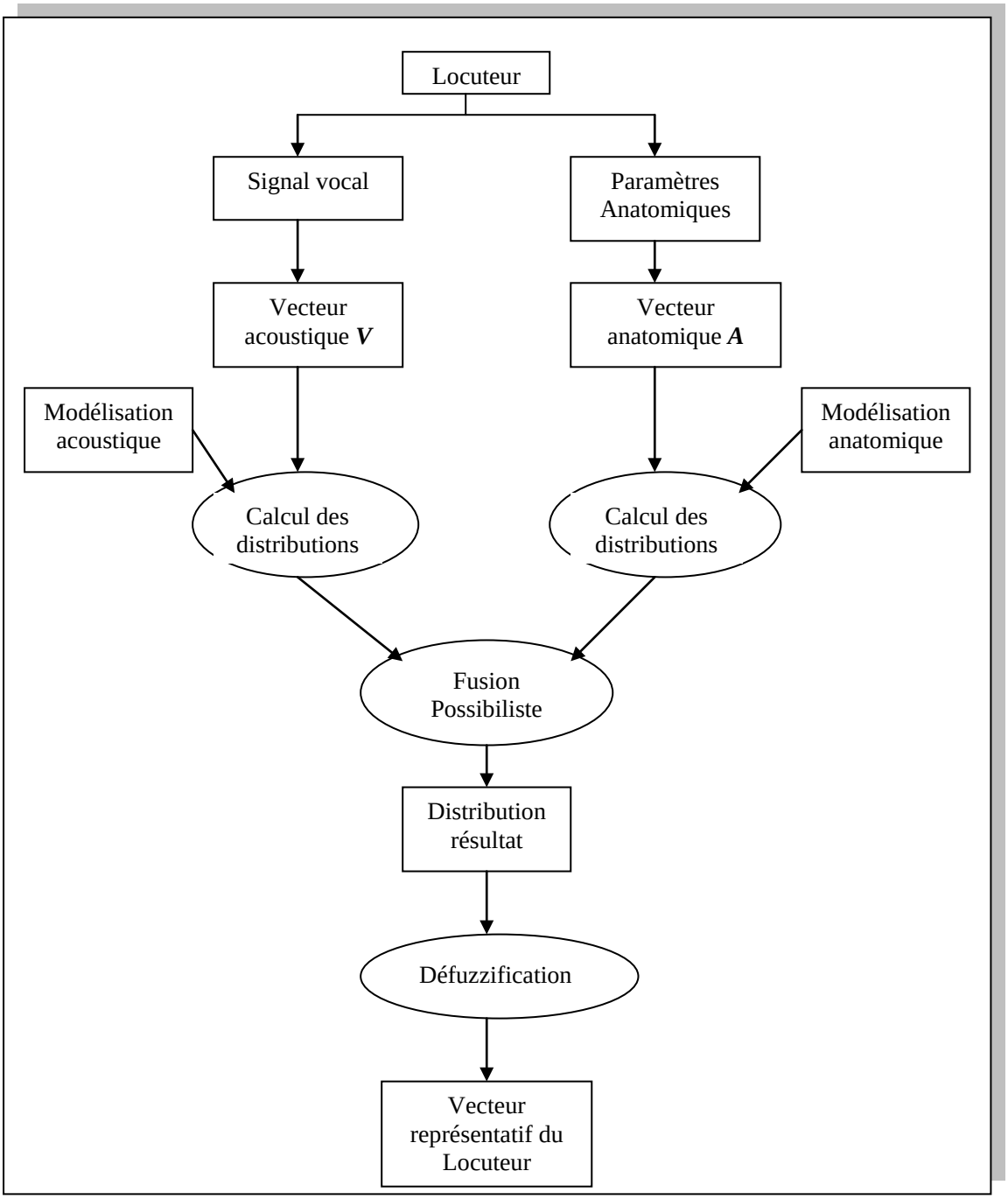

Figure 4.6 Algorithme général de la fusion.

4.2.3.2 Modélisation possibiliste de l'information anatomique

Comme nous l'avons déjà mentionné précédemment, nous exploitons pour l'identification du Locuteur deux dimensions de sa sphère ORL à savoir longueur et épaisseur de ses cordes vocales.

Par conséquent, chaque Locuteur est représenté à côté du vecteur acoustique V par un autre vecteur qu'on appellera vecteur anatomique A tel que :

$$A = (L, d)$$

La distribution de possibilités s'obtient à travers plusieurs séances de mesures effectuées pour chaque Locuteur (voir figure 4.7). Ces distributions sont élaborées subjectivement en demandant la coopération de professionnels en l'occurrence des médecins.

$$\Pi(A) = (\pi_L, \pi_d) \qquad \textbf{(4.2)}$$

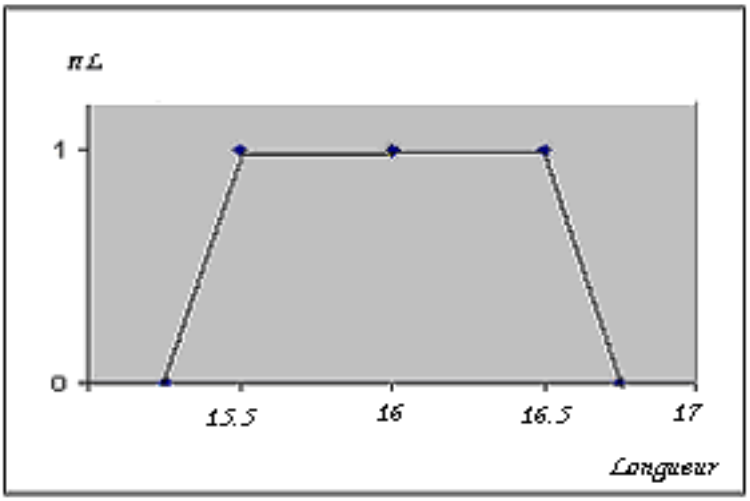

Figure 4.7 Exemples de distributions pour la longueur.

4.2.3.3 Fusion Possibiliste

La fusion des informations acoustiques et anatomiques dans un cadre possibiliste nous permet de faire face à l'imprécision et à l'incertitude des sources.

Pour la méthode de fusion, nous avons choisi l'opérateur de Dubois et Prade (déjà expliqué dans le chapitre 3, section 3.6.3.2).

$$\pi(x) = \max\left(\frac{\pi_n(x)}{h(n)}, \min(\pi_m(x), 1 - h(n))\right) \quad \textbf{(4.3)}$$

Avec

$$h(n) = \sup[h(K), [K] = n] \qquad h(K) = \sup[\min_{i \in K} \pi_i(x)]$$

$$m = \sup[|K|, h(K) = 1] \qquad n = \sup[|K|, h(K) \succ 0]$$

Ce choix est justifié car les sources dont nous disposons ne sont pas fiables, donc à travers cet opérateur, nous basculons du mode conjonctif vers le mode disjonctif.

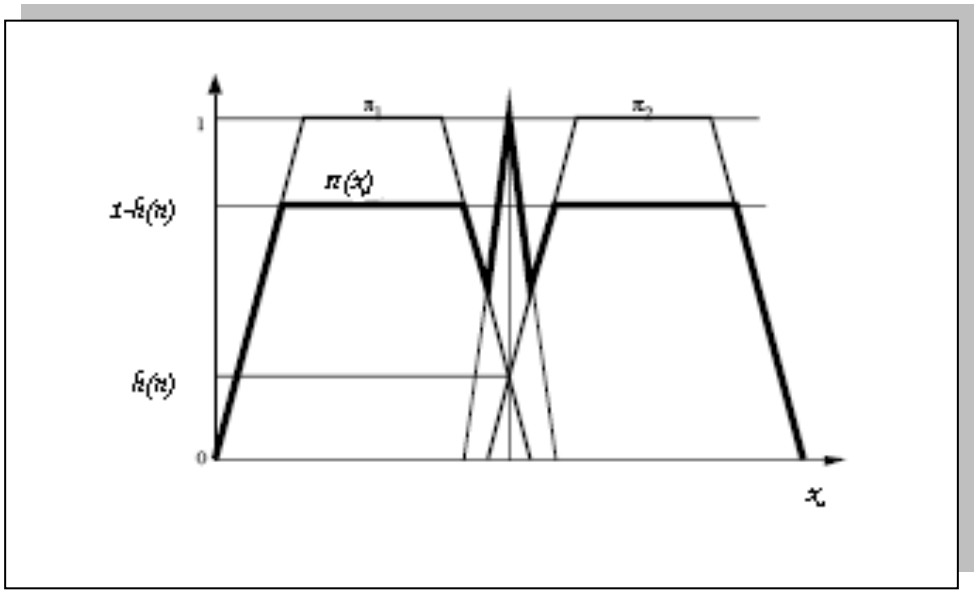

Figure 4.8 Exemples de détermination de la distribution résultant de deux distributions [Dubois et al., 1994].

4.2.3.4 Défuzzfication

Lors de la défuzzification, nous procédons à la transformation de la distribution de possibilités résultant de la fusion en une grandeur précise ; dans notre cas un vecteur représentatif du Locuteur.

Il existe plusieurs méthodes pour la défuzzification : défuzzification basée sur le centre de gravité, méthode des min/max ou celle des somme/produit.

La méthode que nous utilisons est celle du centre de gravité, la méthode la plus utilisée, qui nous permet de calculer l'abscisse du centre de gravité de la distribution résultat.
Le calcul du centre de gravité peut être ramené à la formule suivante :

$$x^* = \frac{\int_0^1 x\pi(x)dx}{\int_0^1 \pi(x)dx} \qquad \textbf{(4.4)}$$

4.2.4 Construction de l'Espace de Représentation

Comme c'est indiqué dans la figure 4.1 et plus exactement dans la partie *off-line,* la phase de construction de l'espace propre exploite les modèles GMM obtenus par l'apprentissage d'un ensemble de Locuteurs différents de nos Locuteurs clients (voir l'Annexe pour plus de détails sur l'approche d'apprentissage).
La construction de l'espace de représentation (espace propre) constitue une étape très importante dans l'approche relative. Elle consiste à créer une base de Locuteurs « virtuels », appelés « voix propres », ou bien un nouvel espace de représentation.
Pour ce faire, nous ne considérons que les moyennes des modèles GMM.
Nous appliquons dans notre système une des méthodes d'analyse de données qui a pour objectif général de décrire, synthétiser et expliquer l'information contenue dans de grands tableaux de données. Cette méthode est l'ACP (Analyse en Composantes Principales).
La méthode est utilisée comme suit :
Nous disposons de ***S*** Locuteurs modélisés par ***M*** gaussiennes dans un espace de dimension ***D.***
D représente la dimension du vecteur représentatif du Locuteur obtenu après la fusion acoustico-anatomique**.**
Par conséquent, nous travaillons sur un tableau de données ***R*** à $p = MxD$ lignes et $n = S$ colonnes.

$$R = \begin{pmatrix} \mu_1(1).....\mu_1(n) \\ \mu_2(1).....\mu_2(n) \\ . \\ . \\ . \\ \mu_{MxD}(1).....\mu_{MxD}(n) \end{pmatrix}$$

L'application de l'ACP nous permet d'obtenir un espace orthogonal Y de faible dimension ***(M x D) x E*** tel que ***E < S*** représentent les Locuteurs propres ou bien voix propres.

Cet espace de faible dimension ajuste au mieux le nuage de points–individus et celui de points-variables.

$$Y = \begin{pmatrix} \bar{\mu}_1(1).....\bar{\mu}_1(E) \\ \bar{\mu}_2(1).....\bar{\mu}_2(E) \\ . \\ . \\ . \\ \bar{\mu}_{MxD}(1).....\bar{\mu}_{MxD}(E) \end{pmatrix}$$

Le processus de la construction de cet espace passe par les étapes suivantes :

Tout d'abord, nous analysons le tableau dit centré plutôt que le tableau d'origine R de terme général r_{ij} .

Par $\bar{r}_j$ nous désignons la moyenne donnée par :

$$\bar{r}_j = \frac{1}{n}\sum_{i=1}^{n} r_{ij} \qquad \textbf{(4.5)}$$

Où n = S (nombres de Locuteurs)

Dans le but de faire jouer à chaque variable *j* un rôle identique, nous effectuons une translation de l'origine au centre de gravité du nuage et nous changeons les échelles sur les différents axes.

Le nouveau tableau de données $\Re$ est défini par le terme général suivant :

$$\Re_{ij} = r_{ij} - \bar{r}_j \qquad \textbf{(4.6)}$$

La recherche des composantes principales se traduit par la recherche des valeurs propres de :

$$\Re^{'}\Re V = V\Lambda \qquad \textbf{(4.7)}$$

Ou bien de :

$$\Re\Re' U = U\Lambda' \qquad \textbf{(4.8)}$$

Où Λ est la matrice diagonale n x n des valeurs propres $(\lambda_1, \lambda_2, ..., \lambda_n)$.

Notons que la matrice V est de dimension *p x n* et U est de dimension *n x n*.

Les composantes principales recherchées correspondent à la matrice :

$$V' = \Lambda^{-1/2} U^{'} \Re \qquad \textbf{(4.9)}$$

Cette matrice est de dimension S x (M x D). Chaque ligne est une voix propre modélisée par M x D paramètres décorrelés.

Nous ne retenons que les E premières lignes au lieu de garder S lignes.

La matrice Y de l'espace propre est donné par :

$$Y = V_E^{'} = (I_E | 0) V^{'} \qquad \textbf{(4.10)}$$

Où $(I_E | 0)$ est une matrice unité complétée par des zéros, donc la matrice Y s'obtient comme suit :

$$Y = \Lambda_E^{-1/2} U_E^{'} \Re \qquad \textbf{(4.11)}$$

4.2.5 Localisation des Locuteurs

Après construction de l'espace propre, nous procédons à la localisation des Locuteurs.

Cette étape représente la phase d'apprentissage de notre système ; elle est aussi réalisée pour le Locuteur de test.

Elle consiste à localiser les Locuteurs clients dans l'espace de représentation (voir figure 4.1).

A chaque Locuteur client, nous associons un vecteur de caractéristiques propres ou bien sa représentation spatiale :

$$w = [w_1 \ldots\ldots w_E]^T$$

Ce vecteur représente les coordonnées du Locuteur dans l'espace des Locuteurs de référence.
Comme nous avons utilisé l'ACP pour la construction de ce dernier qui est un espace orthogonal, nous faisons la localisation par projection orthogonale.

Soit ***s*** un Locuteur représenté par le vecteur $\mathbf{V^s}$ de dimension *M x D.*
La représentation spatiale est donc tout simplement une projection du vecteur $\mathbf{V^s}$ dans l'espace.

$$w = Y \cdot V^s \qquad \textbf{(4.12)}$$

4.2.6 Décision

Dans ce palier du système, nous exploitons les coefficients caractéristiques *w* des Locuteurs, calculés dans la phase précédente.
L'Identification dans ce cas se fait par calcul d'une distance, ce qui se traduit par la proximité des points de projection dans l'espace, c'est-à-dire plus les Locuteurs sont similaires, plus leurs points de projection sont proches et la distance entre eux est petite.
Donc, le Locuteur client dont le point de projection est le plus proche du Locuteur de test constitue le Locuteur reconnu.

Soit ***R*** un Locuteur à reconnaître et ***T*** un Locuteur de test représentés respectivement par les vecteurs $[r_1,\ldots,r_E]^T$ et $[t_1, \ldots,t_E]^T$.
L'identification se fait via la distance de Hamming dont la formule est la suivante :

$$d_1(R,T) = \sum_{i=1}^{E} |r_i - t_i| \qquad \textbf{(4.13)}$$

4.2.7 Corpus Proposé

Dans le cadre de notre projet, lequel est dédié à des environnements de très haute sécurité visant une population restreinte, nous avons construit un corpus comprenant 20 Locuteurs (12 femmes et 08 hommes) ; chaque Locuteur ayant réalisé deux sessions d'enregistrements.
Une session correspondant à un enregistrement de 10 phrases extraites d'un quotidien national et destinées à l'apprentissage, et de 03 phrases pour le test constituées d'une suite de six nombres.
Un deuxième corpus complètement différent du premier (de nos clients) est utilisé pour la construction des modèles GMM(s). Il est composé de 15 Locuteurs ayant prononcé 06 phrases constituées d'une combinaison de six nombres.

4.3 CONCLUSION

Dans ce chapitre, une architecture Acoustico-Anatomique pour l'Identification du Locuteur a été proposée. Notre architecture s'appuie sur une proposition d'une nouvelle Paramétrisation du Locuteur. Cette dernière se résume à la fusion de caractéristiques acoustiques (extraites du signal de parole) et de caractéristiques anatomiques du Locuteur, à savoir longueur et épaisseur de ses cordes vocales.
Prenant en considération que ces données sont imprécises, un algorithme pour la fusion exploitant les principes de la théorie des possibilités a été proposé.
Pour la modélisation des Locuteurs, et vu que nous visons des applications de très haute sécurité où la population est restreinte et où nous disposons de peu de données pour l'apprentissage, nous avons utilisé l'approche relative.

5.1 SYNOPSIS

Au cours de cette thèse, nous avons traité le problème d'Identification Automatique du Locuteur. Nous avons commencé par introduire la Biométrie, domaine de sécurité en pleine effervescence, qui a su remplacer les méthodes traditionnelles telles que : clef, mot de passe, badge, etc.

Cette introduction nous a permis de présenter le contexte général de nos recherches et de ce fait, comprendre l'utilisation et l'exploitation de la voix dans les systèmes de sécurité.

Dans la deuxième partie de ce document, nous avons donné une description détaillée du Locuteur ; dans un second plan nous avons rappelé les principes de la Reconnaissance Automatique du Locuteur (RAL) en présentant les différentes étapes du système de reconnaissance.

La part du lion a été consacrée à la modélisation du Locuteur où nous avons effectué un état de l'art sur les approches adoptées suivi d'une étude comparative.

Comme les systèmes de RAL et plus précisément d'IAL font toujours l'objet de recherche et attirent de plus en plus de chercheurs (car les performances n'ont pas encore atteint le seuil espéré par rapport à d'autres techniques biométriques), nous nous sommes proposé d'étudier une nouvelle paramétrisation du Locuteur basée sur une fusion de données hétérogènes qui pourrait en effet améliorer le taux d'identification.

Par conséquent, le troisième chapitre a été dédié à l'étude du domaine de fusion et principalement des approches qui y sont utilisées et la manière de choisir l'une d'entre elles.

Le dernier chapitre a fait l'objet de plusieurs propositions originales donnant naissance à un système Acoustico-Anatomique pour l'Identification du Locuteur.

5.2 CONTRIBUTIONS

Nos recherches ont abouti à trois principales contributions :

1. Proposition d'une Architecture d'un système Acoustico-Anatomique pour l'Identification Automatique du Locuteur.
2. Proposition d'une nouvelle Paramétrisation du Locuteur basée sur la fusion de paramètres acoustiques et de paramètres anatomiques offrant ainsi un Vecteur Acoustico-Anatomique représentatif du Locuteur.

3. Proposition d'un Algorithme de fusion pour l'obtention du vecteur cité ci-dessus, basé sur les principes de la théorie des possibilités.

5.3 PERSPECTIVES

Nous projetons d'implémenter les modules de notre architecture afin de valider nos objectifs.

[Abidi et al., 1992] Abidi, M.A., Gonzalez, FCC., Data Fusion in Robotics and Machine Intelligence, Academic Press, Boston, MA, 1992.

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

[Bartkova, 2002] Bartkova, K., Production, description et perception vocale, Rapport technique, France Télécom R&D, Lannion, 2002.

[Ben Amor et al., 2004] Ben Amor, S., Martel, J-M., Le choix d'un langage de modélisation des imperfections de l'information en aide à la décision, Faculté des Sc. de l'administration, Université Laval, ASAC, Québec, pages 2-10, 2004.

[Bennani, 1992] Bennani Y., Approches connexionnistes pour la reconnaissance automatique du locuteur : Modélisation et Identification, Thèse de l'Université de Paris Sud, 1992.

[Bennani et al., 1994] Bennani Y., Gallinari P., Connectionist approaches for automatic speaker recognition, ESCA, Workshop on Automatic Speaker Recognition identification verification, Martigny, pages 95-102, 1994.

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

[Bezdek, 1993] Bezdek, J., Fuzzy models—what are they and why? *IEEE Trans. Fuzzy Systems*, volume 1, pages 1–6, 1993.

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

[Biometrie] www.biometrie-online.net

[Bloch] Bloch, I., Fusion d'informations numériques : panorama méthodologique, Ecole Nationale Supérieure des Télécommunications, Paris, France.

[Bloch, 1996] Bloch, I., Information Combination Operators for Data Fusion : A Comparative Review with Classification, *IEEE Transactions on Systems, Man, and Cybernetics*, volume 26(1), pages 52–67, 1996.

[Bloch et al., 2001] Bloch, I., Hunter, A., Fusion : General concepts and characteristics, International Journal of Intelligent Systems, volume 16, pages 1107–1134, 2001.

[Bolt et al., 1970] Bolt, R. H., Cooper, F. S., David, E. E. Jr., Denes, P. B., Pickett, J. M., Stevens, K. N., Speaker Identification by Speech Spectrograms: A Scientists' View of its Reliability for Legal Purposes, Journal of the Acoustical Society of America 47, 2 (2), pages 597-612, 1970.

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

[Campbell, 1997] Campbell, J. P., Speaker Recognition: A Tutorial, Proceedings of the IEEE, 85, pages 1437-1462, 1997.

[Charlet, 1997] Charlet, D., Authentification vocale par téléphone en mode dépendant du texte, Thèse de l'Ecole Nationale Supérieure des Télécommunications, 1997.

[Clusif, 2003] Club de la Sécurité des Systèmes d'Information Français, Techniques de contrôle d'accès par biométrie, Commission Techniques de Sécurité Physique, Juin 2003.

[Dai et al., 1998] Dai, X., Khorram, S., Karimi, H., A neuronal network approach to data fusion in remote sensing, ASPRS-RTI, 1998.

[Dasarathy, 1997] Dasarathy, V. B., Sensor fusion potentiel exploitation–innovative architecture and illustrative applications, Proc . of IEEE Volume 85, pages 24-39, 1997.

[Debbeche et al., 2007a] Debbeche, F., Ghoualmi-Zine, N., Speaker's Atomatic identification: State of the Art, JED'2007, Annaba, 2007.

[Debbeche et al., 2007b] Debbeche, F., Ghoualmi-Zine, N., Towards an Acoustical-Anatomical System for Speaker Identification, ACIT2007, Lattakia Syria, November 26-28, 2007.

[Debbeche et al., 2008a] Debbeche, F., Ghoualmi-Zine, N., Système Acoustico-Anatomique pour l'Identification des Locuteurs par Localisation dans un Espace de Locuteurs de Référence, INFØDays'2008, pages 12-15, Chlef, 15-16 Avril 2008.

[Debbeche et al., 2008b] Debbeche, F., Ghoualmi-Zine, N., Système Acoustico-Anatomique pour l'Identification des Locuteurs par Localisation dans un Espace de Locuteurs de Référence, journée Jeunes Chercheurs en Informatique, JCI'08, page 21, Guelma, 20 Mai 2008.

[Demolin et al., 1996] Demolin, D., Metens, T., Soquet, A., Three-dimensional measurement of the vocal tract by MRI, Université Libre de Bruxelles, Service de Linguistique Générale, Unité de Résonance Magnétique de l'Hopital Erasme, Institut de Phonétique et de langues vivantes, 1996.

[Dempster, 1967] Dempster, A., Upper and lower probabilities induced by multivalued mapping, *AMS*, volume 38, pages 325–339, 1967.

[Dempster, 1968] Dempster, A., A generalization of bayesian inference, Journal of the Royal Statistical Society, volume 30,pages 205-247, 1968.

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

[DSTO, 1994] Defence Science and Technology Organization, Data Fusion Special Interest Group, Data fusion lexicon, Department of Defence, Australia, 21 September 1994.

[Dubois et al., 1988] Dubois, D., Prade, H., Possibility Theory, Plenum Press, New-York, 1988.

[Dubois et al., 1994] Dubois, D., Prade, H., Possibility theory and data fusion in poorly informed environments, Control Engineering Practice, volume2, pages 811-823, 1994.

[Dubois et al., 2004] Dubois, D., Prade, H., On the use of aggregation operations in information fusion process, Fuzzy Sets and Systems, pages 143–161, 2004.

[El Faouzi, 2000] El Faouzi N.E., Fusion de données : Concepts et méthodes. Rapport INRETS-LICIT, No 2002-10-15, Décembre 2000.

[El Faouzi, 2004] El-Faouzi, N.E., Fusion de données, Octobre 2004.

[Flanagan, 1972] Flanagan, J., Speech Analysis Synthesis and Perception, 2nd ed, New York and Berlin: Springer-Verlag, 1972.

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

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

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

[Gis, 2000] OpenGIS, Geospatial fusion services, The Open GIS Consortium (OGC), Wayland, Ma, USA, 2000.

[Hall et al., 1991] Hall, D.L., Llnn, R.J., Llinas, J., A Survey of Data Fusion Systems, Proceedings of the SPIE Conference on Data Structure and Target Classification, Volume 1470, pages 13-36, Orlando, FL, April 1991.

[Hall, 1992] Hall, D., Mathematical Techniques in Multisensory Data Fusion, Artech House, Inc., 1992.

[Hall et al., 1997] Hall, D ., Llinas J ., An introduction to multisensor data fusion, Proc. of IEEE, pages 112-148, 1997 .

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

[Hirano, 1974] Hirano, M., Morphological structure of the vocal cord as a vibrator and its variations, Folia Phoniatrica, Volume 26, pages 89-94, 1974.

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

[JDL, 1991] Data Fusion Lexicon, published by the Data Fusion Subpanel of the Joint Directors of Laboratories Technical Panel for C3 (F. E. White, Code 4202, NOSC, San Diego,CA), 1991.

[Kessler et al., 1992] Kessler et al., Functional Description of the Data Fusion Process, report prepared for the Office of Naval Technology, published by the Naval Air Development Center, Warminster, PA, January 1992.

[Klein, 1993] Klein, L.A., Sensor and Data Fusion Concepts and Applications, SPIE Optical Engineering Press, Tutorial Text, Volume 14, 1993.

[Kob] Kob, M., Physiologie des lèvres et des cordes vocales, Hôpital de phoniatrie, orthophonie et dysfonctionnements de communication, Université d'Aix la Chapelle – RWTH, Allemagne.

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

[Kuhn et al., 1998a] Kuhn, R., Nguyen, P., Junqua, J.-C., Goldwasser, L., Niedzielski, N., Fincke, S., Field, K., Contolini, M., Eigenvoices for speaker adaptation, ICSLP, volume 5, pages 1771-1774, Sydney, 1998.

[Kuhn et al., 1998b] Kuhn, R., Nguyen, P., Junqua, J.-C., Goldwasser, L., Niedzielski, N., Fincke, S., Field, K., Eigenfaces and eigenvoices: dimensionality reduction for specialized pattern recognition, MMSP, 1998.

[Kuhn et al., 1998c] Kuhn, R., Nguyen, P., Junqua, J.-C., Goldwasser, L., Niedzielski, N., Fincke, S., Field, K., Contolini, M., Eigenvoices for speaker adaptation, in ICSLP98, 1998.

[Kuhn et al., 1999] Kuhn, R., Nguyen, P., Junqua, J.-C., Boman, R., Niedzielski, N., Fincke, S., Field, K., Contolini, M., Fast Speaker Adaptation in Eigenvoice Space, ICASSP, 1999.

[Kuhn et al., 2000] Kuhn, R., Junqua, J.-C., Nguyen, P., Niedzielski, N., Rapid speaker adaptataion in eigenvoice space, IEEE transactions on Speech and Audio Processing, volume 8(6), pages 695-707, 2000.

[Ladefoged et al., 1980] Ladefoged, P., Ladefoged, J., The Ability of Listeners to Identify Voices, UCLA Working Papers in Phonetics 49, pages 43-51, 1980.

[Li et al., 1993] Li, H., Manjunath, B. S., Mitra, S. K., Multisensor image fusion using the wavelet transform, *Computer Vision, Graphics, and Image Processing: Graphical Models and Image Processing*, volume 57, pages 235-245, 1993.

[Llinas et al., 1990] Llinas, J., Waltz, E., Multisensory Data Fusion, Artech House, Inc., 1990.

[Magrin Chagnolleau et al., 1995] Magrin Chagnolleau, I., Bonastre, J.-F., Bimbot, F., Effect of utterance duration and phonetic content on speaker identification using Second order statistical methods, European Conference on Speech Comuunication and Technology (EUROSPEECH), volume 1, pages 337-340, 1995.

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

[Mahmoudi, 2000] Mahmoudi, D., Biométrie et Authentification, FI spécial été 2000, 5 Septembre 2000.

[Mami et al., 2002a] Mami,Y., Charlet, D., Identification des locuteurs par regroupement hiérarchique ascendant et modèles d'ancrage, XXIVèmes Journées d'Étude sur la Parole (JEP), pages 225-228, Nancy, 24-27 juin 2002.

[Mami et al., 2002b] Mami,Y., Charlet, D., Speaker identification by location in an optimal space of anchor models, International Conference on Spoken Language Processing (ICSLP), volume 2, pages 1333-1336, Denver (USA), 2002.

[Mangolini, 1994] Mangolini, M., Apport de la fusion d'images satellitaires multi capteurs au niveau pixel en télédétection et photo-interprétation, Thèse de Doctorat, Université Nice - Sophia Antipolis, France, 1994.

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

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

[Merlin et al., 1999] Merlin, T., Bonastre, J.-F., Fredouille, C., Non directly acoustic process for costless speaker recognition and indexation, in COST-254 International Workshop on Intelligent Communication Technologies and Applications, with emphasis on Mobile Communication, 1999.

[Musimem] www.musimem.com

[Nguyen et al., 1999] Nguyen, P[1,2]., Kuhn, R[1]., Junqua, J.-C[1]., Niedzielski, N[1]., Wellekens, C[2]., Voix Propres: Une représentation compacte de locuteurs dans l'espace des modèles, [1]Speech Technology Laboratory, Santa Barbara, Californie. [2]Institut Eurécom, Sophia-Antipolis, France. CORESA'99, 14-15 Juin 1999.

[Nolan, 1983] Nolan, J. F., The Phonetic Bases of Speaker Recognition, Cambridge University Press: Cambridge, 1983.

[Papcun et al., 1989] Papcun, G., Kreiman, J., Davis, A., Long-term memory for unfamiliar voices, J. Acoust. Soc. Am. 85, pages 913-925, 1989.

[Pohl et al., 1998] Pohl, C., Van Genderen, J. L., Multisensor image fusion in remote sensing: concepts, methods and applications, *International Journal of Remote Sensing*, volume 19(5), pages 823-854, 1998.

[Reynaud, 1994] Reynaud, R., La fusion de données, du capteur au raisonnement, Traitement du Signal, Volume 11(6), pages 431-434, 1994.

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

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

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

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

[Rombaut, 2001] Rombaut, M., Fusion : état de l'art et perspectives, Convention DSP 99.60.078, IUT de Troyes, laboratoire LM2S-UTT, pages 1-47, 22 octobre 2001.

[Rombaut, 2008] Rombaut, M., Fusion de données, Département Images et Signal, GIPSA lab, janvier 2008.

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

[Roublot, 2003] Roublot, P., Analyse comparative subjective et objective de la voix avant et après bloc interscalénique du plexus brachial, Université Henri Poincaré, Nancy I, 2003.

[Saporta, 1990] Saporta, G., Probabilités, analyse des données et statistique. Technip, 1990.

[Savoirs] www.savoirs.essonne.fr

[Shafer, 1976] Shafer, G., *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

[Schmidt et al., 2000] Schmidt-Nielsen A., Crystal, T. H., Speaker Verification by Human Listeners: Experiments Comparing Human and Machine Performance Using the NIST 1998 Speaker Evaluation Data, Digital Signal Processing vol. 10, no. 1-3. January/April/July, pages 249-266, 2000. (http://dx.doi.org/10.1006/dspr.1999.0356).

[Sécurite] www.securiteinfo.com

[Smets, 1990a] Smets, P., Constructing the Pignistic Probability Function in a Context of Uncertainty, *Uncertainty in Artificial Intelligence*, volume 5, pages 29–39, 1990.

[Smets, 1990b] Smets, P., The combination of evidence in the transferable belief model, *IEEE Transactions on Pattern ,Analysis and Machine Intelligence*, volume 12(5), pages 447–458,1990.

[Smets, 1991] Smets, P., What is dempster-shafer's model ? Technical Report TR/IRIDIA/91-20, IRIDIA Université Libre de Bruxelles, 50 Av F. Roosevelt. CP194/6.B-1050 Bruxelles, 1991.

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

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

[Stevens et al., 1968] Stevens, K. N., Williams, C. E., Carbonell, J. R., Woods, B., Speaker Authentication and Identification: A Comparison of Spectrographic and Auditory Presentations of Speech Material, J. Acoust. Soc. Am., 44(6), pages 1596-1607, 1968.

[Thyes et al., 2000] Thyes, O., Kuhn, R., Nguyen, P., Junqua, J.-C., Speaker identification and verification using eigenvoices, International Conference on Spoken Language Processing (ICSLP), 2000.

[Valet et al., 2000] Valet, L., Mauris, G., Bolon, P., A Statistic Overview of Recent Literature in Information Fusion, Proceedings of 3th International Conference on Information Fusion, pages MoC3 22-29, Fusion 2000, Paris, France, July10-13 2000.

[Van Lancker et al., 1985] Van Lancker, D., Kreiman, J., Emmorey, K., Familiar voice recognition: Patterns and parameters—Recognition of backward voices, J. Phonetics 13, pages 19-38, 1985.

[Wald, 1998] Wald, L., A European proposal for terms of reference in data fusion, International Archives of Photogrammetry and Remote Sensing, Vol. XXXII, Part 7, pages 651-654, 1998.

[Wald, 1999] Wald, L., Some terms of reference in data fusion, IEEE Transactions on Geosciences and Remote Sensing, volume 37(3), pages 1190-1193, 1999.

[Wald, 2000] Wald, L., The present achievements of the EARSeL - SIG "Data Fusion", In Proceedings of the EARSeL Symposium, held in Dresden, Germany, June 2000.

[Waltz, 1986] Waltz, E., Data Fusion for C31: A Tutorial, Command Control, Communications Intelligence (C31) Handbook, EW Communications, Inc., Palo Alto, CA, pages 217-226, 1986.

[Weber, 1983] Weber, S., A general concept of fuzzy connectives, negations and implications, based on *t*-norms and *t*-conorm , Fuzzy Sets and Systems, volume 11, pages 115-134, 1983.

[Wikipedia] http://fr.wikipedia.org

[Yager, 1987] Yager, R.R., On the Dempster-Shafer framework and new combination rules, *Information Sciences*, 41, pages 93–138, 1987.

[Yager et al., 1994] Yager, R.R., Fedrizzi, M., Kacprzyk ,J., Advances in the Dempster-Shafer theory of evidence,Wiley professional computing, 1994.

[Yarmey et al., 2001] Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., Parliament, L. Commonsense Beliefs and the Identification of Familiar Voices, Appl. Cognit. Psychol. 15, pages 283-299, 2001.

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

[Zadeh, 1965] Zadeh, L. A., Fuzzy sets, Information and Control, Volume 8, pages 338-353, 1965 .

[Zadeh, 1978] Zadeh, L. A., Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems, Volume 1, pages 3- 28, 1978 .

[Zwicker et al., 1981]Zwicker, E., Feldtkeller, R., Psychoacoustique, CENT/ENST, collection technique et scientifique des télécommunications, Mason Paris, 1981.

Les mélanges de gaussiennes

Les mélanges de gaussiennes sont utilisés pour modéliser un Locuteur donné par une somme pondérée de gaussiennes. Un modèle GMM peut être assimiler à un HMM à un seul état.
Cette méthode est la plus utilisée en ce qui concerne le Reconnaissance du Locuteur en mode indépendant du texte.

1. Modèle du mélange

Un mélange de gaussiennes est une somme pondérée de *M* densités gaussiennes.
Soit un Locuteur *s* et un vecteur représentatif de ce dernier *x* de dimension *D,* le mélange de gaussiennes est défini comme suit :

$$p(x|\lambda_s) = \sum_{m=1}^{M} \pi_m^s b_m^s(x) \qquad \textbf{(1)}$$

Où les $b_m^s(x)$ représentent des densités gaussiennes, paramétrées par un vecteur de moyenne μ_m^s et une matrice de covariance Σ_m^s :

$$b_m^s(x) = \frac{1}{(2\pi)^{D/2}.\left|\Sigma_m^s\right|^{1/2}} \times \exp\left[-\frac{1}{2}(x-\mu_m^s)'(\Sigma_m^s)^{-1}(x-\mu_m^s)\right] \qquad \textbf{(2)}$$

Et les π_m^s représentent les poids du mélange, avec $\sum_{m=1}^{M} \pi_m^s = 1$.

Un Locuteur est donc modélisé par un ensemble de paramètres noté λ_s :

$$\lambda_s = \left\{\pi_m^s, \mu_m^s, \Sigma_m^s\right\}_{m=1,\ldots,M} \qquad \textbf{(3)}$$

Ce modèle peut prendre plusieurs formes, notamment en ce qui concerne les matrices de covariance. On peut associer une matrice de covariance à chaque gaussienne ou bien utiliser une matrice de covariance globale commune à toutes les gaussiennes.
Ces matrices peuvent être pleines ou diagonales.

2. Apprentissage du modèle

La phase d'apprentissage consiste à estimer l'ensemble λ des paramètres d'un modèle GMM de Locuteur. La méthode conventionnelle est celle du Maximum de Vraisemblance dont le but est de déterminer les paramètres du modèle qui maximisent la vraisemblance des données d'apprentissage.

Pour une séquence de N vecteurs d'apprentissage $X = \{x_1, ..., x_N\}$ la vraisemblance du modèle GMM est :

$$p(X|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda) = \prod_{n=1}^{N}\sum_{m=1}^{M} p(x_n|\pi_m, \mu_m, \Sigma_m) \qquad \textbf{(4)}$$

☛ *Algorithme Expectation-Maximization (EM)*

Cet algorithme fait intervenir à la fois des observations X et des variables manquantes (l'indice de gaussienne *m = 1,……., M).*

Il maximise de façon itérative la fonction de la vraisemblance. Cette maximisation n'est pas directe, elle fait intervenir la fonction auxiliaire $Q(\theta, \theta^{(t)})$ qui est définie comme étant l'espérance mathématique du logarithme de la vraisemblance jointe (incluant les variables observées et les variables cachées) sur l'ensemble complet des variables d'entraînement, calculée sur la base des paramètres courants.

$$Q(\theta, \theta^{(t)}) = \sum_{m=1}^{M}\sum_{n=1}^{N} p(m|x_n, \theta^{(t)}) \log p(x_n, m|\theta) \qquad \textbf{(5)}$$

Où θ désigne l'ensemble des paramètres à estimer (π_m, μ_m, Σ_m) et $\theta^{(t)}$ l'ensemble des paramètres estimés à l'itération *t.* Ce qui donne après calcul :

$$Q(\theta, \theta^{(t)}) = \sum_{m=1}^{M}\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\log \pi_m - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_m| \right] - \sum_{m=1}^{M}\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\frac{1}{2}(x_n - \mu_m)' \sum_{m}^{-1} (x_n - \mu_m) \right] \qquad \textbf{(6)}$$

Où $\gamma_{n;m}^{(t)}$ est une probabilité a posteriori estimée à l'itération *t :*

$$\gamma_{n,m}^{(t)} = \frac{\pi_m^{(t)} p\left(x_n \middle| \mu_m^{(t)}, \Sigma_m^{(t)}\right)}{\sum_{k=1}^{M} \pi_k^{(t)} p\left(x_n \middle| \mu_k^{(t)}, \Sigma_{k(t)}\right)} \quad \textbf{(7)}$$

En supposant que $p\left(x_n \middle| \theta\right)$ sont des densités gaussiennes à matrices de covariances diagonales, l'expression de la fonction auxiliaire devient :

$$Q\left(\theta, \theta^{(t)}\right) = \sum_{m=1}^{M}\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \log \pi_m - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[Cste + \log \sigma_m^2 + \frac{\left(x_n - \mu_m\right)^2}{\sigma_m^2} \right] \quad \textbf{(8)}$$

Où σ_m^2 est un élément diagonal de la matrice de covariance.

Les paramètres sont estimés en annulant la dérivée partielle de la fonction auxiliaire *Q* par rapport à chacun de ceux-ci.

Le cas des poids des composantes de mélange π_m est assez simple puisqu'il s'agit de paramètres scalaires.

En ce qui concerne les vecteurs des moyennes, les formules de ré estimations sont données par :

$$\mu_m^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(t)} x_n}{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \quad \textbf{(9)}$$

Et les variances :

$$\sigma_m^{2(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(t)} (x_n - \mu_m^{(t)})^2}{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \quad \textbf{(10)}$$

☛ *Apprentissage par Maximum A Posteriori*

L'algorithme EM est un des algorithmes les plus importants et les plus puissants en estimation statistique ; il bénéficie d'une preuve de convergence garantissant que l'itération de l'étape d'estimation et de maximisation converge vers un maximum de la fonction de vraisemblance. Cependant, ses limites apparaissent lorsqu'on dispose de peu de données.
Donc, il est important d'introduire de l'information a priori. Par conséquent, on ne cherche plus à maximiser la vraisemblance des données mais plutôt la probabilité a posteriori.
Dans la littérature, l'apprentissage MAP est utilisé avec un choix des paramètres a priori donnant de ce fait un apprentissage incrémental.
Dans ce cas, les formules de ré estimation pour une gaussienne *m*, sont :

- Les poids des gaussiennes :

$$\pi_m = \frac{n_m^0 + n_m}{\sum_{k=1}^{M}\left(n_k^0 + n_k\right)} \qquad \textbf{(11)}$$

- Les vecteurs des moyennes :

$$\mu_m = \frac{n_m^0 \overline{X_m^0} + n_m \overline{X_m}}{n_m^0 + n_m} \qquad \textbf{(12)}$$

- Les variances :

$$\sigma_m^2 = \frac{n_m^0 \overline{X_m^0 X_m^{0'}}}{n_m^0 + n_m} - \mu_m \mu_m' \qquad \textbf{(13)}$$

Où *n* (respectivement n^0) représente les poids, $\overline{X}$ (respectivement $\overline{X^0}$) le moment d'ordre 1 et $\overline{XX'}$ (respectivement $\overline{X^0 X^{0'}}$) le moment d'ordre 2 des données à adapter X (respectivement des données initiales X^0).

L'apprentissage incrémental consiste à effectuer quelques itérations d'apprentissage sur les données d'adaptation en conservant l'information apportée par les données initiales X^0.
Dans le cas où de nombreuses données sont disponibles, l'apprentissage incrémental converge vers les estimateurs du maximum de vraisemblance. Il permet d'obtenir de nouveaux modèles avec peu de données.
Cette approche est la plus utilisée en reconnaissance du Locuteur en mode indépendant du texte.